Derivatives and Internal Models

Modern Risk Management

Fifth Edition

Hans-Peter Deutsch
Mark W. Beinker

d-fine

Palgrave Macmillan

Finance and Capital Markets Series

The Finance and Capital Markets series is designed to bring you high-quality, cutting-edge information on the latest issues and developments in the financial world. These authoritative yet accessible books, written by experts in their field, provide clear, practical guidance not only for industry professionals, but also serve as a framework of current analysis for scholars and form a comprehensive reference resource for libraries.

More information about this series at http://www.palgrave.com/gp/series/13977

Hans-Peter Deutsch • Mark W. Beinker

Derivatives and Internal Models

Modern Risk Management

Fifth Edition

palgrave
macmillan

Hans-Peter Deutsch
Niedernhausen, Germany

Mark W. Beinker
Friedrichsdorf, Germany

Preface to Fifth Edition

When the fourth edition of this book has been published, the financial crisis was already in full swing. It took a while, though, until its sustainable impact on the pricing of derivatives and the measurement of their risks became obvious. Meanwhile, new standards have been established. This fifth edition follows the goal to explain the most important consequences in a clear and intelligible manner.

This required to add a new chapter on the valuation of credit risk of derivatives. The chapter on the construction of interest rate curves has been completely rewritten, since the current multi-curve universe is substantially different from the former one curve framework of the world prior to the financial crisis. Cross references throughout the whole book refer to these new chapters to demonstrate, where these changes have an impact. The hope was to maintain the character of the book consistently throughout all chapters.

In addition, we took the opportunity to modernize and restructure the entire book. This includes modifications of tiny details (e.g., discount factors have no longer the interest rate as index, since this would cause confusion in a multi-curve setting). Tables and figures have been re-formatted, and the structure of the chapters have been modified. For pedagogical reasons, the section dealing with bi- and trinomial trees precedes now the chapter on finite difference methods. The systematic of financial instruments was re-arranged to better highlight the essentially different features of the different instruments. For the same reason, the part on financial instruments was divided into more chapters and significantly extended, especially with respect to FX and credit derivatives. Some additional day count conventions have been added and links to the (significantly changed and still changing) regulatory requirements have been updated.

The example excel spread sheets have also been updated and extended. These examples are no longer delivered on CD, but can be obtained via a web page accompanying this book [50].

Without the support of my colleagues and of d-fine this new edition would not have surfaced. Especially, I would like to thank Hans-Peter Deutsch, who always supported me with great enthusiasm during this project. Even though, it still took more than 5 years after the publication of the 5th edition in German that this English version of the 5th edition could be finalized. It turned out that the many updates and tiny differences accumulated over the years made it quite difficult to bring the English edition again in line with the new German edition.

Friedrichsdorf, Germany   Mark W. Beinker
March 2019

Preface to Fourth Edition

The philosophy of this book is to provide an introduction to the valuation and risk management of modern financial instruments formulated in precise (and mathematically correct) expressions, covering all pertinent topics with a consistent and exact notation and with a depth of detail sufficient to give the reader a truly sound understanding of the material. An understanding which even places the reader in a position to independently *develop* pricing and risk management algorithms (including actually writing computer programs), should this be necessary. Such tasks will greatly be facilitated by the website [50] accompanying the book. This website contains Microsoft Excel^{TM} workbooks presenting concrete realizations of the concepts discussed in the book in the form of executable algorithms. Of course, the reader has full access to all source codes of the Visual Basic^{TM} modules as well as to all calculations done in the spread sheet cells. The website thus contains a collection of literally thousands of examples providing the reader with valuable assistance in understanding the complex material and serving as the potential basis for the further development of the reader's own particular pricing and risk management procedures.

The book should equip the reader with a wide array of tools needed for all essential topics in the field of modern market risk management and derivatives pricing. The reader is not expected to have previous knowledge of finance, but rather a sound mathematical and analytical background typical of scientists, mathematicians, computer scientists, engineers, etc. The novice is not even required to be familiar with ideas as fundamental as compounding interest. The book, however, is certainly also of interest to the experienced risk manager or financial engineer, since the concepts introduced are widely elaborated upon and analyzed down to the very foundations, making a comprehension of the

material possible which goes significantly beyond the level held to be "common knowledge" in this field.

Since the beauty of a room behind a closed door is of little use if the door itself cannot be found, emphasis has been placed on providing an easy *entry* into the analysis of each of the various topics. As the author does not wish to lose the reader at the outset, or expect the reader to first engage in the study of quoted literature before proceeding, the book is practically self-contained. An explanation of almost every expression or notion needed can be found in the book itself, ranging from compounding interest to term structure models, from expectation to Value at Risk, from time series analysis to GARCH models, from arbitrage to differential equations and exotic options, from the normal distribution to martingales, and so on.

The selection of the topics and the nature of their presentation result to a great extent from my personal experience as a consultant in the world of financial services; first with the *Financial Risk Consulting* division of *Arthur Andersen* in Germany, which it has been my pleasure to establish and direct for many years, and later with the consulting firm *d-fine*, which is in fact this former Financial Risk Consulting division, now operating as a company of its own. In these functions, I have been in a position to observe and identify exactly what knowledge and methods are required in the financial world as well as to see what tools are indispensable for a newcomer to this world.

I would like to take this opportunity to thank many of the (in part former) members of the *Financial Risk Consulting* team and of *d-fine* for their valuable input and many fruitful discussions not only concerning this book, but also in our day to day consulting work.

Niedernhausen, Germany   Hans-Peter Deutsch
October 2008

Contents

List of Figures

List of Tables

Part I

Fundamentals

1

Introduction

The explosive development of derivative financial instruments continues to provide new possibilities and increasing flexibility to manage finance and risk in a way specifically tailored to the needs of individual investors or firms. This holds in particular for banks and financial services companies who deal primarily with financial products, but is also becoming increasingly important in other sectors as well. Active financial and risk management in corporate treasury can make a significant contribution to the stability and profitability of a company. For example, the terms (price, interest rate, etc.) of transactions to be concluded at a future date can be fixed today, if desired even giving the company the option of declining to go ahead with the transaction later on. These types of transactions obviously have some very attractive uses such as arranging a long-term fixed-rate credit agreement at a specified interest rate a year in advance of the actual transaction with the option to forgo the agreement if the anticipated need for money proves to have been unwarranted (this scenario is realized using what is known as a "payer swaption") or providing a safeguard against fluctuations in a foreign currency exchange rate by establishing a minimum rate of exchange today for changing foreign currency into euros at a future date (using a foreign currency option).

However, the complexity of today's financial instruments and markets and the growing pace of technological progress have led to an almost uncontrollable increase in the risks involved in trading and treasury while simultaneously reducing drastically the time available for decision making. Thus, the inappropriate use of financial instruments may quickly result in losses wiping out gains achieved over years in a company's primary business (for example, the production and sale of automobiles or computer chips).

In recent years we have seen an increase in the number of sizable losses incurred in consequence of derivative transactions which, in some cases, have resulted in bankruptcy. This phenomenon has not been restricted to banks but has involved companies in various other sectors as well. Spectacular examples include Metallgesellschaft (oil futures), Procter & Gamble (speculation with exotic "power options"), Orange County (interest rate derivatives, highly leveraged portfolio), Barings (very large, open index futures positions), Daiwa Bank (short-term US Bonds), NatWest Markets (incorrect valuation of options in consequence of incorrect volatility assumptions), the hedge fund LTCM, the Subprime-Crisis in the US 2007/2008 (ABS structures), Societe Generale (speculations of a single equity trader), etc., etc.

Not only have *financial instruments* become increasingly risky ("more volatile"), so have the *markets* themselves. Since the beginning of the 1980s we have seen a fundamental change to the economic framework in the financial world. In today's investment environment, yields, foreign currency exchange rates, commodity and stock prices can shift daily to an extent which would have been inconceivable in previous years. Increased market fluctuation (volatility) is the financial markets' reaction to developments such as the accumulation of capital, the globalization of financial markets, an increase in the budget deficits of the leading industrialized nations and the dismantling of government regulations, to name just a few.

The main prerequisite for the continued success of a bank or treasury department in an environment of highly volatile markets and extremely complex (and thus very sensitive) financial instruments is a sound understanding of the products being traded and efficient management of the risk involved in these transactions. Derivatives are therefore the main *reason for* and the most effective *means of* conducting risk management, and thus can be viewed as the be-all and end-all of risk management.

The *pricing* of derivatives and structured financial instruments is the foundation, the conditio sine qua non, of all risk management. To this end, all derivatives under consideration, regardless of their complexity, are broken down into their basic components and systematically categorized as spot, futures and option transactions. This process is called *stripping*. Understanding

the fundamental instruments and their valuation, as well as "stripping" poses a serious challenge for those involved, requiring skills which in the past were not called for to any extent in the financial world, such as the following

- Theoretical demands: statistics, probability theory, time series analysis, etc., essential to the understanding of concepts such as "value at risk" and the estimation of underlyings and their correlation; differential equations, martingale theory, numerical analysis, financial mathematics, etc., for pricing financial instruments.
- Trading demands: understanding the function of increasingly complicated securities and transactions (e.g., by stripping); making "imaginative" choices of derivatives or constructing new instruments to realize a particular strategy while taking into consideration the often surprising "side effects"; developing complicated hedging strategies; operating highly complex computer systems.
- IT demands: client-server computing, distributed systems, object-oriented programming, new operating systems, network engineering, real time information distribution, Internet and intranet architecture and increasingly support for parallel algorithms on computer grids, multi-core CPUs and graphic processors (GPUs).

The theoretical and trading aspects in particular pervade all business hierarchies, since it is equally important for the trader/ treasurer to be able to understand the risk involved in his or her portfolio as it is for a managing director or the board of directors whose decisions are based on this information.

Thus, the aim of this book is to provide a detailed presentation of the methods and procedures used in pricing financial instruments and identifying, quantifying and controlling risk. The book is structured as follows:

Part I introduces the fundamental market parameters which govern the price and risk of financial instruments. These include the price of stocks, commodities, currencies, etc. and naturally interest rates. A stochastic model, known as the *random walk* and its generalization, the *Ito process*, will be used to model these fundamental risks, also known as *risk factors*, and will serve to describe both the deterministic and the random aspects of the risk factors. In doing so, an additional parameter, called *volatility* will, similar to the above mentioned risk factors, play an important role in the price and risk accorded to a financial instrument.

Finally, in Part I, the most common instruments for trading the risk factors are introduced. *Trading* is defined as the acceptance of risk or risk

factors, through a financial instrument, in return for a certain yield from the instrument and/or specified payments from a contracting party. Conversely, a financial instrument can of course also be used to transfer risk to a contracting party in exchange for paying a specified yield and/or amount to the risk-taking party. In Part I, the *functionality* of several of the most important financial instruments will be introduced without entering into a discussion of their *valuation*, which is frequently very complicated.

Once the fundamental risks and the financial instruments used for trading them have been introduced, the building blocks for further analysis will have been defined. This will allow us to proceed with the actual topic of this book, namely determining the *prices* and the *risks* of these building blocks. For this, the same fundamental *methods* will be repeatedly applied to many different financial instruments. In order to make this clear, Part II introduces the most important methods for pricing and hedging in their full generality—independent of any specific instrument. As a result, this section is rather theoretical and technical. The concepts elaborated upon in Part II will then be applied to specific financial instruments in Part III. This separate treatment of general methods and specific financial instruments will contribute to a clearer understanding of this rather complex material.

Once the pricing of the most common financial instruments has been dealt with in Part III, the determination of the risks associated with these instruments will be presented in Part IV. The information about the prices and risks of financial instruments can then be used for decision making, specifically of course for trading decisions and the management of investment portfolios. This is demonstrated in Part V. Finally, methods for determining and analyzing the market data and historical time series of market risk factors will be the topic of Part VI.

2

Fundamental Risk Factors of Financial Markets

The fundamental risk factors in financial markets are the market parameters which determine the price of the financial instruments being traded. They include foreign currency exchange rates and the price of commodities and stocks and, of course, interest rates. Fluctuations in these fundamental risks induce fluctuations in the prices of the financial instruments which they underlie. They constitute an inherent market risk in the financial instruments and are therefore referred to as *risk factors*. The risk factors of a financial instrument are the market parameters (interest rates, foreign currency exchange rates, commodity and stock prices), which, through their fluctuation, produce a change in the price of the financial instrument. The above mentioned risk factors do not exhaust the list of the possible risk factors associated with a financial instrument nor do all risk factors affect the price of each instrument; for example, the value of a 5 year coupon bond in Swiss Francs is not determined by the current market price of gold. The first step in risk management is thus to identify the relevant risk factors of a specified financial instrument.

2.1 Interest Rates

Various different *conventions* are used in the markets to calculate interest payments. For example, interest rates on securities sold in the US money markets (*T-bills*, *T-bill futures*) are computed using *linear compounding*, whereas in the European money market, *simple compounding* is used. Interest rates in the

capital markets are calculated using *discrete* or *annual* compounding while option prices are determined using the *continuous* compounding convention. While these conventions are not essential for a principle understanding of financial instruments and risk management they are of central importance for the *implementation* of any pricing, trading or risk management system. Before entering into a general discussion of interest rates in Sect. 2.1.5, we will therefore introduce the most important compounding conventions here.

2.1.1 Day Count Conventions

Before one of the many compounding conventions are applied to calculate the interest on a certain amount over a period between the time (date) t and a later time T, the number of days between t and T over which interest is accrued must first be counted. The beginning, the end and the length of this time period (measured in years) must be precisely specified. To do this there are again different conventions used in different markets, known as *day count conventions*, *DCC* for short.

Therefore, a time difference $T - t$ measured in years depends in general on the chosen day count convention. To make this explicit, we will often use the expression $\tau(t, T)$ (year fraction) for the time period from time t to T measured in years with a given day count convention. These market usances are usually specified by making use of a forward slash notation: the method for counting the days of the month are specified in front of the slash, the number of days of the year after the slash. A list of the most commonly used conventions is presented in Table 2.1. These conventions compute the length of an interest rate period as follows:

- *Actual/365f:* The actual number of calendar days between t and T are counted and divided by 365 to obtain the interest period in years regardless

Table 2.1 The commonly used day count conventions

Common designation	Alternative designation
Act/365f	Act/365(f), Act/365(fixed), Act/365fixed
Act/360	
30/360	BondBasis, Bond, 30
30E/360	EuroBondBasis, EuroBond, 30E
Act/Act	Act/Act ISDA, Act, Act/365(l), Act/365leap
ICMA	Act/Act, Act/Act ICMA, Act/Act ISMA
BD/252	Brasilian

of whether the year concerned is a leap year. This distinguishes the Actual/365f convention from the Act/Act in years.

- *Actual/360*: The actual number of calendar days between t and T are counted and divided by 360 to obtain the interest period in years.
- *30/360:* The days are counted as if there were exactly 30 days in each month and exactly 360 days in every year. In addition, the following holds:
  - If the *beginning* of the interest period falls on the 31st of the month, the beginning is moved from the 31st to the 30th of the same month for the purpose of the calculation.
  - If the *end* of the interest period falls on the 31st of the month, it is moved forward for the purposes of the calculation to the 1st of the next month **unless** the *beginning* of the interest period falls on the 30th or 31st of a month. In this case, the end of the interest period is moved from the 31st to the 30th of the same month.

- *30E/360:* The days are counted as if there were exactly 30 days in each month and exactly 360 days in each year. In addition, the following holds:
  - If the *beginning* of the interest period falls on the 31st of the month, the beginning is moved from the 31st to the 30th of the same month for the purpose of the calculation.
  - If the *end* of the interest period falls on the 31st of the month, the end of the period is moved to the 30th of the same month for the purpose of the calculation (this differentiates this day count convention from the 30/360 convention).

- *Act/Act*: The actual number of calendar days are counted and divided by the actual number of days in the year, i.e. parts of the time period falling in leap years are divided by 366, otherwise the divisor is 365. Often, the same notion is used for other day count conventions as well, so you need to be cautious.
- *ICMA* is a specific convention proposed by the International Capital Market Association, which is frequently used as standard convention for bond emissions. The number of days equals the actual number of days. The calculation of the divisor is based on the coupon payment frequency m (i.e. number of coupon payments per year) of the bond and the rolling day D_R, which is the start date for the roll out of all coupon periods, which are subsequently adjusted due to business day rules (see Sect. 2.1.2). Next, the natural length L (i.e. without any further adjustments) of the considered time period is calculated. The divisor is set to the product mL of natural length L and coupon frequency m. This method guarantees that

any coupon period, which start and end days fall together with the rolled out and according to business day rules adjusted dates, has exactly the length $L/12$ in years. Only periods, which have been shortened or lengthened by some days may differ.

- *BD/252* counts only the business days, divided by the number of business days per year, which is fixed to 252.

These conventions for calculating the time between two dates can also be expressed in formulas:

$$\text{Act/365f} : \frac{D_2 - D_1}{365}$$

$$\text{Act/360} : \frac{D_2 - D_1}{360}$$

$$\text{30/360} : J_2 - J_1 + \frac{M_2 - M_1}{12} \tag{2.1}$$

$$+ \frac{T_2 - \min(T_1, 30) - \max(T_2 - 30, 0) * \text{feb}(T_1 - 29)}{360}$$

$$\text{30}E\text{/360} : J_2 - J_1 + \frac{M_2 - M_1}{12} + \frac{\min(T_2, 30) - \min(T_1, 30)}{360}$$

$$\text{Act/Act} : J_2 - J_1 + \frac{D_2 - \text{Date}(J_2, 1, 1)}{\text{Date}(J_2 + 1, 1, 1) - \text{Date}(J_2, 1, 1)}$$

$$- \frac{D_1 - \text{Date}(1, 1, J_1)}{\text{Date}(J_1 + 1, 1, 1) - \text{Date}(J_1, 1, 1)}$$

$$\text{ICMA} : \frac{D_2 - D_1}{m \left(\text{Adjusted}(D_R + m/12 \text{ months}) - \text{Adjusted}(D_R) \right)}$$

$$\text{BD/252} : \frac{\text{business days between } D_1 \text{ and } D_2}{252} \, .$$

The notation should be read as follows: D_1 denotes the start date of the interest period, D_2 the end date. D_1 consists of the number T_1 for the days, M_1 for the months and J_1 for the years; D_2 is defined analogously. m is the number of coupon payments per year and D_R the rolling date, which is the start date for the roll out of all coupon start and end dates. The function "Date" delivers the running number for each given date. The counting of this running number begins at some time in the distant past, for example January 1, 1900. Then the

date function (as defined in Microsoft Excel, for instance; please note that for compatibility reasons Microsoft Excel counts the (actually non-existing) date February 29th, 1900 by default also as a valid date) yields the value 35,728 for Date (1997, 10, 25). This is the running number corresponding to October 25, 1997. The function "feb(x)" is defined as equal to zero for $x \leq 0$, and 1 otherwise. The function $\min(x, y)$ yields the smaller of the two values x and y, and the function $\max(x, y)$ yields the larger of the two. In general, the difference $D_2 - D_1$ is actually the number of days between D_1 and D_2, including D_1 but excluding D_2. The single exception is the last premium period of a *Credit Default Swap* (*CDS*), which includes both, D_1 and D_2 and is thus one day longer than, e.g., the equivalent interest rate period. The function Adjusted(D) adjusts the date D according to a given business day rule (see Sect. 2.1.2) to fall on a business day.

The computation of time periods using these day count conventions is demonstrated in Table 2.2 and in the Excel-Sheet Usance.xls available in the download section [50] accompanying this book. The time period has been chosen to yield different results for different conventions. The period starts at February 15th, 2012 and ends on December 31st, 2012. The chosen business day calendar is TARGET2 (the official calendar for the EUR zone) with February 15th, 2012 as roll date and period length of one month. The time period has been intentionally selected to contain the 29th of February in order to demonstrate the difference between the *Act/Act* and *Act/365f* conventions. Furthermore, the interest period was chosen to end on the 31st of the month to illustrate the difference between the *30/360* and *30/E360* conventions.

Table 2.2 Determining the length of a time period using different day count conventions. The time period was chosen to yield a different length (in years) in each day count convention

| | Time period | |
	Days	Years
Act/365f	320	0.87671233
Act/360	320	0.88888889
30/360	316	0.87777778
30E/360	315	0.875
Act/Act	320	0.87431694
ICMA	320	0.87356322
BD/252	223	0.88492063

2.1.2 Business Day Conventions

Establishing a day count convention is not sufficient to uniquely determine interest periods. The *value dates* of the cash flows must also be defined, i.e., the number of days following the end of an interest period T on which the interest payment must be settled. Several different conventions govern this calculation as well. Furthermore, there are conventions to account for weekends and holidays. If the value date falls on a *bank holiday*, for example, should payment be made on the day before or after the holiday? And finally, bank holidays themselves vary from country to country. The conventions governing these questions are called *business day conventions*, *BDC* for short. These conventions transfer the value dates of a cash flow away from weekends and bank holidays in accordance with the rules compiled in Table 2.3.

Table 2.4 shows as an example the adjustment of March 31st, 2012 and January 1st, 2012. These example dates fall on either a weekend or bank holiday (according to TARGET2 calendar). For the EUR zone, *TARGET2* is the official bank holiday calendar. TARGET2 is the short form of Trans-European Automated Real-Time Gross Settlement Express Transfer System (2nd version), the settlement system for money transfers between banks operating in the EUR zone. TARGET2 replaced on May 19th, 2008 the preceding system (or collection of systems) TARGET. The TARGET2 holidays are January 1st, Good Friday, Easter Monday, May 1st, and December 25th and 26th. All other days, except for Saturdays and Sundays, are business days.

Table 2.3 Business day conventions

Convention	Adjustment to
Following	Following business day
Modified following	Following business day as long as this falls within the same month; otherwise the preceding business day
Preceding	Preceding business day
Modified preceding	Preceding business day as long as it falls within the same month; otherwise the following business day (used rarely or never)

Table 2.4 Effects of business day conventions

Convention	Adjustment March 31st, 2012	Adjustment January 1st, 2012
Following	April 2nd, 2012	January 2nd, 2012
Modified following	March 30th, 2012	January 2nd, 2012
Preceding	March 30th, 2012	December 30th, 2011
Modified preceding	March 30th, 2012	January 2nd, 2012

For interest rate instruments, a further distinction is made between whether the adjustment convention holds solely for the *payment* date of an interest period or for its *maturity* date as well. The *maturity* date determines the length of the interest rate period and thus affects the *amount* of the interest payment (if the length of an interest period is longer, the amount of the interest payment to be made is naturally higher). The *payment* date determines *when* the interest payments are actually made (usually one or two business days after the maturity date) and therefore affects how strongly a payment is discounted, in other words, today's *value* of the payment (a later payment is obviously worth less than an earlier one); it is thus relevant when the payment is actually made and not when it was due.

If the maturity date of a financial instrument is specified as *fixed*, i.e., non-moveable, it is not adjusted. However, the *payment* date is still adjusted in accordance with the business day convention for the instrument concerned. The *rollover day* of an interest rate instrument determines on which day and month of each year the *rollover* from one interest period into the next is to take place, i.e., when the maturity and payment dates of individual interest payments occur. Depending on the selected business day convention, a decision is made as to how the maturity and payment dates, derived from the rollover date, are to be adjusted. For example, federal bonds are commonly agreed to be *fixed.* This means that only the *payment* date is adjusted to the next valid business day while the *maturity* date is always the same as the rollover day. For swaps, on the other hand, both the payment and maturity dates are adjusted to the next valid business day subsequent to the rollover day.

All these conventions make trading substantially more complicated without causing a fundamental change in the properties of the instruments being traded. They are actually unnecessary but the markets are inconceivable without them because of strong historical ties. This is particularly true of holiday calendars, some of which even have religious roots.

2.1.3 Discount Factors

In order to concentrate on the essentials, a general notation will be observed in this book which holds for *all* compounding, day count, business day and other market conventions. To accomplish this, discount factors rather than interest rates will be employed throughout. The *discount factor* is the value by which a *cash flow* to be paid at a time T is multiplied in order to obtain the value of the cash flow at an earlier time t (for example, today). In general, an amount of money today is worth more than the same amount of money which

is accessible not before a future date T, because of the missing choice to either spend the money, invest it in a profit-gearing asset, or just leave it alone until T. This option has a positive value, the so-called *time value of money*.[1] Thus, the discount factor is generally less than 1 (but greater than 0). A discount factor for discounting from a time T back to an earlier time t will be referred to using the notation

$$B(t, T) \tag{2.2}$$

The letter B is used since a discount factor is identical to the price at time t of a zero bond with maturity T and nominal of one currency unit (i.e. 1 EUR).

The discount factor yields the value of a *future* payment *today* (*discounting*). Conversely, the *future value* of a *payment today* (*compounding*) is obtained by multiplying the payment by $B^{-1}(t, T)$. The *interest* accrued between times t and T is thus the difference between the compounded value and the original amount, i.e., the original amount multiplied by the factor $(B^{-1}(t, T) - 1)$.

2.1.4 Compounding Methods

Over time, various methods for calculating interest have been established. The explicit form of the discounting and compounding factors and of the interest accrued (based on a notional=1) are shown for the four most common *compounding methods* in Table 2.5. At each stage in this book, the results expressed in the general notation given in the first line of Table 2.5 can be converted directly into the explicit expressions of the desired compounding method by replacing the general expression with the appropriate entries in Table 2.5, given the fact that discount factors offer a convention-independent way to express these factors.

These discount factors are obtained from intuitive considerations which are now described in detail for each compounding method. Note that interest rates quotes are always relate to a specific *time unit*. Typically, interest rates are given as a percentage per year (i.e. per annum), quoting interest rates per month or even per day is the exception. The actual length of the time period over which interest is paid, and which has to be measured in terms of day count convention consistent with the quoted interest rate, is independent from the above mentioned time unit. For example, an interest rate quote of 6% is only

---

[1]In case of negative interest rate, the time value may become negative, since negative interest rates can be viewed as storage costs.

Table 2.5 Interest rate factors in general notation and their specific form for the three most commonly used compounding methods. Linear compounding is an exception, since it should not be seen as an additional compounding method rather than an approximation for short time periods of the simple compounding method and is only mentioned for sake of completeness. Calculation of interest requires always the specification of a compounding method, therefore, it is not possible to calculate accrued interest if only the discount factor is given

	Discount factor	Compounding factor	Interest accrued
General	$B(t, T)$	$B^{-1}(t, T)$	Not defined
Continuous	$e^{-R(T-t)}$	$e^{R(T-t)}$	$e^{R(T-t)} - 1$
Discrete	$(1 + \frac{R}{m})^{-m(T-t)}$	$(1 + \frac{R}{m})^{m(T-t)}$	$(1 + \frac{R}{m})^{m(T-t)} - 1$
Simple	$[1 + R(T-t)]^{-1}$	$1 + R(T-t)$	$R(T-t)$
Linear	$1 - R(T-t)$	$1 + R(T-t)$	$R(T-t)$

complete if the related time unit (e.g. per annum, if nothing else is given) and day count convention (e.g. *act/365*) is also specified.

Simple Compounding

For *simple compounding* the interest paid at the end of the term agreed upon is calculated by simply applying the rule of three: if an interest rate R per time unit has been agreed upon and the interest period $(T - t)$ spans n time units, the interest payable is simply the product n multiplied by R. If the period $(T - t)$ is measured in the same *unit* (e.g., years) as that used to quote the interest rate (e.g., interest rate is quoted per annum) then we simply have $n = (T-t)$ and the interest paid simply equals $R(T-t)$ as indicated in Table 2.5. If capital in the amount K_0 has been invested, the interest earned resulting from simple compounding is obtained by multiplying K_0 by $R(T-t)$. The capital held at maturity from an investment paying, for example, 6% per annum over a quarter of a year is thus

$$K = K_0(1 + 6\frac{\%}{\text{year}} * 0.25\text{year}) = K_0(1 + 6\% * 0.25) = K_0[1 + R(T-t)] \ .$$

The compounding factor is thus $1 + R(T - t)$. The discount factor is the reciprocal of the compounding factor. Simple compounding is often used in money markets where interest periods $(T - t)$ are usually less than one year or for fixed income instruments with periodic interest rate payments (without compounding interest rate over more than on period).

Discrete Compounding

In contrast to simple compounding, *discrete compounding* takes compounding effects into account. It is used for long term investments (longer than one interest period), if during the investment period no interest is actually paid out. Therefore, the length of the interest period or the frequency of assumed interest periods per year is required as an additional parameter. Typical interest periods are annual, semi-annual, quarterly, monthly, or daily, though again the interest rate quote refers to interest paid per annum, with the full investment period $T - t$ given in fraction of years. In the following, m denotes the number of interest periods per year. Then, interest is calculated as if after each interest period the interest is immediately re-invested at the same interest rate. Or, to put it otherwise, the invested capital is increased by the interest due. The assumed interest amount for each period is calculated by means of *simple* compounding. An interest rate of 5% with $m = 2$ (semi-annual compounding) would thus yield after half a year (one interest period) an invested capital K_1 of

$$K_1 = K_0 \left(1 + \frac{5\%}{2}\right)^{2 \times 0,5} = K_0(1 + 2,5\%) = K_0(1 + R/2) \ .$$

This capital is reinvested over the next time unit (the second year) in exactly the same manner. Thus, after one year the investor's capital has increased to

$$K_2 = K_1(1 + R/2) = K_0(1 + R/2)(1 + R/2) = K_0(1 + R/2)^2 \ .$$

In the second period, interest is accrued on the interest earned in the first period in addition to the initial investment capital. This is called *compounded interest*. Likewise, after the third period, the investor's capital is given by

$$K_3 = K_2(1 + R/2) = K_0(1 + R/2)^3$$

and so on. The compounding factor thus obtained is $(1 + R/m)^{mn}$ where n is the length of the full investment period in years and m the number of interest periods per year, or, more general, the compounding factor is equal to $(1 + R/m)^{m(T-t)}$ as in Table 2.5. Here, $m(T - t)$ is not necessarily a whole number. The calculation remains the same for $T - t = 3.5$, for example.

Continuous Compounding

In the case of *continuous compounding*, the calculation is performed as if interest payments were made after each *infinitesimal* small time increment (each payment calculated using simple compounding) with the accumulated interest being immediately reinvested at the same rate. Therefore, it can be considered as the limit of the discrete compounding method for infinitesimal small interest periods. This method is frequently used for modeling derivatives, because of its convenient mathematical features, easing the life of financial mathematicians significantly. Since different compounding methods can be easily converted into each other, you have the free choice of using the most convenient method for doing your math. However, the interest rates for real interest coupons paid by fixed income instruments are never quoted as continuous compounding rates.

The total capital accumulated on an investment over a time period of $T - t$ calculated by this method is then

$$K_{(T-t)} = \lim_{m \to \infty} K_0 \left(1 + \frac{1}{m} R \right)^{(T-t)m} = K_0 e^{R(T-t)} . \qquad (2.3)$$

The compounding factor is thus given by $e^{R(T-t)}$ as indicated in Table 2.5. Here, *Euler's number e*, also called the *natural number*, arises. Its value is approximately

$$e \approx 2,718281828459\ldots$$

Euler's number to the power of some number x is called the *exponential function*, which is defined by means of the limit in Eq. 2.3:

$$\exp(x) := e^x := \lim_{m \to \infty} \left(1 + \frac{x}{m} \right)^m .$$

Linear Compounding

Linear compounding is justified for very short periods of time $T - t$. For such times, the product $R(T - t)$ is also very small. For example, if $R = 3\%$ per annum and the time to maturity $T - t$ is one month $= 0.083$ years, the product $R(T - t) = 0.0025$. The square of this product is considerably smaller, namely $R(T - t)^2 = 0.00000625$. Thus, in the case of linear compounding,

only terms of order $R(T-t)$ are of any importance, i.e., all non-linear terms are simply neglected. If we represent the discount factor, which is always the inverse of the compounding factor, as a geometric series[2] neglecting all terms of higher order, we obtain the discount factor given in Table 2.5:

$$[1 + R(T-t)]^{-1} \approx \underbrace{1 - R(T-t)}_{\text{linear terms}} + \underbrace{(R(T-t))^2 \pm \cdots}_{\text{higher order terms are neglected!}} \quad .$$

It follows, in strict linear approximation, that the general rule that the reciprocal is identical to the inverse discount factor holds also for linear compounding. However, since this condition is only approximatively true, linear compounding is inferior to other compounding methods and should be used only as an approximation for simple compounding. Since the advent of computers, it's practical importance has diminished. Therefore, we won't use this method within the rest of the book.

Convention-Dependent Interest Rates

Today's value of a future cash flow is determined by the discount factor for the proper time period. On the other hand, the discount factor could easily be calculated as today's value of the cash flow divided by its amount (or its value at pay date). However, expressing discount factors in terms of interest rates requires the application of compounding methods and day count conventions, which is demonstrated by the very different formulas for discount factors in Table 2.5. After all, today's value of a monetary unit paid in the future must be independent of the convention used for discounting.

Other than discount factors, *interest rates* depend on the methods and conventions used. The impact of conventions is fully "absorbed" in the interest rates. As a useful consequence, equalizing discount factors expressed in different compounding methods as shown in Table 2.5 enables the transformation of a given interest rate from one compounding convention into another:

$$e^{-R_{\text{continuous}}(T-t)} = \frac{1}{(1 + R_{\text{discrete}}/m)^{m(T-t)}} = \frac{1}{1 + R_{\text{simple}}(T-t)} \quad .$$

---

[2]The expansion used here is $(1+x)^{-1} = 1 - x + x^2 - x^3 + x^4 - x^5 \pm \cdots$. The result is now obtained by substituting $R(T-t)$ for x.

For example, the interest rate necessary to generate a discount factor in discrete discounting (with $m = 1$, i.e. annual compounding) with the same numerical value as a given discount factor in continuous compounding is

$$R_{\text{annual}} = e^{R_{\text{continuous}}} - 1, \quad R_{\text{continuous}} = \ln(1 + R_{\text{annual}}) \ .$$

If for discrete compounding the interest period is identical to the compounding period, the relation is very simple, i.e. if $\frac{1}{m} = T - t$ we have:

$$\frac{1}{\left(1 + \frac{R_{\text{discrete}}}{m}\right)^{m(T-t)}} = \frac{1}{1 + R_{\text{discrete}}(T-t)} = B(t, T) \ ,$$

or

$$R(T - t) = B^{-1}(t, T) - 1 \ . \tag{2.4}$$

In this special case, the expression for the discount factor is identical to the case of simple compounding. This relation will be used frequently throughout the book.

However, not only the effects of the compounding methods but also the effects of the day count convention are absorbed in the interest rates. This is demonstrated in the Excel-sheet USANCE.XLSX available from the download area and in Table 2.6. The interest period is the same as in Table 2.2, but the different day count conventions generate different time lengths (measured in years). The discount factor must be the same for all conventions. The interest rates associated with this single discount factor, however, are strongly

Table 2.6 Interest rates for the same discount factor based on different day count and compounding conventions. The interest period is the same as in Table 2.2. The value of the discount factor for this period is $0,98261858$

	Zeitperiode		Zero-Raten [in %]			
	Tage	Jahre	Linear	Einfach	Diskret	Contin.
Act/365f	320	0.87671233	1.983	2.018	2.020	2.000
Act/360	320	0.88888889	1.955	1.990	1.992	1.973
30/360	316	0.87777778	1.980	2.015	2.018	1.998
30/E360	315	0.875	1.986	2.022	2.024	2.004
Act/Act	320	0.87431694	1.988	2.023	2.026	2.005
ICMA	320	0.87356322	1.990	2.025	2.027	2.007
BD/252	223	0.88492063	1.964	1.999	2.001	1.981

influenced by both the day count convention as well as the compounding convention. They vary between 1.955% and 2.027%.

As already mentioned, all of the conventions introduced here are actually unnecessary for understanding financial instruments. In order to concentrate on the essentials, discount factors rather than interest rates will be predominantly used in this book. For any concrete calculation, the reader should be able

- to write down explicitly the required discounting factor using Table 2.5 and
- to calculate the time length $T - t$ using Eq. 2.1 (along with the Excel-sheet Usance.xlsx)
- after the exact dates t and T in the appropriate business day convention as specified in Table 2.3 have been determined.

In what follows, we will therefore work only with the general discount factor given in Eq. 2.2.

2.1.5 Spot Rates

Spot rates are the *current* yields on securities which generate only one single payment (cash flow) upon maturity. A zero coupon bond is an example of such a security as are coupon bonds whose last coupon payment prior to maturity has already been made. The spot rates as a function of time to maturity T is called the *spot rate curve* or the *term structure*. These spot rate curves can be represented by the discount factors $B_R(t, T)$.

2.1.6 Forward Rates

Forward rates are the *future* spot rates from today's point of view which are consistent with the current spot rates. The pay-off profile of many interest rate products is based on a forward interest rate $F(t, t_{\text{fix}}, T, T')$, which is fixed at a future time t_{fix} and which relates to the future interest period from T to T' with $t_{\text{fix}} \leq T < T'$. Of course, from today's point of view with $t < t_{\text{fix}}$, the actual value of this interest rate is yet unknown.

The forward rate as of today can be derived from the following arbitrage argument: One monetary unit is to be invested today (time t) until a specified

Fig. 2.1   The sequence of the times t, T and T'

maturity date T'. The investor can consider the following two investment strategies:

1. Invest the monetary unit without adjusting the position until the maturity date T', or
2. Invest the monetary unit until some time T where $T < T'$, and at time T immediately reinvest the interest earned together with the original amount until the maturity date T'.

If the investor were able to fix *today* the interest rate for the future period between T and T' (see Fig. 2.1), then the *total return* for both strategies must be the same. If this were not the case, the investor would have an opportunity to earn a profit without risk, i.e., an *arbitrage* opportunity (see Sect. 6.1). If the yield from the first strategy were higher, an investor could raise capital according to the second strategy and invest this capital according to the first. Financing the investment using the second strategy requires less interest than the total return on investing in the first. Thus, since all interest rates are fixed at time t, a profit would have been made without risk or investment capital at time t. Conversely, if the total return of the first strategy is lower than that of the second, the investor could finance the second investment strategy according to the terms of the first. Again, this gives the investor an arbitrage opportunity.

The interest rate fixed at time t for the future time period between T and T' which eliminates any possible arbitrage opportunity, i.e., which is consistent with the spot rates at time t, is the forward rate $R(T, T'|t)$. The notation for the analogous *forward discount factor*, i.e. the price of a zero bond at the future time T which pays one monetary unit at $T' > T$ is

$$B(T, T'|t) . \tag{2.5}$$

The condition eliminating the arbitrage opportunity can be summarized as follows: the compounding factor using the *spot* rate from t to T' must be equal to the compounding factor of the spot rate from t to T multiplied by the compounding factor of the forward rate from T to T'. The compounding

factors are the inverse of the corresponding discount factors as indicated in Table 2.5. Consequently, the following equality must hold to prevent arbitrage:

$$\underbrace{B^{-1}(t,T')}_{\text{Spot}} = \underbrace{B^{-1}(t,T)}_{\text{Spot}}\underbrace{B^{-1}(T,T'|t)}_{\text{Forward}}$$

or equivalently,

$$\underbrace{B(t,T')}_{\text{Spot}} = \underbrace{B(t,T)}_{\text{Spot}}\underbrace{B(T,T'|t)}_{\text{Forward}} \quad \forall\, t \le T \le T' \,. \tag{2.6}$$

The forward discount factors can thus be uniquely determined from today's discount factors:

$$B(T,T'|t) = \frac{B(t,T')}{B(t,T)} \qquad \forall\, t \le T \le T' \,. \tag{2.7}$$

The last two equations are quite fundamental. They will be used repeatedly in what follows to simplify formulas involving terms with products and quotients of discount factors. With the aid of these equations, many essential properties of financial instruments depending on interest rates can be derived without ever having to specify the compounding convention. This helps to clarify questions as to whether certain properties under consideration are intrinsic properties of the instruments themselves or merely results of the compounding convention being used. For example, the difference between *Macaulay* Duration and *Modified* Duration (see Chap. 5) is not an inherent property of bonds but merely an effect resulting from applying a particular compounding convention. In the case of continuous compounding, for example, there is no difference between the two!

At this point we provide an example demonstrating the usefulness of Eq. 2.6 and Table 2.5. We give explicitly the forward rates as a function of the spot rates for each of the three common compounding conventions:

$$R(T,T'|t) = \begin{cases} \dfrac{R(t,T')(T'-t) - R(t,T)(T-t)}{T'-T} & \text{Continuous Compounding} \\[3mm] \dfrac{[1+R(t,T')]^{\frac{T'-t}{T'-T}}}{[1+R(t,T)]^{\frac{T-t}{T'-T}}} - 1 & \text{Discrete Compounding} \\[3mm] \left[\dfrac{1+R(t,T')(T'-t)}{1+R(t,T)(T-t)} - 1\right]/(T'-T) & \text{Simple Compounding} \end{cases} \quad .$$

term	spot rate	start	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
																	forward rates
1	2.20%																
2	2.50%	1	2.80%	3.25%	3.67%	4.08%	4.36%	4.65%	4.83%	5.02%	5.20%	5.28%	5.37%	5.45%	5.49%	5.52%	5.54%
3	2.90%	2	3.70%	4.11%	4.51%	4.76%	5.03%	5.18%	5.34%	5.51%	5.56%	5.63%	5.70%	5.72%	5.74%	5.74%	
4	3.30%	3	4.51%	4.91%	5.11%	5.36%	5.47%	5.61%	5.77%	5.80%	5.84%	5.90%	5.90%	5.91%	5.90%		
5	3.70%	4	5.32%	5.41%	5.65%	5.71%	5.83%	5.98%	5.98%	6.01%	6.06%	6.04%	6.04%	6.02%			
6	4.00%	5	5.51%	5.82%	5.85%	5.96%	6.11%	6.10%	6.11%	6.15%	6.12%	6.11%	6.08%				
7	4.30%	6	6.12%	6.01%	6.11%	6.26%	6.21%	6.21%	6.24%	6.20%	6.18%	6.14%					
8	4.50%	7	5.91%	6.11%	6.31%	6.24%	6.23%	6.26%	6.21%	6.18%	6.14%						
9	4.70%	8	6.31%	6.52%	6.35%	6.31%	6.33%	6.26%	6.22%	6.17%							
10	4.90%	9	6.72%	6.36%	6.31%	6.33%	6.25%	6.21%	6.15%								
11	5.00%	10	6.01%	6.11%	6.21%	6.13%	6.10%	6.05%									
12	5.10%	11	6.21%	6.31%	6.17%	6.13%	6.06%										
13	5.20%	12	6.41%	6.15%	6.10%	6.02%											
14	5.25%	13	5.90%	5.95%	5.90%												
15	5.30%	14	6.00%	5.89%													
16	5.33%	15	5.78%														

Fig. 2.2 Forward rates for periods starting in $T = 1, 2, \ldots, 15$ years for terms $T' - T = 1, 2, \ldots, 15$ years. From line to line the start points T of the forward periods change by one year. From column to column the lengths $T' - T$ of the forward periods change by one year. The inset graphic shows, from top to bottom, the current term structure along with the forward term structures in 1, 5 and 10 years

This clearly demonstrates the advantage of using the general notation for discount factors 2.7 introduced above. In Fig. 2.2, the forward rates are calculated from the spot rates taken from the Excel sheet PLAINVANILLA.XLS (see examples on accompanying website [50]) using the above formulas for annual, discrete compounding over a period of 15 years.

Since coupon payments are generally calculated by means of the simple compounding methods, we demonstrate here explicitly how $F(t, t_{\text{fix}}, T, T')$ depends on the (forward) discount factor:

$$F(t, t_{\text{fix}}, T, T') = \frac{1}{\tau(T, T')}(B^{-1}(T, T'|t) - 1) . \qquad (2.8)$$

The forward rate calculated in this way is the *risk neutral* or *risk free*[3] forward rate, which can be calculated by means of risk neutral discount factors $B(t, T)$ only. For actually traded forward rates, an additional *tenor basis* has to be taken into account, which is an add-on of a few basis points to the risk neutral forward rate. A major cause for the tenor basis is that unsecured interest payments bear credit default risk which is the higher the longer the term to payment date is. The tenor basis could be specified as an explicit add-on (the

---

[3]"risk free" in the sense of free from credit default risk.

tenor basis spread) or by replacing $B(t, T)$ by a *forward yield curve* or *tenor curve* (see Sect. 15.2 also).

2.2   Market Prices

Let $S(t)$ be the *spot price* at time t of a stock, a commodity, or a currency. The *dividend-adjusted* spot price $\widetilde{S}(t, T)$ at time t is the price net of the value of all dividends paid between the times t and T. It is given by

$$
\widetilde{S}(t, T) = \begin{cases} S(t) & \text{no dividend} \\ S(t) - D(t')B(t, t') & \text{dividend payment } D \text{ due at time } t' \\ S(t)e^{-q(T-t)} & \text{dividend yield } q \\ S(t) - D(t')S(t, t') & \text{rel. discrete dividend } D(t') \text{ due at time } t' \end{cases} \quad .
$$

(2.9)

The *dividend adjustment* is thus accomplished by subtracting the value of dividends, discounted back to time t at the spot rate R, from the spot price, or—in the case of a dividend yield q—discounting the spot price from T back to t using the dividend yield q. In the following, the notation $B_q(t, T) = \exp[-q\tau(t, T)]$ might be used as an alternative way to express the dividend yield, making use of the formal similarity to discount factors. For example, currencies are mathematically equivalent to stocks with dividend yields, where q represents the risk-free interest rate (in continuous compounding) in the foreign currency. Similar, for commodities q is replaced by the difference between the *convenience yield* and the *cost of carry* (expressed as a yield, see later). Relative or proportional discrete dividends are proportional to the underlying price at the dividend payment date. The dividend cash flow therefore depends on the forward price $S(t, t')$.

When considering stock indices, the dividend payments from the assets in the index are commonly averaged out to result in a dividend yield q of the index rather than considering each dividend of each stock as an individual payment. However, *performance indices*, such as the German *DAX*, which require that dividends be reinvested in the index should not be adjusted for dividend payments.[4]

---

[4]Except for effects caused by taxation (corporate tax).

2.3   An Intuitive Model for Financial Risk Factors

2.3.1 Random Walks as the Basis for Pricing and Risk Models

The random walk is a mathematical model which is frequently used to characterize the random nature of real processes. The concept of the random walk has attained enormous importance in the modern financial world. Most option pricing models (such as the Black-Scholes model) and several methods used in modern risk management (for example, the variance-covariance method) and, of course, Monte Carlo simulations are based on the assumption that market prices are in part driven by a random element which can be represented by a random walk. It is therefore worthwhile to acquire an understanding of random walks if only to develop an intuitive comprehension of *stochastic processes* which are at the heart of this book.

A *random walk* can be described as follows: starting from some point in space, we travel a random distance in a randomly selected direction. Having arrived at a new point, another such step of random length and direction is taken. Each individual step in the procedure has a length and direction and thus can be represented as a vector as shown in Fig. 2.3. The completed random walk is a series of such vectors. Each base point of a vector is the end point of its predecessor.

At this point, we ask the following important question: what is the distance from the starting point after having completed a random walk consisting of n steps?[5] In other words: how large is the "end-to-end distance" represented by the length of the vector $\mathbf{R}$ in Fig. 2.3? The length and direction of the vector $\mathbf{R}$ are random since $\mathbf{R}$ is the sum of random steps. As a result, only statistical statements are available to describe the properties of this vector. For example, the length of this vector cannot be determined with certainty but we could calculate its mean length. This requires a large number of random walks with the same number of steps.

For each of these random walks, the square of the Euclidean norm of the end-to-end vector $\mathbf{R}$ is determined and used to calculate the mean $\langle \mathbf{R}^2 \rangle$. The mean end-to-end distance is than defined as the square root of this value. A Monte Carlo simulation (see Chap. 11) could be carried out to generate the random walks and obtain an estimate for the mean $\langle \mathbf{R}^2 \rangle$ by measuring $\mathbf{R}^2$, the square of the end-to-end vector, of each simulated random walk and then take

---

[5]Figure 2.3 for instance shows a random walk with $n = 8$.
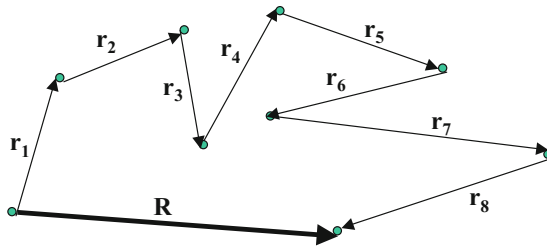
Fig. 2.3   A random walk with 8 steps in two dimensions

the average of these. In doing so, it can be observed that the square of the end-to-end vector is, on average, proportional to the number of steps taken in the random walk.[6]

$$\mathrm{E}\left[\mathbf{R}^2\right] \approx \langle \mathbf{R}^2 \rangle \propto n \ ,$$

Here $\mathrm{E}[x]$ denotes the expectation of a random variable x and $\langle x \rangle$ denotes its mean value.[7] This holds irrespective of the dimension of the space in which the random walk occurs. The same result holds for a random walk on a line, in a plane as in Fig. 2.3 or in a 15-dimensional Euclidean space [52].

The expectation of the end-to-end vector itself is equal to zero, i.e., $\mathrm{E}[\mathbf{R}] = 0$. This is immediate since the end-to-end vector $\mathbf{R}$ points in any direction with equal likelihood and therefore all these vectors cancel each other out in the average. In consequence, the variance $\mathrm{Var}[x]$ of the end-to-end vector is given by:

$$\mathrm{Var}(\mathbf{R}) \equiv \mathrm{E}\left[(\mathbf{R} - \mathrm{E}\,[\mathbf{R}])^2\right] = \mathrm{E}\left[\mathbf{R}^2\right] \propto n \ .$$

The variance of the end-to-end vector is thus also proportional to n. The standard deviation, defined as the square root of the variance, is consequently proportional to $\sqrt{n}$. This is the reason why the uncertainty in future market prices increases proportionally to the square root of time since, as will be shown in Sect. 2.3.2, the time corresponds to the number of steps when a random walk is used to model and interpret price movements. Many well-known rules in financial mathematics, for example, that the overnight value at risk is to

---

[6]The symbol "$\propto$" means "is proportional to".

[7]As shown in Sect. 31.1, the expectation converges to the mean value in the limit of an infinitesimal number of measurements.

be multiplied by a factor consisting of the square root of the liquidation period or that the conversion of daily volatilities into monthly volatilities is accomplished by multiplying by the square root of the number of days in the month, have their origin in the fact that the variance of the end-to-end vector of a random walk is proportional to the number of random steps!

The relation $E[\mathbf{R}^2] \propto n$ is so fundamental, that it is worthwhile to understand its theoretical meaning. As mentioned above, a random walk consists of a series of 'step' vectors. According to the rules of vector addition, the end-to-end vector is merely the sum of these vectors: $\mathbf{R} = \sum_{i=1}^{n} \mathbf{r_i}$. Since each individual step vector, $\mathbf{r_i}$ points in any direction with the same likelihood, the expectation of each step is $E[\mathbf{r_i}] = 0$ just as for the end-to-end vector. Since each of the steps are independent of one another and therefore uncorrelated, we have

$$E[\mathbf{r_i} \cdot \mathbf{r_j}] = \underbrace{E[\mathbf{r_i}]}_{=0} \cdot \underbrace{E[\mathbf{r_j}]}_{=0} = 0 \quad \forall\, i \neq j, i, j = 1, \ldots, n \,.$$

With this information about the individual steps, we immediately obtain

$$E\left[R^2\right] = E\left[\sum_{i=1}^{n} \mathbf{r}_i \cdot \sum_{j=1}^{n} \mathbf{r}_j\right] = \sum_{i,j=1}^{n} E\left[\mathbf{r}_i \cdot \mathbf{r}_j\right]$$

$$= \sum_{i=1}^{n} E\left[\mathbf{r}_i \cdot \mathbf{r}_i\right] + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \underbrace{E\left[\mathbf{r}_i \cdot \mathbf{r}_j\right]}_{0} = n\, b^2$$

where the constant b denotes the mean length of a single step:

$$b^2 = \frac{1}{n} \sum_{i=1}^{n} E[\mathbf{r_i^2}]$$

and as such is the constant of proportionality in the relation $E[\mathbf{R}^2] \propto n$. At no point did the dimension enter into the above derivation. The equation $E[\mathbf{R}^2] = b^2 n$ thus holds in any dimension and is an expression of a fundamental property of random walks, namely their *self-similarity*: the statistical properties of a random walk are always the same, regardless of the degree of detail with which they are observed. In other words, a step in a random walk can itself be represented as the end-to-end vector of a random walk with

"smaller" steps. Likewise, the end-to-end vector of a random walk can itself be considered a single step of a "coarser" random walk (*coarse graining*).

As yet, only the so called moments (expectation, variance, etc.) of the probability distribution $p(\mathbf{R})$ have received mention. However, the distribution of $\mathbf{R}$ itself can be determined as well. In general, the end-to-end vector of a random walk in a d-dimensional space has a normal distribution. The concrete expression of this fact in one dimension (the most important case for the financial applications) is given by [52]

$$p(\mathbf{R}) = \frac{1}{\sqrt{2\pi \operatorname{Var}(\mathbf{R})}} \exp\left[-\frac{(\mathbf{R} - \operatorname{E}[\mathbf{R}])^2}{2\operatorname{Var}(\mathbf{R})}\right] \qquad (2.10)$$

where

$$\operatorname{E}[\mathbf{R}] = 0 \quad \operatorname{Var}(\mathbf{R}) = b^2 n \;. \qquad (2.11)$$

Equations 2.10 and 2.11, together with the principle of self-similarity are the quintessential properties of the theory of random walks introduced here. They comprise all that is necessary for their application in the field of finance. For example, the "*normal distribution assumption*" for price movements is an immediate consequence of the theory of random walks. Contrary to popular belief, the normality of the distribution of relative price changes need not be *assumed*. It follows automatically that if market prices (or more precisely, the logarithm of the market prices, see below) behave as random walks, they must be normally distributed with a density function as given in 2.10. In financial literature, random walks are often *defined* having normally distributed individual step vectors. By doing so the "normal distribution assumption" is "injected" into the definition of the random walk. This assumption is equally unnecessary! As mentioned above, a random walk is defined as a series of completely random steps. Absolutely no assumptions have been made on the distribution of the individual steps or their sum. The fact that the end-to-end vector has a normal distribution follows *automatically* as a consequence of the central limit theorem.

2.3.2   Risk Factors as Random Walks

A *random variable* $z(t)$, whose values change randomly with time t is called a *stochastic process*. The process is called *Markov* if its future behavior is influenced solely by its value at the present time t. This means intuitively

that the future behavior is independent of the path taken to reach the present value. Assuming that the current value of a risk factor, such as a stock price or an interest rate, contains all the information about its historical development (this is called *weak market efficiency*), it follows that the subsequent values taken on by such a risk factor depend only on the current price and other external effects, such as politics, but not on the past prices or rates. Market prices can then assumed to be Markov processes.

In order to derive a model for the Markov process $S(t)$ representing the time-evolution of a risk factor, we assume that the process can be split into a random and a deterministic component. The deterministic component is called the *drift*. We will begin our discussion with an analysis of the random component.

The derivation of the model describing the random component given here is fundamentally different from that which is commonly found in the related literature and is based on the general properties of the random walk, Eq. 2.10 and Eq. 2.11. The literature usually begins with the introduction of a model and proceeds with a presentation of calculations and results following from the model assumption. It is more intuitive, however, to begin with a derivation of the model itself, i.e., to illustrate the steps and describe the motivation leading to its construction. Thus, this section serves the dual purpose of introducing a model for future analysis and an example of the modeling process. What follows is a detailed discussion of how we can construct a model describing a real phenomenon by means of abstraction (and intuition!) which is simple enough to allow a mathematical and/or computational analysis but still complex enough to capture the essential features of the real phenomenon.

A real process, for example the closing prices of a stock on 500 days, might look like the points shown in Fig. 2.4. We propose to model this process. To do so, several questions must first be answered:

- *What is the fundamental idea behind the model?*
  As mentioned above, the random properties of many processes can be traced back to the general concept of the random walk. We therefore use the random walk as the basis for our model.
- *A random walk in which dimension?*
  A market price can rise or fall, i.e., can change in only two directions. A random walk in d-dimensional space allows for an upward or downward change in d linearly independent directions, i.e., there are $2d$ possible changes in direction. Thus, the dimension required for the description of just upward or downward changes is $d = 1$.
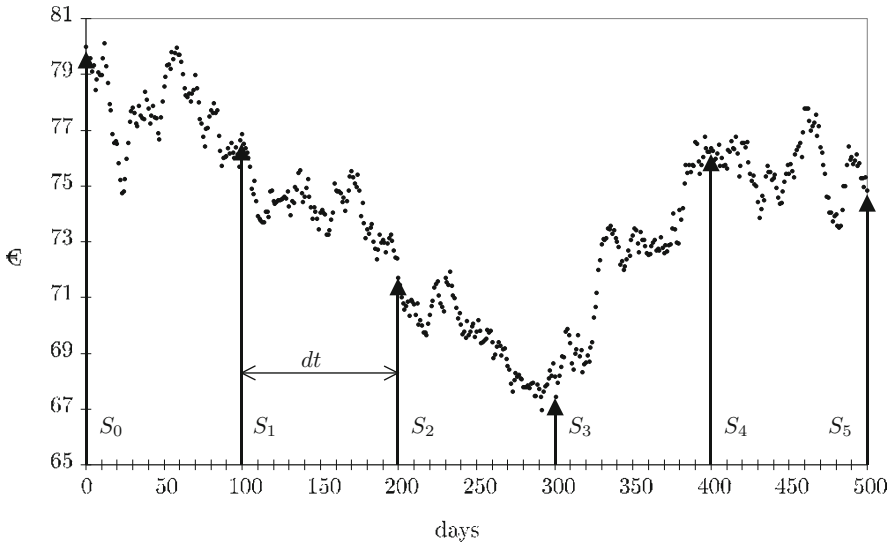
Fig. 2.4 End-of-day values of a stock price over a period of 500 trading days. The values after $0, 100, 200, \ldots, 500$ days are denoted by $S_0, S_1, S_2, \ldots, S_5$

- *Which real parameter is described by the number of steps n in the random walk?*
  In order to observe a change in price (in other words for a step in a random walk to be taken), one thing must occur: time must pass. If the price is observed in regular, fixed time intervals dt (for example, every 100 days as in Fig. 2.4, or daily or hourly, etc.), then the amount of time passing between steps is dt. If the entire random walk occurs between t (=today) and a future date T then

$$T - t = n dt . \tag{2.12}$$

  Since dt is a constant, the number of steps n is proportional to the time in which the random walk occurs, i.e., proportional to $T - t$.
- *Which real parameter should be modeled by a random walk?*
  At first glance, we might take the market price of a risk factor. The market price evolution $S_5 - S_0$ over the entire period in Fig. 2.4 can be decomposed into individual steps as follows:

$$S_5 - S_0 = (S_0 - S_1) - (S_2 - S_1) + (S_3 - S_2) + (S_4 - S_3) + (S_5 - S_4),$$

or more generally,

$$S_n - S_0 = \sum_{i=1}^{n} dS_i \quad \text{with } dS_i = S_i - S_{i-1} \ .$$

If the market price itself were a random walk, then as a result of the self similarity property, the individual steps dS_i would also be random walks. The price differences however are real cash amounts given in euros, for example. This would mean that a security costing 1000 euros would experience the same fluctuations (in euros) as one costing only 10 euros. This is surely not the case. It would make much more sense to consider *relative* fluctuations. Our next candidate for a step in our random walk could therefore be the ratio S_i/S_{i-1}. The ratio of the last price to the first is given by

$$\frac{S_5}{S_0} = \frac{S_1}{S_0}\frac{S_2}{S_1}\frac{S_3}{S_2}\frac{S_4}{S_3}\frac{S_5}{S_4},$$

or more generally,

$$\frac{S_n}{S_0} = \prod_{i=1}^{n} \frac{S_i}{S_{i-1}} \ .$$

This is the *product* of the individual steps and not their sum. A random walk however, is a vector and as such must always be the sum of its component steps. In light of this fact, the ratios S_i/S_{i-1} are completely unsuitable for the steps of a random walk as they are not even vectors! However, the ratios S_i/S_{i-1}, which make economic sense, can be utilized by converting the products into sums by taking the *logarithm* of both sides of the above equation. The *functional equation* for the logarithm is given by

$$\ln(a \times b) = \ln(a) + \ln(b) , \quad \ln(a/b) = \ln(a) - \ln(b) \ .$$

Thus taking the logarithm of both sides of the above product yields

$$\ln(\frac{S_5}{S_0}) = \ln(\frac{S_1}{S_0}) + \ln(\frac{S_2}{S_1}) + \ln(\frac{S_3}{S_2}) + \ln(\frac{S_4}{S_3}) + \ln(\frac{S_5}{S_4}),$$

or more generally,

$$\ln(\frac{S_n}{S_0}) = \sum_{i=1}^{n} \ln(\frac{S_i}{S_0 i - 1})$$

and thus,

$$\ln(S_n) - \ln(S_0) = \sum_{i=1}^{n} d \ln(S_i) \quad \text{with} \quad d \ln(S_i) = \ln(S_i) - \ln(S_{i-1}) .$$

This looks exactly like our first attempt with the sole exception that the market price S has been replaced with its logarithm $\ln(S)$. This small change makes it possible to satisfy both the economic requirement that the proportional changes in the market price be modeled and the mathematical requirement that a random walk be the sum of its component steps.

We have thus completed the fundamental construction of our model: *The random component of a market price is modeled by interpreting the logarithm of the price as a one-dimensional random walk with independently and identically distributed (iid) random steps and the number of these steps being proportional to the length of time during which the random walk takes place.*

Now we are in a position to apply what we know about random walks to draw conclusions about the evolution of market prices. It follows from Eq. 2.10 that the end-to-end vector $\mathbf{R} = \ln(S(T)/S(t))$ is normally distributed with expectation and variance as given in Eq. 2.11. Because in our model the time and the number of steps in the random walk are related as in Eq. 2.12, the time dependence of the end-to-end vector's variance can be calculated as

$$\mathrm{E}\left[\ln\left(\frac{S(T)}{S(t)}\right)\right] = 0 \quad , \quad \mathrm{Var}\left[\ln\left(\frac{S(T)}{S(t)}\right)\right] = b^2 n = \sigma^2(T - t) .$$

$$(2.13)$$

The logarithm of the market price is normally distributed, its variance being proportional to the time. The constant of proportionality σ^2 is well defined, even in the limiting case of the time intervals dt becoming infinitesimal. The *standard deviation*, defined as the square root of the variance, is thus proportional to the square root of the time. The constant of proportionality

σ is called *volatility*. Making use of Eqs. 2.13 and 2.12 gives us

$$\sigma^2 = \frac{b^2}{dt} = \frac{1}{T-t} \mathrm{Var}\left[\ln\left(\frac{S(T)}{S(t)}\right)\right] . \qquad (2.14)$$

Note again that the normal distribution has not been *assumed* at any point in this argument. Nor has it been assumed that the standard deviation increases proportionally to the square root of the time. The model consists solely of the statement above, that the logarithm of the market price is modeled as a random walk. The rest follows automatically! Owing to the self-similarity property, each step of the random walk, regardless of its length can be interpreted as a random walk itself. It follows that all changes $d\ln(S)$ in the logarithm of the market price over an infinitesimally small time interval dt are also normally distributed with an expectation of zero and the following variance:

$$\mathrm{Var}\left[d\ln(S(t))\right] = \mathrm{Var}\left[\ln(S(t+dt)) - \ln(S(t))\right] \qquad (2.15)$$

$$= \mathrm{Var}\left[\ln\left(\frac{S(t+dt)}{S(t)}\right)\right] = \sigma^2 dt .$$

Therefore the process $d\ln(S(t))$ is very similar to a very prominent and well understood process appearing time and again in physics, biology and other sciences, called the *Wiener process* which, for instance, describes what is known as *Brownian motion* in physics. Such a Wiener process W is a stochastic process which changes randomly by an amount dW over a time interval dt. These changes dW are normally distributed with a mean of zero and with a variance equal to the length of the time interval passed during the change, i.e., with a variance equal to dt. Or written more compactly:

$$dW \sim X\sqrt{dt} \quad \text{with} \quad X \sim \mathrm{N}(0,1) . \qquad (2.16)$$

Here, as always in this book, the notation $\mathrm{N}(x,y)$ denotes the *normal distribution* with mean x and variance y, i.e., $\mathrm{N}(0,1)$ denotes the *standard normal distribution*.[8] The sign "$\sim$" in these circumstances is to be read as "is distributed as". Equation 2.16 thus means in words: "dW is distributed as $\sqrt{dt}$ times a random number X. This random number X in turn is distributed according to the standard normal distribution." The definition 2.16 together

---

[8]The normal distribution and the standard normal distribution are presented in detail in Sect. A.4.3.

with the property 2.15 allows us to write the process for the infinitesimal random change $d \ln (S(t))$ very compactly as

$$d \ln (S(t)) = \sigma \, dW \ .$$

We have now arrived at the point where the discussion in the literature usually begins.

The only model parameter which has as yet been introduced is the volatility σ. This, however, is insufficient to provide an adequate description of the behavior of risk factors. For example, two parameters (denoted by u and d) are required for the binomial model dealt with in Chap. 9. All risk factors will be required to have an additional essential property, namely that they be martingales.[9] This subject will receive detailed discussion in Chap. 13. The intuitive meaning of the martingale property is that the current value of a risk factor can be represented as the appropriately discounted expected future value. In order to ensure that this property is satisfied when modeling market movements with a random walk, two parameters are required. To provide a motivation[10] for the second parameter in our random walk, we consider the following: the end-to-end vector in our 1-dimensional random walk is given by $\mathbf{R} = \ln [S(T)] - \ln [S(t)]$. If our random walk is to serve as a model for a stock price which, in the long term, should show a positive mean return, then the expectation of the end-to-end vector should not be equal to zero but should grow with time. This is incorporated into the model by introducing a deterministic term, called the *drift*:

$$d \ln [S(t)] = \mu dt + \sigma \, dW \ . \tag{2.17}$$

Since no additional random component has been introduced into the model, $\ln(S(t))$ retains the normal distribution with variance $\sigma^2(T - t)$ after the passage of time from t to T. The expectation, however, is now given by $\mu(T - t)$. The addition of the drift into our random walk has the consequence that the mean of the random walk is no longer 0 but rather $\mu(T - t)$.

---

[9]At least as long as the risk factors are tradeable, see Chap. 13.

[10]This is only a motivation based on the example and not a general explanation for a drift. For example, when considering an interest rate as a risk factor, there is no reason to believe that the change in the interest rate should show a "mean return". The more general and profounder explanation for the drift is that the risk factors must modeled in such a way that there exists a probability measure in which all (tradeable and properly normalized) financial instruments are martingales.

Table 2.7 Statistical properties of the logarithm of a risk factor and of the risk factor itself

	Normal distribution	Lognormal distribution
Random variable	$x = \ln\left(\frac{S(t+dt)}{S(t)}\right)$	$x = \frac{S(t+dt)}{S(t)}$
Density	$\frac{1}{\sqrt{2\pi\sigma^2 dt}}e^{-\frac{(x-\mu dt)^2}{2\sigma^2 dt}}$	$\frac{1}{x\sqrt{2\pi\sigma^2 dt}}e^{-\frac{(\ln(x)-\mu dt)^2}{2\sigma^2 dt}}$
$\mathbf{P}(\mathbf{x} \leq \mathbf{a})$	$\frac{1}{\sqrt{2\pi\sigma^2 dt}}\int_{-\infty}^{a} e^{-\frac{(x-\mu dt)^2}{2\sigma^2 dt}}\, dx$	$\frac{1}{\sqrt{2\pi\sigma^2 dt}}\int_{-\infty}^{\ln(a)} e^{-\frac{(x-\mu dt)^2}{2\sigma^2 dt}}\, dx$
Expectation	μdt	$e^{(\mu+\sigma^2/2)dt}$
Variance	$\sigma^2 dt$	$e^{2\mu dt}(e^{2\sigma^2 dt} - e^{\sigma^2 dt})$
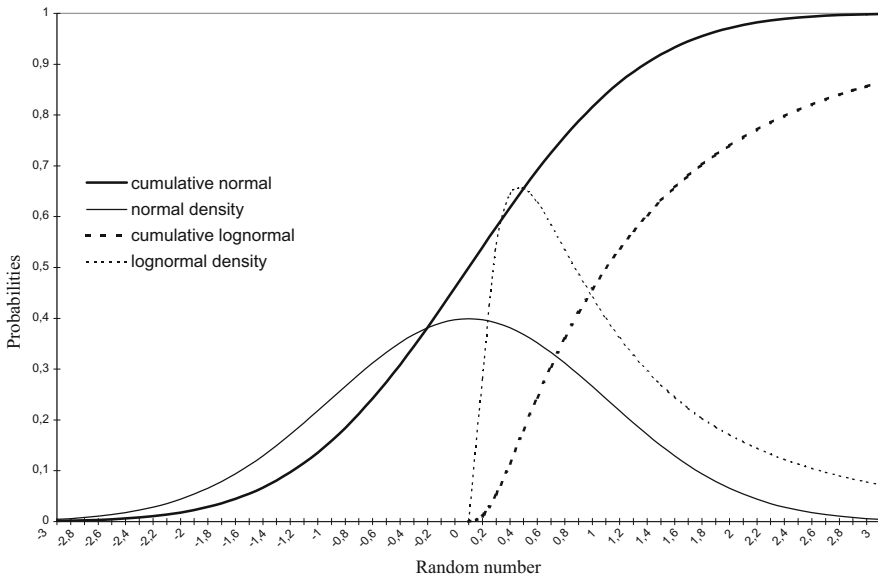
Fig. 2.5 The distribution from Table 2.7 with $\mu dt = 0$ and $\sigma\sqrt{dt} = 1$. With these parameters the normal distribution has mean 0 and variance 1 while the lognormal distribution has mean $\sqrt{e} \approx 1,65$ and variance $e^2 - e \approx 4,67$

A random variable whose logarithm has a normal distribution has a *lognormal distribution*. The price $S(t)$ is thus lognormally distributed. The most important properties of $\ln(S)$ and S summarized in Table 2.7, see also Sect. A.4.

In Fig. 2.5, the density functions and cumulative probabilities of the normal and lognormal distribution are shown. Here, the parameters μdt and $\sigma\sqrt{dt}$ have been set equal to 0 and 1, respectively.

The random walk model just derived can, of course, be generalized by allowing for non-constant drifts and volatilities. If σ is expressed as in Eq. 2.14 in terms of the variance and if the yield is left in its most general form as a function of time t and the market price $S(t)$ at time t, Eq. 2.17 can be expressed in the following generalized form:

$$d \ln (S(t)) = \mu(S(t), t)dt + X\sqrt{\mathrm{Var}[d \ln (S(t))]} \qquad (2.18)$$

$$\text{with } X \sim \mathrm{N}(0, 1) \ .$$

This equation is now in the general form of a stochastic diffusion process as given by Eq. 2.19 and is the starting point for more general stochastic models for the market parameters. These will receive a detailed treatment in the following chapter. Concrete examples of such generalized stochastic processes can be found in Sect. 32.1.

2.4   Ito Processes and Stochastic Analysis

In the previous section, risk factors in financial markets were introduced and a motivation and derivation of an intuitive model (the random walk) was provided to describe them. Now, this section will be devoted to the more theoretical fundamentals underlying these concepts, namely to stochastic analysis. *Stochastic analysis* is the branch of mathematics dealing with the investigation of stochastic processes. Particularly close attention will be paid to results which find application in finance. This subject is naturally quite theoretical, but should give the interested reader a deeper understanding of relationships between different fundamental concepts in financial mathematics. Nevertheless, the reader who is less interested in the mathematical details may skip over the rest of this Chapter and continue directly on to Chap. 3.

2.4.1 General Diffusion Processes

All stochastic processes in this book which will be used to model risk factors satisfy—as long as there is only one single random variable involved—an equation of the following form:

$$dS(t) = a\,(S, t)\,dt + b(S, t)\,dW \qquad \text{with} \quad dW \sim X\sqrt{dt}\,, \quad X \sim \mathrm{N}(0, 1) \ .$$
$$(2.19)$$

Here, W denotes—as always—a Brownian motion and X a standard normally distributed random variable. Processes satisfying an equation of this type are called *diffusion processes* or *Ito processes*. These quite general stochastic processes have long since been the subject of research in the field of stochastic analysis.

The parameters $a(S, t)$ and $b(S, t)$ are called the drift rate and the volatility of the Ito process. They may depend on the time t, on the stochastic process S or on both. The interpretation of the variable S depends on the particular application under consideration. In the simple models derived in Sect. 2.3, for example in Eq. 2.17, the logarithm of the risk factor was modeled as the stochastic variable. These models are of the form 2.19 with the stochastic variable being given by $\ln(S)$, where $b(S, t) = \sigma$ and $a(S, t) = \mu$.

The first moments of the conditional probability distribution of the general Ito process are

$$\mathrm{E}\left[dS(t)\right] = a(S, t)\, dt$$

$$\mathrm{Var}\left[dS(t)\right] = \mathrm{E}\left[(dS - \mathrm{E}[dS])^2\right] = \mathrm{E}\left[(b(S, t)\, dW)^2\right] = b(S, t)^2 dt$$

$$\mathrm{E}\left[dS(t)^2\right] = \mathrm{E}\left[(a(S, t)dt + b(S, t)dW)^2\right] \approx b(t, S)^2 dt + \text{higher order terms in } dt\ .$$

2.4.2 Ito's Lemma

A very important question is which stochastic differential equation is satisfied by a function $f(S)$ of a stochastic variable S, i.e., what change df is induced by an infinitesimal change dS in S. This question is answered by Ito's famous lemma, the proof of which will be sketched in this section. Typically, when we are interested in small changes in a function f, a *Taylor series* expansion is calculated for df. The only subtlety in our case (and this is critical) is that, because of Eq. 2.19, a *stochastic* variable is now involved:

$$\begin{aligned}
df(S, t) &= \frac{\partial f}{\partial S} dS + \frac{\partial f}{\partial t} dt + \frac{1}{2}\frac{\partial^2 f}{\partial S^2}(dS)^2 + \frac{1}{2}\frac{\partial^2 f}{\partial t^2}(dt)^2 + \cdots \\
&= \frac{\partial f}{\partial S}\underbrace{[a(S, t)\, dt + b(S, t)\, dW]}_{dS} + \frac{\partial f}{\partial t} dt \\
&\quad + \frac{1}{2}\frac{\partial^2 f}{\partial S^2}\left[b^2(S, t)\, dW^2 + 2a(S, t)\, b(S, t)\, dW dt\right] + O(dt^2)\ .
\end{aligned}$$

In this case we can *not*, as is usually the case, neglect the non-linear terms since the Brownian motion is of the order $dW \sim \sqrt{dt}$ and thus dW^2 is *not* small in comparison with dt. In fact, it can be shown that the square of Brownian motion has the following properties:

$$E\left[dW^2\right] = dt$$
$$\mathrm{Var}\left[dW^2\right] \sim dt^2 \ .$$

Thus, the variance of dW^2 vanishes in the linear order of dt. This means that, in linear order, dW^2 is no longer stochastic but can be set equal to its expectation, i.e.,

$$dW^2 \approx dt \ . \tag{2.20}$$

This is the central idea of Ito's lemma. Making this substitution and *then* ignoring all terms of order greater than dt, making use of the fact that $dW \sim \sqrt{dt}$ immediately gives the equation for infinitesimal (limit $dt \to 0$) changes in the function $f(S, t)$

$$df(S, t) = \left[\frac{\partial f}{\partial S}a(S, t) + \frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial S^2}b^2(S, t)\right]dt + \frac{\partial f}{\partial S}b(S, t)\,dW \ . \tag{2.21}$$

This is Ito's famous lemma. This equation has the same structure as Eq. 2.19. The function f is thus also an Ito process with its drift rate and volatility given by

$$a_f(S, t) = \frac{\partial f}{\partial S}a(S, t) + \frac{\partial f}{\partial t} + \frac{1}{2}b^2(S, t)\frac{\partial^2 f}{\partial S^2}$$
$$b_f(s, t) = \frac{\partial f}{\partial S}b(s, t) \ .$$

Equation 2.21 appears rather complicated. In practice, it is sufficient to remember the essential idea, namely Eq. 2.20. The equation can then be derived from a "mechanical" application of Taylor's theorem up to the linear term in dt and quadratic term in dW and making the substitution $dW^2 = dt$. A useful form of Eq. 2.21 can be obtained by observing that the terms involving

$\frac{\partial f}{\partial S}$ can be collected and expressed in terms of the process dS in accordance with Eq. 2.19:

$$df(S, t) = \left[\frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial S^2} b^2(S, t) \right] dt + \frac{\partial f}{\partial S} dS(t) . \tag{2.22}$$

The Process for the Risk Factor Itself

Ito's lemma is a powerful tool which will be applied frequently in this book. For example, we can immediately obtain the process for the market price from the process describing the *logarithm* of the market price given by Eq. 2.17. To do so, we define the stochastic variable[11] $y = \ln(S(t))$ and choose for f the function given by $f(y, t) = e^{y(t)}$. The original process $y(t) = \ln(S(t))$ satisfies the Eq. 2.17, i.e.,

$$dy(t) = \mu dt + \sigma \, dW .$$

The required derivatives of f can be easily calculated and applying Ito's lemma we obtain

$$f(y, t) = e^y \quad \Rightarrow \quad \frac{\partial f}{\partial y} = f \quad , \quad \frac{\partial f}{\partial t} = 0 \, , \quad \frac{\partial^2 f}{\partial y^2} = f$$

$$df(y, t) = \left(f(y, t)\mu + \frac{1}{2}\sigma^2 f(y, t) \right) dt + f(y, t) \sigma \, dW .$$

Setting $S(t) = f(y, t)$ yields:

$$dS(t) = S(t) \left(\mu + \frac{\sigma^2}{2} \right) dt + S(t)\sigma \, dW \quad . \tag{2.23}$$

In the literature, an equivalent approach is sometimes taken to arrive at this equation. A random walk model for the risk factor $S(t)$ is introduced in the first place (and not for its logarithm as done here) with a subsequent application of Ito's lemma to derive the process for $\ln(S(t))$. Explicitly, one begins with the process

$$dS(t) = S(t)\tilde{\mu} \, dt + S(t)\sigma \, dW \tag{2.24}$$

---

[11]Instead of S, another letter is used for the stochastic variable here in order to avoid confusion in the notation.

obtaining from Ito's lemma[12]

$$d \ln(S(t)) = (\widetilde{\mu} - \frac{\sigma^2}{2}) \, dt + S(t)\sigma \, dW \quad . \tag{2.25}$$

This result corresponds to Eqs. 2.23 and 2.17 in our development with a somewhat modified drift.

$$\widetilde{\mu} = \mu + \frac{\sigma^2}{2} . \tag{2.26}$$

The model is thus exactly the same, only the drift must be reinterpreted, see Eqs. 2.30 and 2.32 below.

The formulation of 2.23 and 2.24, in which the risk factor (and not its logarithm) is directly modeled as a stochastic process, of course also has the form of a diffusion process as in Eq. 2.19, with the choice of a and b given by $b(S,t) = S(t)\sigma$ and $a\,(S,t) = S(t)\widetilde{\mu}$ or equivalently, $a(S,t) = S(t)(\mu + \sigma^2/2)$.

The Process for the Risk Factor over a Finite Time Interval

Equation 2.19 describes the *infinitesimal* change in S and thus determines the *differential* of S. We are therefore dealing with a (partial) *differential equation*. Because it contains the stochastic component dW, it is referred to as a *stochastic partial* differential equation, often abbreviated by *SDE*. Special cases such as Eq. 2.23, for example, are SDEs as well.

With the aid of Ito's lemma and the general diffusion process, Eq. 2.19, an equation for *finite* changes in S (over a finite, positive time span δt) can be derived by solving the SPDE 2.23 (which holds for *infinitesimal* changes dS). For this purpose, we use the process 2.19 with $a\,(y,t) = 0$ and $b(y,t) = 1$, i.e., simply[13] $dy(t) = dW(t)$. Now we construct a function S of the stochastic variable y by

$$S(y,t) := S_0 \exp(\mu t + \sigma y)$$

---

[12] Equation 2.21 with $f(S,t) = \ln(S(t))$.

[13] In order to avoid confusion in the notation, we denote the stochastic variable by y.

where $y(t) = W(t)$ is the value of the Wiener process at time t, and S_0 is an arbitrary factor. Ito's lemma gives us the process for S induced by the process dy:

$$dS = \left[\underbrace{\frac{\partial S}{\partial y} \underbrace{a(y,t)}_{0}}_{\sigma S} + \underbrace{\frac{\partial S}{\partial t}}_{\mu S} + \underbrace{\frac{1}{2}\underbrace{b(y,t)^2}_{1}\underbrace{\frac{\partial^2 S}{\partial y^2}}_{\sigma^2 S}} \right] dt + \underbrace{\frac{\partial S}{\partial y}\underbrace{b(y,t)}_{1}}_{\sigma S} dW$$

$$= (\mu + \frac{\sigma^2}{2})S\,dt + \sigma S\,dW \ .$$

This corresponds exactly to the process in Eq. 2.23. This means that the process S thus constructed satisfies the stochastic differential equation 2.23, i.e., is a solution of this SPDE. Simply making the substitution $t \rightarrow t + \delta t$ we obtain

$$S(t + \delta t) = S_0 \exp(\mu t + \mu \delta t + \sigma y(t + \delta t))$$
$$= S_0 \exp(\mu \delta t + \sigma W(t + \delta t))$$
$$= S_0 \exp(\sigma W(t) + \mu \delta t + \sigma \delta W) \ ,$$

where in the second step we absorbed $\exp(\mu t)$ in the (still arbitrary) S_0 and in the third step we adopted the notation δW for a change in a Brownian motion after the passing of a finite time interval δt:

$$\delta W := W(t + \delta t) - W(t) \Longrightarrow \delta W \sim N(0, \delta t) \ . \tag{2.27}$$

The first term in the exponent refers to (already known) values at time t. It can also be absorbed into the (still arbitrary) factor S_0, i.e.,

$$S(t + \delta t) = S_0 \exp(\mu \delta t + \sigma \delta W) \ .$$

Finally, the (still arbitrary) S_0 is chosen so that $S(t + \delta t) \overset{\delta t \rightarrow 0}{=} S(t)$ holds. This corresponds to the initial condition for the solution of the SPDE. Thus, we obtain the change in S corresponding to Eq. 2.23 for a finite, positive time span δt:

$$S(t + \delta t) = S(t) \exp(\mu \delta t + \sigma \delta W) \quad \text{where } \delta W \sim \mathrm{N}(0, \delta t) \ . \tag{2.28}$$

Analogously, Eq. 2.24 gives the corresponding change in S over a finite, positive time span δt as:

$$S(t + \delta t) = S(t) \exp\left((\widetilde{\mu} - \frac{\sigma^2}{2})\delta t + \sigma \delta W\right) \quad \text{where } \delta W \sim N(0, \delta t) .$$

(2.29)

The Drift and the Expected Return

With the risk factor evolution over finite time spans at our disposal, we are now able to answer a question which often confuses market participants: which is the *mean* (or *expected*) *return* of the risk factor over a finite time span? Is it the drift μ of $\ln S$ as in Eq. 2.17 or the drift $\widetilde{\mu}$ of S itself as in Eq. 2.24? The answer to this question does in fact depend on the compounding methods used.

In continuous compounding, if a security is worth $S(t)$ at time t and worth $S(t + \delta t)$ at time $t + \delta t$ then the return R of this security over this time span is *defined* by the equation $S(t + \delta t) = S(t)e^{R\delta t}$, i.e.,

$$R \equiv \frac{1}{\delta t} \ln\left(\frac{S(t + \delta t)}{S(t)}\right)$$

$$= \frac{1}{\delta t} \ln\left(\frac{S(t) \exp(\mu \delta t + \sigma \delta W)}{S(t)}\right)$$

$$= \mu + \frac{\sigma}{\delta t} \delta W ,$$

where we have used Eq. 2.28 for $S(t + \delta t)$. The *mean* (or *expected*) return is of course simply the expectation of this return. Since $\delta W \sim N(0, \delta t)$, i.e., $E[\delta W] = 0$, this expectation is

$$E[R] = \mu = \widetilde{\mu} - \frac{\sigma^2}{2} .$$

(2.30)

Thus, in continuous compounding the drift parameter μ is the mean return.[14]

In other compounding methods, things are different. We look at linear compounding as an example, since this method is especially important for

---

[14]Note that the expectation of the *logarithm* of the market price was required for the determination of the mean return not the expectation of the market price itself.

the following reason: Very often, the return R is defined as the relative price change over a time period. This very natural definition is in fact equivalent to linear compounding:

$$R \, \delta t \equiv \frac{S(t + \delta t) - S(t)}{S(t)}$$

$$\Longleftrightarrow \qquad S(t)(1 + R\delta t) \qquad = S(t + \delta t) \,. \tag{2.31}$$

In the last line we recognize the linear compounding factor as in Table 2.5. Using now Eq. 2.28 for $S(t + \delta t)$ we find:

$$1 + R\delta t = \exp(\mu \delta t + \sigma \delta W) \,.$$

In contrast to the situation in continuous compounding, we now do not have a logarithm at our disposal to conveniently get rid of the (non-linear) exponential function. Thus, before we can take any expectations we need to expand the exp-function:

$$\exp(\mu \delta t + \sigma \delta W) = 1 + \mu \delta t + \sigma \delta W + \frac{1}{2}(\mu \delta t + \sigma \delta W)^2 + \mathcal{O}(\delta W^3)$$

$$= 1 + \mu \delta t + \sigma \delta W + \frac{1}{2}\sigma^2 \delta W^2 + \mathcal{O}(\delta W^3)$$

$$= 1 + \mu \delta t + \sigma \delta W + \frac{1}{2}\sigma^2 \delta t + \mathcal{O}(\delta W^3) \,.$$

where in the last step we have used Eq. 2.20. Now we can take the expectations and exploit $\mathrm{E}[\delta W] = 0$ to find that up to linear order in δt we have

$$1 + \mathrm{E}[R]\delta t = \mathrm{E}\left[1 + \mu \delta t + \sigma \delta W + \frac{1}{2}\sigma^2 \delta t + \mathcal{O}(\delta W^3)\right]$$

$$= 1 + \mathrm{E}[\mu \delta t] + \underbrace{\mathrm{E}[\sigma \delta W]}_{0} + \mathrm{E}\left[\frac{1}{2}\sigma^2 \delta t\right] + \underbrace{\mathrm{E}[\mathcal{O}(\delta W^3)]}_{\mathcal{O}(\delta W^4) = \mathcal{O}(\delta t^2)}$$

$$= 1 + \mu \delta t + \frac{1}{2}\sigma^2 \delta t + \mathcal{O}(\delta t^2) \,.$$

Extracting $E[R]$ on the left side yields

$$E[R] = \frac{1 + \left(\mu + \sigma^2/2\right)\delta t - 1}{\delta t} + \mathcal{O}(\delta t) \,.$$

Therefore, in linear compounding

$$E[R] \approx \mu + \frac{\sigma^2}{2} = \widetilde{\mu} \,. \tag{2.32}$$

2.4.3 Transition Probabilities, Forward and Backward Equation

Since S in Eq. 2.19 is a stochastic process, its *exact* evolution over time cannot be described even by solving the stochastic differential equation. This becomes clear in the above examples where the solutions to Eqs. 2.28 and 2.29 still contain a random component $\delta W \sim N(0, \delta t)$. However, probabilities can be given for the stochastic process at time t' to take on values between a and b, given that the value of S at an earlier time t is known. A formal notation for this probability is

$$P(a \leq S(t') \leq b \,|\, S(t) = S) = \int_a^b p(S', t' \,|\, S, t)\, dS' \,. \tag{2.33a}$$

where the *transition probabilities* denoted by $p(S', t' \,|\, S, t)$ indicate the probability that the stochastic variable at time t' is exactly equal to S', on the condition that it was equal to S at time t. The transition probability[15] contains all information about the associated stochastic process 2.19. This fact prompts a more detailed investigation of transition probabilities.

The Forward Equation

The time evolution of the transition probabilities driven by the stochastic process 2.19 can be described using two differential equations. One of these equations involves the derivative with respect to the future variables S', t' and

---

[15] This concept of transition probabilities is quite similar to the concept of (Feynman-) propagators used in quantum field theories of elementary particle physics.

answers the question: if the stochastic process is equal to S at time t, what distribution does it have at a later time $t' > t$? This equation is called the *forward equation*, also known as the *Fokker-Planck equation*. It is explicitly given by[16]

$$\frac{\partial p}{\partial t'} - \frac{1}{2}\frac{\partial^2}{\partial S'^2}\left[b(S',t')^2 p\right] + \frac{\partial}{\partial S'}\left[a(S',t')p\right] = 0 \ . \tag{2.34}$$

The initial condition that the value of the process at time $t' = t$ is exactly equal to S can be formally expressed with the help of the *Dirac delta function*

$$p(S',t' = t \mid S,t) = \delta(S'-S) \ .$$

The solution to the forward equation with this initial condition describes the widening of the probability distribution of the process over time starting from the "sharp" delta function. The forward equation is useful in situations where S has an established value at a fixed time and we are interested in the probability distribution of S at a later time, in other words, when information available now is to be used to calculate *forward* in time.

The Backward Equation

In mathematical finance, we are more often interested in the opposite situation: information at a future point in time is known (for example, at the maturity of an option) and we wish to calculate values backwards to an earlier point in time (today, say). To do so, we require a differential equation involving derivatives with respect to the earlier time variables S and t which answers the question: given that the stochastic variable is equal to S' at time t', how is the stochastic process distributed at an earlier point in time $t < t'$? This question can be answered using the *backward* or *Kolmogorov equation* This is given explicitly by

$$\frac{\partial p}{\partial t} + \frac{1}{2}b^2(S,t)\frac{\partial^2 p}{\partial S^2} + a(S,t)\frac{\partial p}{\partial S} = 0 \ . \tag{2.35}$$

---

[16]To simplify the notation, we neglect here the arguments of p, a habit we will continue to observe often in the remainder of this chapter. If p stands in an equation without an argument, it should be read as $p(S',t' \mid S,t)$.

The associated initial condition is given by

$$p(S', t' \,|\, S, t = t') = \delta(S - S') \,.$$

Note that for all process whose parameters a and b are not dependent on S, the only difference between the forward and backward equation is the sign of the term involving the second derivative.

A Derivation of the Forward and Backward Equations

Both the forward equation 2.34 and the backward equation 2.35 are of such a fundamental nature that it is possible to complete their derivation on the basis of first principles. This derivation involves quite extensive calculations (essentially, taking products of Taylor series) which the reader need not necessarily work through (the reader who does not wish to take the time is recommended to continue on to Sect. 2.4.4). However, experience shows that one tends to feel somewhat ill at ease when such fundamental equations simply appear out of thin air and are accepted without a sound explanation. Therefore, the derivation of the forward and backward equation is presented explicitly here in greater detail than is commonly found in the literature on mathematical finance. This gives the reader the opportunity to obtain a real understanding of these two fundamental equations.

We begin by deriving the *forward* equation. We split the probability that the stochastic process will travel from (S, t) to (S', t') into the probability that by time $t' - \delta t$ it will have arrived at $S' - \delta S$, $S' + \delta S$ or S' and then will proceed to S' in the remaining time δt with the probabilities δp_+, δp_- and δp_0, respectively, thus

$$
\begin{aligned}
p(S', t' \,|\, S, t) = {}& p(S' - \delta S, t' - \delta t \,|\, S, t)\delta p_+(S' - \delta S, t' - \delta t) \quad (2.36) \\
& + p(S', t' - \delta t \,|\, S, t)\delta p_0(S', t' - \delta t) \\
& + p(S' + \delta S, t' - \delta t \,|\, S, t)\delta p_-(S' + \delta S, t' - \delta t)
\end{aligned}
$$

where

$$
\begin{aligned}
\delta p_\pm(S, t) &:= p(S \pm \delta S, t + \delta t \,|\, S, t) \\
\delta p_0(S, t) &:= p(S, t + \delta t \,|\, S, t) = 1 - \delta p_+(S, t) - \delta p_-(S, t) \,.
\end{aligned}
$$
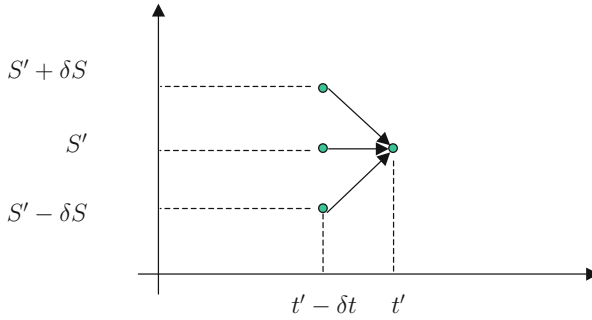
Fig. 2.6   The trinomial tree used to derive the forward equation

Intuitively δp_+ is the probability that the process will increase by δS in the short time span δt, δp_- the probability that it will decrease by δS in the time span δt, and δp_0 is the probability that it will remain constant in the time span δt. Other possibilities for this stochastic process within this time span are not admitted. This split of the probabilities is graphically represented by the trinomial tree in Fig. 2.6.

The relationship between these δp and the stochastic process 2.19 results from the requirement that the first two moments, i.e., the expectation and the variance of the discrete process defined for the trinomial tree, agree with the first two moments of the continuous Ito process in the limit as $\delta t \to dt \to 0$. The expectation of δS is

$$\mathrm{E}\,[\delta S] = \delta p_+ \delta S + \delta p_0 0 - \delta p_- \delta S = (\delta p_+ - \delta p_-)\delta S \ .$$

By definition, its variance is given by

$$\mathrm{var}\,[\delta S] = \mathrm{E}\left[(\delta S - \mathrm{E}\,[\delta S])^2\right]$$
$$= \delta p_+ (\delta S - \mathrm{E}\,[\delta S])^2 + \delta p_- (-\delta S - \mathrm{E}\,[\delta S])^2 + \delta p_0 (0 - \mathrm{E}\,[\delta S])^2 \ .$$

Substituting $\mathrm{E}[\delta S] = (\delta p_+ - \delta p_-)\delta S$ and $\delta p_0 = 1 - \delta p_+ - \delta p_-$ yields, after several simple algebraic manipulations

$$\mathrm{Var}\,[\delta S] = [\delta p_+ (1 - \delta p_+ + \delta p_-)^2 + \delta p_- (1 + \delta p_+ - \delta p_-)^2$$
$$+ (1 - \delta p_+ - \delta p_-)(\delta p_+ - \delta p_-)^2]\delta S^2$$
$$= (\delta p_+ + \delta p_-)(1 - (\delta p_+ + \delta p_-))\delta S^2$$
$$\approx (\delta p_+ + \delta p_-)\delta S^2 \ .$$

where only terms up to order $\delta p \delta S^2$ were considered in the last step. The requirement that the expectation and the variance be equal to those of the Ito process 2.19 as $\delta t \to dt \to 0$, i.e., up to linear order in δt means that

$$(\delta p_+ - \delta p_-)\delta S = \mathrm{E}\,[\delta S] \overset{\delta S \to dS}{\longrightarrow} \mathrm{E}\,[dS] = a(S,t)dt$$

$$(\delta p_+ + \delta p_-)\delta S^2 \approx \mathrm{Var}\,[\delta S] \overset{\delta S \to dS}{\longrightarrow} \mathrm{Var}\,[dS] = b(S,t)^2 dt .$$

This is obviously achieved by choosing

$$\delta p_\pm(S,t) = \frac{1}{2}\frac{\delta t}{\delta S^2}\left[\pm a(S,t)\delta S + b(S,t)^2\right] . \tag{2.37}$$

And, in order to retain the equality $\delta p_0 = 1 - \delta p_+ - \delta p_-$ we have

$$\delta p_0(S,t) = 1 - \frac{\delta t}{\delta S^2}b(S,t)^2 .$$

Substituting these values in our original equation 2.36 and expanding *all* terms appearing in this expression in a Taylor series about the point (S',t') up to order δt finally yields the forward equation. As a result of Ito's lemma we have $\delta S^2 \sim \delta t$. This means that all terms of order 1, δS, δS^2 and δt are retained and all higher order terms can be neglected. Note that $\delta t/\delta S^2$ in Eq. 2.37 is of order 1. Explicitly, the Taylor series for p is given by:

$$p(S' \pm \delta S, t' - \delta t \,|S,t) = p(S',t'\,|S,t) \pm \frac{\partial p}{\partial S'}\delta S + \frac{1}{2}\frac{\partial^2 p}{\partial S'^2}\delta S^2 - \frac{\partial p}{\partial t'}\delta t + \cdots$$

$$p(S',t' - \delta t \,|S,t) = p(S',t'\,|S,t) - \frac{\partial p}{\partial t'}\delta t + \cdots$$

But now, all terms involving δp must also be expanded in consideration of Eq. 2.37:

$$\delta p_+(S' - \delta S, t' - \delta t) \approx \delta p_+(S',t') - \frac{\partial \delta p_+}{\partial S'}\delta S + \frac{1}{2}\frac{\partial^2 \delta p_+}{\partial S'^2}\delta S^2 - \frac{\partial \delta p_+}{\partial t'}\delta t + \cdots$$

$$\approx \frac{1}{2}\frac{\delta t}{\delta S^2}\left[b^2 + a\delta S\right] - \frac{1}{2}\frac{\delta t}{\delta S^2}\left[\frac{\partial b^2}{\partial S'} + \frac{\partial a}{\partial S'}\delta S\right]\delta S$$

$$+ \frac{1}{4}\delta t\left[\frac{\partial^2 b^2}{\partial S'^2} + \frac{\partial^2 a}{\partial S'^2}\delta S\right] - \frac{1}{2}\frac{\delta t}{\delta S^2}\left[\frac{\partial b^2}{\partial t'} + \frac{\partial a}{\partial t'}\delta S\right]\delta t + \cdots$$

$$\approx \frac{1}{2} \frac{\delta t}{\delta S^2} b^2 + \frac{1}{2} \frac{\delta t}{\delta S^2} \left[a - \frac{\partial b^2}{\partial S'} \right] \delta S$$

$$+ \frac{1}{2} \left[\frac{1}{2} \frac{\partial^2 b^2}{\partial S'^2} - \frac{\delta t}{\delta S^2} \frac{\partial b^2}{\partial t'} - \frac{\partial a}{\partial S'} \right] \delta t$$

where in the last step, the terms are written in increasing order and those of order $\sim \delta S \delta t$ or smaller are neglected. Analogously we obtain

$$\delta p_-(S' + \delta S, t' - \delta t) \approx \frac{1}{2} \frac{\delta t}{\delta S^2} b^2 - \frac{1}{2} \frac{\delta t}{\delta S^2} \left[a - \frac{\partial b^2}{\partial S'} \right] \delta S$$

$$+ \frac{1}{2} \left[\frac{1}{2} \frac{\partial^2 b^2}{\partial S'^2} - \frac{\delta t}{\delta S^2} \frac{\partial b^2}{\partial t'} - \frac{\partial a}{\partial S'} \right] \delta t \ ,$$

and finally, for the probability of remaining at S':

$$\delta p_0(S', t' - \delta t) \approx \delta p_0(S', t') - \frac{\partial \delta p_0}{\partial t'} \delta t + \cdots \approx 1 - \frac{\delta t}{\delta S^2} b^2 + \frac{\delta t}{\delta S^2} \frac{\partial b^2}{\partial t'} \delta t \ .$$

These expansions substituted in Eq. 2.36 give

$$p \approx \left[p - \frac{\partial p}{\partial S'} \delta S + \frac{1}{2} \frac{\partial^2 p}{\partial S'^2} \delta S^2 - \frac{\partial p}{\partial t'} \delta t \right]$$

$$\times \left(\frac{1}{2} \frac{\delta t}{\delta S^2} b^2 + \frac{1}{2} \frac{\delta t}{\delta S^2} \left[a - \frac{\partial b^2}{\partial S'} \right] \delta S + \frac{1}{2} \left[\frac{1}{2} \frac{\partial^2 b^2}{\partial S'^2} - \frac{\delta t}{\delta S^2} \frac{\partial b^2}{\partial t'} - \frac{\partial a}{\partial S'} \right] \delta t \right)$$

$$+ \left[p - \frac{\partial p}{\partial t'} \delta t \right] \left(1 - \frac{\delta t}{\delta S^2} b^2 + \frac{\delta t}{\delta S^2} \frac{\partial b^2}{\partial t'} \delta t \right)$$

$$+ \left[p + \frac{\partial p}{\partial S'} \delta S + \frac{1}{2} \frac{\partial^2 p}{\partial S'^2} \delta S^2 - \frac{\partial p}{\partial t'} \delta t \right]$$

$$\times \left(\frac{1}{2} \frac{\delta t}{\delta S^2} b^2 - \frac{1}{2} \frac{\delta t}{\delta S^2} \left[a - \frac{\partial b^2}{\partial S'} \right] \delta S + \frac{1}{2} \left[\frac{1}{2} \frac{\partial^2 b^2}{\partial S'^2} - \frac{\delta t}{\delta S^2} \frac{\partial b^2}{\partial t'} - \frac{\partial a}{\partial S'} \right] \delta t \right)$$

where all values appearing in the equation are evaluated at (S', t'). Taking the product and neglecting all terms of order $\sim \delta S \delta t$ or smaller, we obtain, in

linear order of δt:

$$p \approx p + p\delta t \left(\frac{1}{2} \frac{\partial^2 b^2}{\partial S'^2} - \frac{\partial a}{\partial S'} \right) - \frac{\partial p}{\partial S'} \delta S \frac{\delta t}{\delta S^2} \left[a - \frac{\partial b^2}{\partial S'} \right] \delta S$$

$$+ \frac{1}{2} \frac{\partial^2 p}{\partial S'^2} \delta S^2 \frac{\delta t}{\delta S^2} b^2 - \frac{\partial p}{\partial t'} \delta t \ .$$

This immediately yields the forward equation:

$$0 = \frac{1}{2} \frac{\partial^2 p}{\partial S'^2} b^2 + p \left(\frac{1}{2} \frac{\partial^2 b^2}{\partial S'^2} - \frac{\partial a}{\partial S'} \right) - \frac{\partial p}{\partial S'} \left(a - \frac{\partial b^2}{\partial S'} \right) - \frac{\partial p}{\partial t'}$$

$$0 = \frac{1}{2} \frac{\partial^2 p}{\partial S'^2} b^2 + \frac{\partial p}{\partial S'} \frac{\partial b^2}{\partial S'} + \frac{1}{2} p \frac{\partial^2 b^2}{\partial S'^2} - p \frac{\partial a}{\partial S'} - \frac{\partial p}{\partial S'} a - \frac{\partial p}{\partial t'}$$

$$0 = \frac{1}{2} \frac{\partial^2}{\partial S'^2} (b^2 p) - \frac{\partial}{\partial S'} (a\, p) - \frac{\partial p}{\partial t'} \ .$$

The derivation of the *backward* equation is analogous but requires considerably less effort: since the variable in a trinomial tree has only three possibilities to travel from a point S' at time t' over the next time step δt, the probability that the variable will be equal to S' at time t' is equal to the sum of the probabilities that it will travel to either $S' + \delta S$, $S' - \delta S$ or S' under the condition that it started at S at time t, see Fig. 2.7.
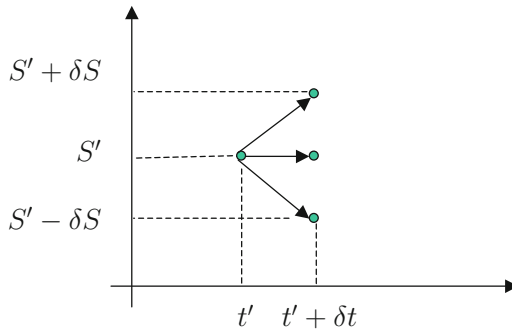
Fig. 2.7 The trinomial tree used to derive the backward equation

Thus we can state

$$
\begin{aligned}
p(S', t' \,|\, S, t) ={}& p(S' + \delta S, t' + \delta t \,|\, S, t)\delta p_+(S', t') \\
&+ p(S', t' + \delta t \,|\, S, t)\delta p_0(S', t') \\
&+ p(S' - \delta S, t' + \delta t \,|\, S, t)\delta p_-(S', t') \,.
\end{aligned}
$$

This equation means in words: the probability of being S' at time t' equals

- the probability of being $S' + \delta S$ at time $t' + \delta t$ multiplied by the probability of having made an "up move" in the time interval from t' bis $t' + \delta t$,
- plus the probability of being $S' - \delta S$ at time $t' + \delta t$ multiplied by the probability of having made a "down move" in the time interval from t' to $t' + \delta t$
- plus the probability of being S' at time $t' + \delta t$ multiplied by the probability of having made a "null move" in the time interval from t' to $t' + \delta t$.

We can now see immediately that the backward equation is more easily derived than its counterpart since the probabilities δp are needed at the point (S', t'), and consequently, a Taylor series expansion about this point is unnecessary. We only need the Taylor series expansion of p given by

$$
\begin{aligned}
p ={}& \left[p + \frac{\partial p}{\partial S'}\delta S + \frac{1}{2}\frac{\partial^2 p}{\partial S'^2}\delta S^2 + \frac{\partial p}{\partial t'}\delta t \right] \frac{1}{2}\frac{\delta t}{\delta S^2} \left[b^2 + a\delta S \right] \\
&+ \left[p + \frac{\partial p}{\partial t'}\delta t \right] \left[1 - \frac{\delta t}{\delta S^2}b^2 \right] \\
&+ \left[p - \frac{\partial p}{\partial S'}\delta S + \frac{1}{2}\frac{\partial^2 p}{\partial S'^2}\delta S^2 + \frac{\partial p}{\partial t'}\delta t \right] \frac{1}{2}\frac{\delta t}{\delta S^2} \left[b^2 - a\delta S \right] \,.
\end{aligned}
$$

All terms appearing in the above expression are evaluated at the point (S', t'). Taking the products and neglecting all terms of order $\sim \delta S \delta t$ or smaller gives, in linear order of δt,

$$
p = p + \frac{\partial p}{\partial S'}\delta S \frac{\delta t}{\delta S^2} a\delta S + \frac{1}{2}\frac{\partial^2 p}{\partial S'^2}\delta S^2 \frac{\delta t}{\delta S^2} b^2 + \frac{\partial p}{\partial t'}\delta t \,.
$$

Which, after rearranging the terms, immediately yields the backward equation 2.35.

2.4.4 Forward and Backward Equation in the Black-Scholes World

We consider the simple example of the process given by Eq. 2.24, namely $dS(t) = S(t)\widetilde{\mu}\,dt + S(t)\sigma\,dW$, i.e., we simply have

$$a(S', t') = S'\widetilde{\mu}$$
$$b(S', t') = S'\sigma \ .$$

This process, or its equivalent 2.17, is, as has been shown above, the simplest process of relevance in mathematical finance. Reasonable and often even analytical expressions for the prices of financial instruments can be obtained by solving these equations. For example, we will later show that such processes form the basis for the famous Black-Scholes option pricing formula. We thus speak of the *Black-Scholes world* when referring to the description of market parameters using such processes. With these assumptions, the backward equation becomes

$$\frac{\partial p}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 p}{\partial S^2} + \widetilde{\mu} S \frac{\partial p}{\partial S} = 0 \ ,$$

and likewise the forward equation is simply[17]

$$\frac{\partial p}{\partial t'} - \frac{\sigma^2}{2} S'^2 \frac{\partial^2 p}{\partial S'^2} + \left[\widetilde{\mu} - \sigma^2\right] S' \frac{\partial p}{\partial S'} + \left[\widetilde{\mu} - \sigma^2\right] p = 0 \ .$$

---

[17] The derivation of this equation is a simple application of the product rule:

$$\frac{\partial p}{\partial t'} - \frac{1}{2}\frac{\partial^2}{\partial S'^2}\left[S'^2\sigma^2 p\right] + \frac{\partial}{\partial S'}\left[S'\widetilde{\mu}\,p\right] = 0$$

$$\frac{\partial p}{\partial t'} - \frac{\sigma^2}{2}\frac{\partial}{\partial S'}\left[2S'p + S'^2\frac{\partial p}{\partial S'}\right] + \widetilde{\mu}\left[p + S'\frac{\partial p}{\partial S'}\right] = 0$$

$$\frac{\partial p}{\partial t'} - \frac{\sigma^2}{2}\left[2p + 2S'\frac{\partial p}{\partial S'} + S'^2\frac{\partial^2 p}{\partial S'^2}\right] + \widetilde{\mu}\left[p + S'\frac{\partial p}{\partial S'}\right] = 0 \ .$$

The solution to this differential equation with the above initial condition $p(S', t \mid S, t) = \delta(S' - S)$ provides an explicit example for a transition probability:

$$p(S', t' \mid S, t) = \frac{1}{S'\sqrt{2\pi\sigma^2(t'-t)}} \exp\left\{ \frac{-\left[\ln(S'/S) - (\widetilde{\mu} - \sigma^2/2)(t'-t)\right]^2}{2\sigma^2(t'-t)} \right\}.$$
(2.38)

The corresponding equation and its solution for the equivalent process 2.23 can of course be found by simply making the substitution $\widetilde{\mu} = \mu + \sigma^2/2$ in accordance with 2.26.

3

Financial Instruments: A System of Derivatives and Underlyings

As mentioned in the introduction, *trading* can be defined as an agreement between two parties in which one of the two consciously accepts a financial risk in return for the receipt of a specified payment or at least the expectation of such a payment at same future time from the counterparty. *Financial instruments*, also called *financial products*, are instruments which make such a *risk mitigation* by *risk transfer* possible. The purpose of this section is to present a classification of such instruments in a system of underlyings and derivatives, with a special focus on interest rate instruments, with their broad variety of complex structures. However, instruments on other risk factors, e.g. equity prices or foreign exchange (FX) rates, can be classified in a mostly similar manner. Otherwise, the particular differences will be mentioned explicitly. The classification of financial instruments and transactions will be based on criteria like term to maturity, optionality, or risk structure.

3.1 Issuer and Counterparties

An institution in need of capital may decide instead of negotiating a debt contract with its bank to sell securities to the public. Such securities are often *tradable* or *fungible*. In the following, we will use the term security for any such tradable assets, regardless of the concrete legal definition.

In this case, the creditor, who bought the securities, may decide and any time to sell them (in full or partially) a third party. This transaction would include three parties: the buyer, the debtor or *issuer* of the security, and the

seller or original creditor. The buyer and seller are the two *counterparties* of the transaction. For *over-the-counter* or *OTC* trades, both parties of the transaction negotiate directly the conditions of the transaction. This is in contrast to exchange traded contracts, where the conditions are defined by the exchange, who is an intermediary between buyer and seller of the contract. Therefore, exchange traded products are more standardized. If an institution decides to sell a financial product via an exchange instead of OTC, the financial product needs to be securitized, i.e. turned into a security of which the institution becomes the issuer. This require to fulfill some additional requirements, however.

The further classification of financial transactions differentiates the term until the transaction can be considered as closed (difference between spot trades and non-spot trades like forwards, futures, swaps, and options) and the term, until all risks involved in this transaction terminate (e.g., difference between *money market* and *capital market*). In general, both time horizons differ, e.g. in the case of spot transaction of a long term security.

3.2   Spot Transactions

In the *spot market*, transactions take place in which financial instruments are immediately (more precisely, as of the current *value date* or *spot date*) exchanged and paid for. Selling or buying a security at spot day is a spot transaction, regardless of the complexity or term to maturity of the underlying security, which could be a share, bond or even a securitized credit portfolio (e.g. an *Asset Backed Security* or *ABS*).

Foreign exchange transaction are denoted as *FX spot trade* is a, exchanging an amount of money in one currency against an amount in another currency of equal value at spot day, given the current foreign exchange rate. E.g., counterparty A pays to counterparty B the amount of €100 and gets in exchange from counterparty B the amount 135$.

In the FX and money market, it is quite common to trade at even shorter maturities, i.e. settlement day is before spot day, to invest or borrow money in home or foreign currency for just one day (or more accurate: one night). E.g., for an *over-night deposit* or *o/n deposit*, counterparty A would pay an amount N at the trading day and gets back one business day later the same amount N plus interest for one day $R_{on}\tau_{on}$ from his counterparty B. Here, R_{on} is the o/n-interest rate and τ_{on} the timeperiod until the next business day, usually one day, over weekends, it could be 3 calendar days, measured in years. Similar, a *tomorrow-next* deposit or *t/n* deposit starts on the next business day and ends

one business day later (usually spot day). These instruments are used to roll over liquidity cash positions from one day to the next and still earning interest on these cash amounts. Cash positions in foreign currencies are rolled forward in much the same way by means of *FX swaps* (see next section).

3.3   Forward Transactions

Derivative securities, or short *derivatives*, can be divided into *conditional* and *unconditional* forward transactions. A conditional forward transaction grants one party of the contract certain rights whereas the other party assumes certain obligations. In contrast, an unconditional forward transaction is an agreement that is binding on both parties. Warrants, options, and instruments similar to them, e.g., caps and floors, can be classified as conditional derivatives. Futures and forwards, on the other hand, are assigned to the unconditional derivatives.

3.3.1 Forward Rate Agreements and Forwards

Forward rate agreements (abbreviated *FRA*s) are the oldest and thus probably the most simple interest rate instruments. A forward rate agreement is an agreement between to parties to lend or borrow short-term money at a fixed rate of interest at some time (usually a few months) in the future. The agreed upon fixed rate is referred to as the *FRA rate* or *forward rate*.

   If the reference interest rate, e.g., EURIBOR, exceeds the agreed FRA rate at maturity, the buyer realizes a profit and receives a cash settlement payment from the seller. However, if the reference rate at maturity is lower, the buyer suffers a loss and compensates the seller (in form of a cash settlement). It follows that the buyer pays the FRA rate and receives in return the current interest rate from the seller.

   A FRA is usually purchased in combination with a loan or other credit obligation. The *buyer* of a FRA protects herself against *rising* interest rates. If the reference interest rate (for example, the EURIBOR) at maturity lies above the FRA rate, the buyer makes a profit (which can be used to compensate for the higher interest rate on the loan). If the reference rate lies below the FRA rate, the buyer takes a loss (but on the other hand has to pay lower interest on the loan).

   Conversely, the *seller* of a FRA protects him or herself against *falling* interest rates. He or she could for instance sell a FRA to lock in (at least for the FRA period) a certain interest yield on a floating rate investment. If the reference

rate (e.g., EURIBOR) at the FRA's maturity is lower than the FRA rate, the seller receives a cash settlement which can be used to increase the (then low) yield from the investment. If the EURIBOR at FRA maturity is higher than the FRA rate, the seller has to pay a cash settlement (but then will have a high yield from his investment).

If the reference rate at the fixing time is equal to the FRA rate, neither a profit nor a loss is realized. Two additional factors should be taken into account with respect to FRAs. Both parties of the contract face a counterparty credit risk (credit default risk). This is because depending on the interest rate level at maturity of the FRA, either party receives or is obliged to make a cash settlement payment. The second aspect is that no liquid funds are exchanged at the transaction or during the life of the contract. Only at maturity of the FRA will the contract be settled if the current interest rate does not equal the FRA rate.

If the underlying in the contract is not an interest *rate* (e.g., LIBOR, EURIBOR) but a fixed rate security (for example, a government obligation, government bond, mortgage bond, promissory note, etc.) then such a forward transaction is simply referred to as a *forward*. An *FX forward* or *FX outright* is an exchange of cash flows in different currencies, which differs from a FX spot transaction only by the fact, that the settlement date is not equal to, but later than the spot date. In the inter-banking-market, the most liquidly traded type of FX forward transactions are *FX swaps*, which is a combination of a FX spot and a FX outright. We will have a closer look at these FX transactions in Sect. 16.2.

3.3.2 Financial Futures

Although *commodity futures* have been traded on organized exchanges since 1860, *financial futures* are a rather recent addition to the market. In the US, active trading in *commodity* futures began at the beginning of the nineteenth century. It was no coincidence that the *Chicago Board of Trade* (*CBOT*) was founded as early as 1848; this was where wheat contracts were first traded. In the course of the following years, additional commodity exchanges were founded, for example in New York and London.

Trade in *financial* forward transactions is relatively recent; trading in these instruments commenced in Chicago in 1972 on the *International Monetary Market* (*IMM*). The first transactions were concluded for foreign currencies. In Germany, futures have been traded since the establishment of the *Deutsche Terminbörse* (*DTB,* now called the *EUREX*) in 1990.

The basic idea behind *futures* is identical to that of forward rate agreements and forwards. Futures are a further development in forward transactions, which go back as far as the seventeenth century. Hence, futures have the same payoff profile as forwards. The essential difference between forwards and futures is that contract elements are not individually negotiated. A future is a *standardized* forward transaction. The underlying security, the volume, the time of settlement, and other payment and delivery conditions are standardized and are set by the exchange (for example, *EUREX*, *LIFFE*). Futures are usually not exercised but closed out prior to maturity by entering into a trade opposite to the original one.

Similar to forward transactions, both parties to a futures contract take on a counterparty risk. The counterparty risk in the case of futures, however, is lower by orders of magnitude; it is in fact negligible. Futures markets have developed two mechanisms for the elimination of counterparty risk. The first involves the settlement of a potential payment not at maturity but daily. If, for example, the forward price of the futures underlying increases, the holder of a futures contract makes a profit. For a forward transaction, the profit in the form of a settlement payment cannot be realized until the maturity of the forward transaction, i.e., at the end of the forward's lifetime. For futures, in contrast, the profit is credited *directly* to a so-called *margin account* on a daily basis. Similarly, in the case of a loss, the corresponding amount is debited from the account. In the language of the futures market, these daily profit and loss payments are referred to as *daily settlement*. The daily sum credited or debited to the margin account is called the *variation margin*. Since the time to repayment is reduced to one day by the daily settlement, the counterparty risk is accordingly smaller.

In addition to the variation margin, a second mechanism reducing the counterparty risk is the *risk-based margin* which every market participant has to set aside. These are securities which must be set aside by the investors to cover potential settlement risk. Clear rules establish the amount of margin required by each exchange. These rules are based on price fluctuations associated with futures. For example, the EUREX requests payment of a so-called *additional margin* (originally called the *initial margin*) for every Bund future contract (contract size of 100,000 euros) entered into, to guarantee the ability to make any settlement payments later on. If the value of the contract increases in the course of a business day, the profit is credited to the margin account. On the other hand, if the value of the contract falls, the loss is charged to the account. Should the margin account balance fall below the additional margin, the market participant must deposit additional funds to ensure the minimum

level of the additional margin. This subsequent payment request is referred to as a *margin call.* If the market participant fails to meet this obligation, his or her position is perforce closed out. These two precautionary measures serve to further reduce the counterparty risk.

There is another difference between forward transactions and futures. While two parties close a contract directly with one another in the case of a forward, the exchange plays the role of a *clearing house* for futures and *guarantees* the settlement of the contract. The market participant thus need not estimate the counterparty risk of an arbitrary trading partner but solely that represented by the exchange itself. The counterparty risk is thus transferred to the exchange.

A further advantage of futures lies in their liquidity (market depth) in the markets. The standardization results in a concentration on trading in only a few instruments. This high liquidity leads to a lower demand-dependent price fluctuations and to lower transaction costs.

The very low credit default risk of exchanges is the primary reason why regulators enforce market participants to transfer standardized OTC trades to *central counterparties* or short *CCPs*, as set by the *European Market Infrastructure Regulation* or *EMIR*. E.g., (see Sect. 3.3.3) fall already under this regulation. CCPs reduce credit default risks by similar means as an exchange, but do not originate the trades. Market participants may settle their trades directly (as *direct member* of a CCP) or through other direct members via a CCP. The effective reduction of the number of counterparty increases further the netting efficiency, i.e. the netting of future payments to be sold or received from various transactions) , since a larger part of the portfolio is settled with a single counterparty.

3.3.3 Swaps

A *plain vanilla swap* is an agreement in which two parties agree to exchange a series of fixed interest payments for a series of floating interest payments. While the *fixed leg* of the swap has the character of a fixed rate bond, the *floating leg* refers to a variable, short-term interest rate (the 6-month EURIBOR for instance) which is "fixed" at some agreed upon future dates.

Thus, swaps combine two portfolios (the legs) and the interest is usually earned on the same nominal principal for both legs. However, different compounding and payment conventions might be specified. A standard or plain vanilla swap consists of series of coupon payments, called the *coupon strip* in exchange for a series of EURIBOR payments, the *EURIBOR strip.*

For a plain vanilla swap in EUR, both strips have the same maturity but may roll over at different dates. Typically, the fixed leg pays interest annually in accordance with the *30/360* compounding convention. The floating leg, in contrast, pays semi-annually and the *Act/360* compounding convention is used. The *TARGET holiday calendar* is used to determine bank holidays. Swaps and their most important variants will be discussed in Sect. ff in more detail.

3.4   Options

With FRAs, forwards, futures and swaps, both parties enter into a contractual *obligation*. All agreed upon payments will have to be paid (as long as no party defaults before maturity), though the payment amount may vary with the interest rate level. In contrast to this, the buyer of an *option* obtains a *right*. Options furnish the buyer with a *claim* to payment or delivery which is conditional on market events and terms agreed upon in advance. The buyer of the option then has the choice of exercising this right or allowing it to expire without exercise. The buyer will only do exercise if this would be to his advantage. If exercising has a doubtless and evident advantage to the buyer, he might have to notify the option seller explicitly to exercise the option, if exercise is evidently and obviously to the advantage of the buyer, the option could also be automatically exercised. depending on the contract. The seller of the option always has an obligation.

Options are divided into two basic types: options to *buy* and options to *sell*, referred to as *calls* and *puts,* respectively.

Options which are traded on an exchange are called *exchange traded options.* On the EUREX, for example, options on stocks, the *DAX* index, the Bobl future, and the Bund future are traded.

In addition to options traded on the various exchanges, individually designed options, called *OTC-options* (*OTC* is an abbreviation for *over the counter*), are traded directly between financial institutions and corporations rather than on an exchange. All the features of an OTC option can be freely specified between the counterparties, independent of the standardization required for options traded on an exchange. Options or *option rights* can also be embedded in other instruments. These options are referred to as *embedded options* or *embeddos*. The most common types are warrant bonds, which are composed of straight bonds and attached *warrants*. Shortly after issue the warrants are usually detached from the bonds and traded independently as securities, viz. as warrants. Additionally, there exist securities with attached

rights that resemble options. In these instruments, the attached rights also exhibit option-like properties. In this case, however, these cannot be detached from the underlying security. Examples are *callable* and *putable* bonds.

Generally speaking, an option is an agreement between two contract parties in which the seller guarantees the buyer the right to buy (call) or sell (put) an *underlying* at a previously agreed upon price either within a certain time period (*American option*) or on a certain date (*European option*). The seller of the option, guarantees this right in exchange for payment of a certain cash amount called the *option premium*. The price for which the underlying is to be bought or sold is called the *strike* or *strike price*. The date up to which (American options) or on which (European options) the option may be exercised by the buyer is called the *maturity date*.

An option is said to be *at the money* when its strike is equal to the current underlying price. If the current underlying price is such that the option holder would make money by exercising the option *now*, then the option is said to be *in the money*. If the current underlying price is such that the holder would *not* make money by exercising the option *now*, then the option is said to be *out of the money*.

Sometimes there is confusion among investors about the expressions put and call. For stocks, a long position in calls wagers on an increase in the stock price, whereas the holder of a long put position hopes for a decline in the stock price. For interest rate instruments, the situation is somewhat more complicated. For example: A bank has issued a floater and wishes to hedge itself against rising short-term rates (6-month EURIBOR). One possibility would be to take a long position in a call on the 6-month EURIBOR. The higher the 6-month EURIBOR increases, the more valuable the call. The underlying is the interest rate. The same hedging result is obtained when taking a long position in puts on an interest rate *instrument* (for example, on the 3-month EUR future). The future *loses* value if short-term rates *rise*. The further the price of the future falls the higher the value of the put on the future). The underlying of the put is the future and not the interest rate. The desired result is attained by both strategies: the bank accomplishes the hedge against rising short-term rates since it can compensate for the higher interest rates it has to pay for the floater through the profits resulting from the long call on the 6-month EURIBOR or the long put on the 3-month Euro *future.*

A *cap* is an agreement between the seller of the cap and the buyer according to which the seller pays the buyer the difference in the interest earned on a nominal with respect to a *reference* rate (e.g., a floating market rate such as EURIBOR) and that earned on an agreed upon fixed *strike* rate, should this fixed rate be *exceeded* by the reference rate. Different modes of payment are

used in practice: in some cases, the discounted payoff is made at the maturity of the option ("early"), otherwise, payment is made upon maturity of the respective EURIBOR period ("standard"). Viewed analytically, a cap can be interpreted as a series (portfolio) of European call options on a reference rate with various maturities. The buyer of a cap (long cap) profits from an increase in interest rates. Caps are employed primarily to hedge against interest rate risk arising in consequence of variable *financing* (floaters, for instance). They can also be used when speculating on rising interest rates. Moreover, variations in the term structure effect a corresponding change in the value of the option.

Floors are the counterpart of a cap. While caps limit a floating rate from above, the floor limits them from below (lower interest rate limit). A floor is an agreement between the seller of the floor (short floor) and the buyer (long floor) in which the seller pays the buyer the difference in interest on a nominal calculated with respect to a floating market rate (for example, the 6-month EURIBOR) and a fixed *strike* rate established in the floor, should the market rate fall *below* this fixed rate. Floors are employed in hedging against risks involved in floating rate *investments* (such as floating rate notes). From the analytical point of view, a floor can be interpreted as a series (a portfolio) of European put options on the reference market rate with various maturities.

Collars are a combination of caps and floors. The purchase of a collar (long collar) corresponds to simultaneously buying a cap (long cap) and selling a floor (short floor) where the strike of the floor is lower than that of the cap. With the purchase of a collar, the buyer obtains the right to receive a settlement payment from the seller should the reference rate exceed the cap rate, must however make payment to the seller should the reference rate sink below the floor rate. In this way, the buyer of a collar ensures that for him the interest rate effectively stays within a specified range whose upper and lower limits are established by the cap and floor rates, respectively. The purpose of a collar is to reduce the costs of the cap involved. The long position in the collar receives a premium through the short floor position which lowers the price of the long cap position. However, an investor can then only profit from falling interest rates up to the lower interest rate limit of the collar.

Swaptions are interest rate instruments having a swap as an underlying. They thus give the holder the right to enter into the underlying swap. The strike of the swap is quoted in the form of the specified *swap rate*.[1] Both "cash" and "swap" are typical as modes of settlement; upon exercise of the swap, either a

---

[1]Swap rates should not be mixed up with swap points, which is the difference between the FX forward rate and the FX spot rate.

cash payment is made in settlement or the underlying swap is actually entered into.

In Part III, the structure and function of these and other financial instruments will be discussed in far greater detail. Perhaps the most important point in understanding financial instruments is their valuation and in this connection, the concept of *hedging*. Hence, the most important methods of pricing and hedging will be introduced in the following Part II before we subsequently apply them to today's most common financial products in Part III.

3.5   Classification of Fixed Income Securities

The main characteristics of interest rate instruments is that they can be represented as a right to one or several payments (*cash flows*) which may be either fixed or variable and will occur at some future dates. The amount of these cash flows is determined from the length of the respective interest periods and the associated interest rates quoted according to the convention particular to the market under consideration. The following parameters may be used to define the cash flow structure of an instrument, in particular, to establish the times on which payments are to be made: valuation date, rollover date, maturity date, payment date, reset or fixing date, frequency of the coupon payments per year (e.g., 1, 3, 6 and 12 months), over-long or partial first or final period, number of interest periods, conventions (day count conventions, business day conventions, etc.).

We differentiate between the start of an interest period, the maturity of the interest period and the payment date. The start of the interest period is the date upon which interest begins to accumulate. The *maturity* is the date on which interest stops accumulating, in other words, it is the end of the interest period. The *payment date* is the date on which the payment is actually made.

The *rollover date* is the date on which the current interest period ends and the next begins. Knowledge of the rollover date is quite important since it specifies the first date as seen from today after which regular cash flows occur with a specified frequency. The interest, accrued over the full term of the interest period, is paid at the payment date. Payments dates are either identical to the period end date or are delayed by a few days, e.g. by one or two business days after period end date. Payment dates need to be adjusted according to the specified business day rules, since they must not fall on a bank holiday, to enable the bank to settle the payment on the payment day. Start and end date may also be adjusted, but this is not a requirement. As an alternative to

the specification of the roll date, every start, end and payment date might be specified explicitly, e.g. if interest periods are irregular.

Fixed Income instruments are differentiated by the term to maturity, where instruments with a term to maturity of less than 2 years belong to the *money market* and instruments with longer terms to the *capital market*. If the instrument is a security, they are called *money market security* or *capital market security*.

3.5.1 Money Market Securities

In contrast to the capital market, where securities with terms of up to 30 years or more are traded, the money market encompasses all short-term instruments, with *terms* (i.e., times to maturity) of up to one year. Occasionally, instruments with terms of up to two years are referred to as money market instruments. The money market itself can be divided into two categories, depending on whether or not the money market instrument pays regular interest. Figure 3.1 shows examples of commonly traded money market instruments.
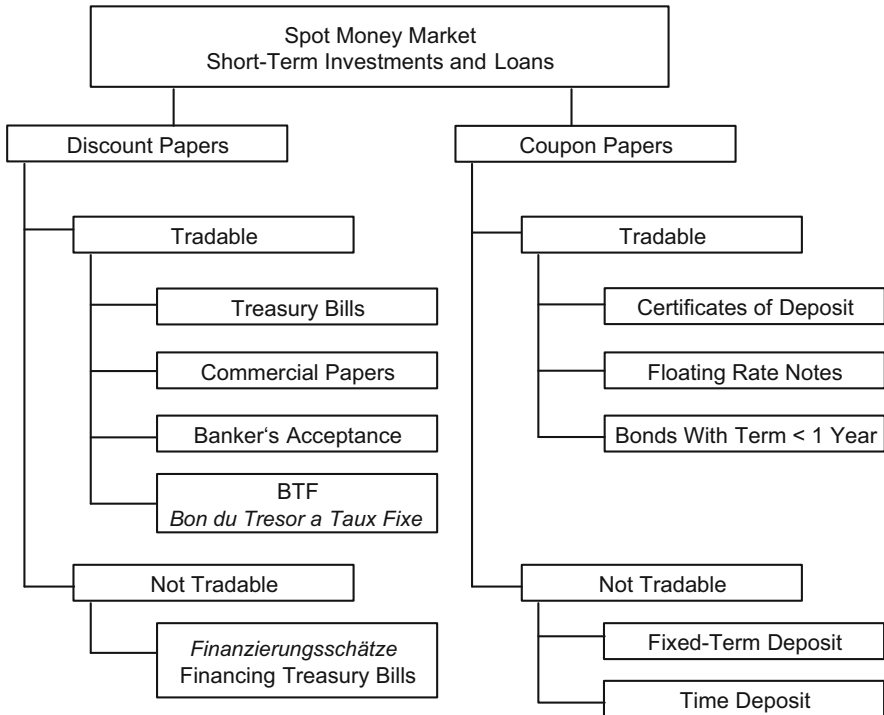
Fig. 3.1 Examples of common money market instruments

Discount Papers

Papers which are traded at a discount of their nominal value and do not regularly pay interest are referred to as *discount papers*. The interest rate at issue is determined by the difference between the issue price and the nominal value. The buyer pays the current price equal to the nominal less the discount and receives the entire nominal at maturity in return. Typical examples of discount papers are *treasury bills* and *commercial papers* in the US or the UK or *BTFs* (*bons du Tresor a taux fixe*) in France.

Commercial papers, or *CPs* for short, are short-term, unsecured, tradable obligations issued by banks and corporations. In Germany, commercial papers can be issued with terms ranging from seven days up to two years. A commercial paper program represents the general framework for an agreement between the issuer and the banks appointed to place the paper. It gives the issuer the right but not the obligation to issue short-term papers into the market at any time. Such a program has the character of a continual emission since the CP can be issued at various times in various amounts over a longer period of time. The terms of commercial papers have an upper limit of two years and are closely related to the so-called *medium-term note* program under which partial obligations with a term of at least two years are issued. The commercial paper program fills the gap to the medium-term note program.

Interest Bearing Securities

Papers having one or more regular interest payments are referred to as *interest bearing securities* or *coupon bearing securities*. They are issued at face value. *Certificates of deposit*, *floating rate notes*, *reverse floaters* and short-term capital market papers are included in this category.

Certificates of deposit (*CD*), are tradable money market papers issued by banks having a term between 30 days and 4 years. Essentially, CDs are securitized time deposits in banks.

Floating rate notes (also called *floaters* or *FRNs*) are *bonds* whose interest payments vary depending on a current market rate. The interest rates are regularly adjusted to a reference rate (for example, the *LIBOR* or *EURIBOR*), in general, to a money market rate. Hence, floaters, despite often having longer terms are also assigned to the category of money market papers. As opposed to other floaters (reverse floaters, for example), these normal floaters are referred to as plain vanilla. *Plain vanilla* is an American term for simple papers without

any particular attributes such as options or convertibility rights. These simple floaters are to be considered as floating rate notes in their most basic form.

The most important difference between a floater and a *fixed rate* instrument is the nominal interest rate. While for a fixed rate paper, the nominal interest rate is fixed over the entire lifetime of the paper, the nominal rate of the floater fluctuates. The nominal interest rate changes according to the current level of the reference market rate; a decrease in short-term rates results in a corresponding decline in the interest yield of a floater. Investors expecting a decline in money market rates will therefore usually close out their positions in plain vanilla floaters.

A more complex variant of plain vanilla floaters are *reverse floater*, sometimes referred to as *bull floating rate notes*. Like plain vanilla floaters, reverse floaters are papers paying interest at a variable rate, depending on the level of the current money market rate. Also like normal floaters, the interest is paid in regular intervals, e.g., every six months. However, while the interest yields of a plain vanilla floater increases with rising money market rates, the opposite is true of the reverse floater (hence the name). For a reverse floater, the current money market rate is *subtracted* from a fixed base rate.

Long-term papers on the capital market such as federal bonds and obligations, mortgage bonds, jumbo bonds and municipal bonds are indeed issued with longer terms than money market papers. However, the residual *time to maturity* naturally decreases constantly with the passing of time. Therefore, these papers, initially designated as capital market papers, are comparable with money market papers in the last year of their lifetime.

Time and Notice Deposits

The name *time deposit* signifies a deposit at a bank which is invested for a certain time (*fixed deposits*) or with an agreed upon term of notice (*deposits at notice*). *Fixed deposits* are deposits at a bank or savings institution with individual, fixed and unalterable maturities, amounts and interest rates, all agreed upon in advance. The conditions depend on current interest rates as well as the term of the deposit. The amount deposited also plays a deciding role; usually, the more invested, the higher the interest paid.

Investors gain somewhat more flexibility in comparison to fixed deposits by placing their savings in *deposits at notice*. For these deposits, a period is agreed upon which the investor must give notice in advance if he or she wants to withdraw the invested funds. Withdrawal notice can be one day, 48 hours, seven days or even three months. In comparison to fixed deposits, deposits at notice play a subordinate role.

Trading Conventions for Money Market Instruments

Similar to papers sold on the capital markets, the yields of money market papers are quoted publicly. *Yields* are always per year, in other words, they are *annualized*. However, different conventions for calculating yields are observed on the different international markets, as was illustrated in Sect. 2.1. The annualized yields are influenced by the day count convention and compounding method used. The day count conventions vary in the calculation of the number of days per year as well as the days of the month. The usual day count convention in the money market is the *Act/360* convention. Examples of instruments whose interest is calculated in accordance with this day count convention are commercial papers, treasury bills and floating rate notes in the US. In Germany, quotes have been published in accordance to this convention since July 1, 1990 for inter-bank trading. *Act/360* is employed elsewhere in Europe as well. Floaters are also calculated employing this method. The method *Act/360* is also referred to as the *international* or *French* method. It is frequently called the *euro interest convention* as well. Alternatively, *Act/365* is used in several markets. Sterling commercial papers, for instance, are quoted in accordance with this method; it is sometimes referred to as the *English* method.

In addition, the *30/360* convention was employed in particular in trading bonds and obligations; with the introduction of the euro, it was phased out and replaced with the *Act/Act* or *ICMA* convention. The *30/360* convention assumes that each year has 360 days, each month 30 days. The *30/360* convention is also referred to as the *German* method. The *Act/Act* method calculates with the actual number of days of the month and year. In Germany, the *30/360* convention is still applied, in particular for mortgage bonds issued before 1999. Some existing issues have already changed over to the *Act/Act* convention (for example, all government bonds and all jumbo bonds) New issues in the euro zone are computed in accordance with the *Act/Act* convention.

If yields are determined for money market instruments which pay interest several times within a single year (or in its lifetime if this happens to be less than one year in duration), a conversion into an annual yield (*annualization*) can be accomplished with or without *compounding interest effects*. Yields which do not take compounded effects into consideration (linear and simple compounding) are referred to as *nominal yields* or *nominal rates*. These methods of computing yields predominate in most money markets. We already provided a detailed description of the different compounding conventions in Sect. 2.1.

Table 3.1 Some examples of international money market instruments

Instrument	Yield calculation	Day count	Typical term on issue
Treasury bill	Discount rate	Act/360	13, 26, 52 weeks
EURIBOR	MMY	Act/360	1, 2, 3,…,12 months
Commercial Paper USD (CP)	Discount rate	Act/360	7 days to 2 years
Commercial Paper GBP (CP)	Discount rate	Act/365	7 days to 2 years
Certificate of Deposit USD (CD)	MMY	Act/360	7 days to 1 year
Federal Financing Treasuries	Discount rate	Act/Act	1 or 2 years

On the international money markets, money market papers are traded in accordance with linear (*discount rate*) or simple compounding (money market yields).

Discount papers are generally traded on the basis of the discount rate (*DR*). The discount rate is not a yield since, in contrast to a yield, it is not calculated by reference to invested capital, but on the nominal principal of the instrument. Examples of discount papers traded in foreign markets are treasury bills in the United States

The *money market yield* (*MMY*) or *CD-equivalent yield* is computed for money market instruments having interest payments. An example is the *certificate of Deposit* (*CD*). Table 3.1 provides an overview of the most important details of several money market instruments on the international money markets.

3.5.2 Capital Market Securities

Capital market papers are long-term interest rate instruments with a *term* (time to maturity) longer than one year. Papers with terms longer than 30 years are available, for instance in Germany. Typically, however, such papers have a term of ten years. The prices and yields of these long-term instruments are computed with compounded interest. Figure 3.2 illustrates some examples of common capital market papers.

Long-term interest rate instruments can also be divided into discount instruments and coupon bearing securities. The only difference between long-term to short-term interest rate instruments is that multiple interest payments are involved, instead of only one for most money market instruments.

Zero coupon bonds (or *zero bonds* for short) are tradable interest instruments which can be classified as discount securities. Like financing treasury notes, zero bonds are interest instruments without regular interest payments. The
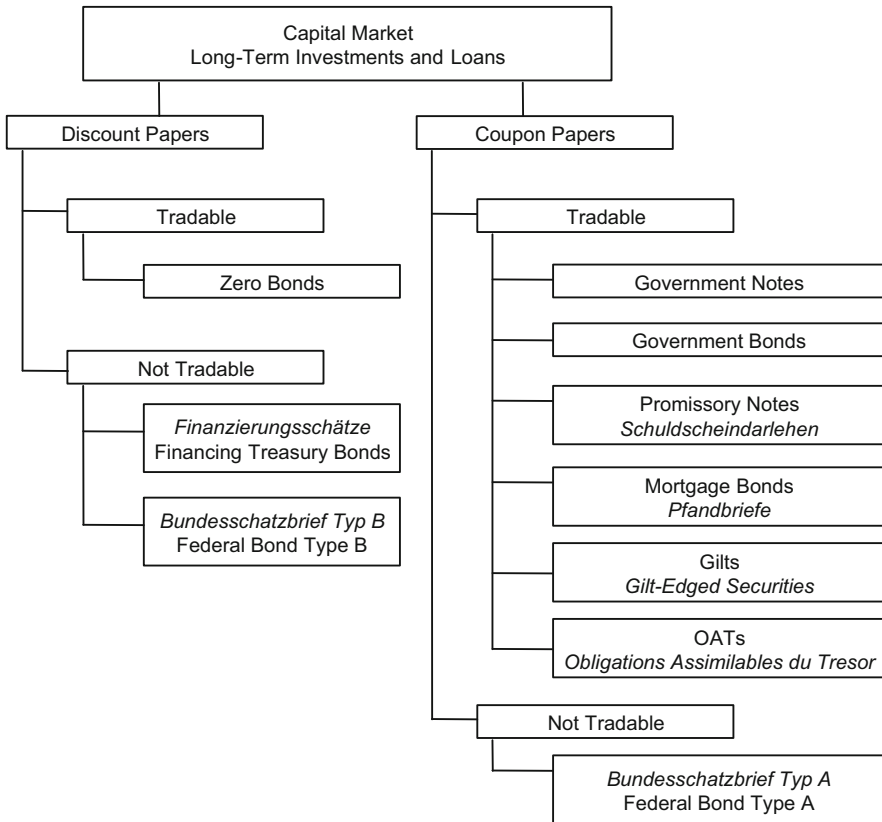
Fig. 3.2 Examples of common capital market securities

issue price is significantly lower than the redemption price. The difference between both determines the yield to maturity. We can distinguish between *zeros issued with discount*, which are issued well below par, e.g., at 60%, and *capital growth bonds*, which are repaid well above par, e.g., 230%. Zero coupon bonds with discount are issued most frequently.

The interest instruments introduced below are exclusively *straight bonds* (one exception is *Bundesschatzbrief Typ B*). Straight bonds are instruments with the following features:

- Fixed *coupon* interest rate, which is constant during time to maturity
- Redemption at par (i.e., 100%)
- Fixed life span
- Payment of the full price at issue or purchase
- No embedded option rights

Treasury notes (in Germany called *Bundesschatzanweisungen*) are securities issued by the government with maturity periods of up to two years. Treasury notes are fixed-rate coupon-bearing securities with 100% repayment at maturity. *Bundesobligationen* (or *BOBL* for short) have times to maturity of five years and are also issued by the German government. The interest is paid annually and redemption is also made at face value. Both instruments are medium-term and therefore comparable to *US Treasury notes*.

Mortgage bonds and *jumbo bonds* are interest instruments issued by mortgage banks and public sector banks used for refinancing mortgage loans. Public mortgage bonds and *municipal bonds* are issued for financing tasks of public institutions (cities, communities). Mortgage bonds and municipal bonds pay annual interest and are repaid in full at maturity.

Government bonds are securities issued by the government with terms up to 30 years, annual interest payments and redemption at par. Examples of these are the *Bundesanleihen* in Germany, *US Treasury bonds* in the United States or the French *OATs* (*obligation assimilable du Trésor*).

Schuldscheindarlehen in Germany are *promissory notes* or *loans* issued by banks, states and the federal government which are not traded on an exchange. A promissory loan is a transferable contract between the lender and borrower. A promissory loan, from a legal point of view, does not fall into the category of *securities*. The lifetime can be as long as 15 years.

On many international government bond markets the interest payment is made semi-annually. This convention can be found especially in Anglo-American bond markets. Table 3.2 gives an overview.

Table 3.2 Interest payment modes in international bond markets

National bond market	Coupon frequency	Yield computation
Australia	Semi-annual	Semi-annual yield
Belgium	Annual	Annual yield
Federal Republic of Germany	Annual	Annual yield
France	Annual	Annual yield
United Kingdom	Semi-annual	Semi-annual yield
Japan	Semi-annual	Simple yield-to-maturity
Italy	Semi-annual	Semi-annual yield
Canada	Semi-annual	Semi-annual yield
Netherlands	Annual	Annual yield
Austria	Annual	Annual yield
Sweden	Annual	Annual yield
Switzerland	Annual	Annual yield
Spain	Annual	Annual yield
United States	Semi-annual	Semi-annual yield

Part II

Methods

4

Overview of the Assumptions

To apply the common methods for pricing and risk management we need to make assumptions which are necessary for the construction of the associated models. A complete list of all model assumptions made in this book are summarized here in order to provide an overview of the numerous conditions and assumptions arising in the various methods. For each pricing and risk management method discussed in the following chapters, we will specify which of these assumptions are needed for their application. Beforehand, it should be noted that all these assumptions can at best approximate reality more or less accurate and in times of a crisis often not even that.

1. Given identical costs and risks, investors always prefer the strategy earning the greatest profit.
   This is one of the fundamental motivations for trading. No one deliberately forgoes a profit.
2. There are no arbitrage opportunities.
   Should opportunities exist allowing a profit to be made without risk and additional costs, they would be exploited by arbitrageurs until an "equilibrium" is reestablished in the market in short term.
3. The markets have infinite liquidity.
   Unlimited quantities of financial instruments are available for purchase and sale at any time without affecting their price. As is indicated in its the formulation, this assumption becomes increasingly questionable the more illiquid the market for the instrument becomes. In practice, liquidity risks

are reflected in broad bid/ask spreads. Thus, Assumption 5 usually implies Assumption 3.

4. There is no counterparty risk.

   The price of an instrument is solely a function of market parameters (prices of the risk factors, underlyings, volatility) and is in no way dependent on the instrument's seller. This is a good approximation for products traded on an exchange which also acts as a clearing house. This, of course, is not the case when dealing with OTC transactions. The quantitative treatment of credit and counterparty risks has in recent years rapidly developed. Nowadays, it is market best practice to include default risk of derivatives in valuation and risk management (see Chap. 20).

5. There is no "friction" in the market.

   Market friction is a general term referring to all costs involved in trading. These include transaction costs, bid/ask spreads, (opportunity) costs of margins, taxes, etc. All such costs not belonging to the actual *investment* are neglected. This is by no means a bad approximation for large institutions.

6. Continuous trading is possible.

   This means that the time between two trades can be arbitrarily small as can the differences in price and the number of instruments traded. The assumption that trading can take place in arbitrarily short time intervals is an approximation for the simple reason that no exchange is *always* open for trade (weekends, bank holidays). Furthermore, infinitely many adjustments in a continuous hedging strategy would cause infinitely high transaction costs if Assumption 5 did not hold. This means that Assumption 5 must hold if Assumption 6 is to make sense. The assumption that arbitrarily small price differences are possible can be problematic for options which are well out of the money. For example, the smallest possible price change of bond futures ("tick size") usually is 0.01 which is small in comparison to the futures quote, say 99.8. Therefore, the assumption of continuously changing prices is good for futures. However, for an *option* on the future which is far out of the money, a change in the futures price by one tick can result in a change in the price of the option of 50% of its value. Thus, such options do not have continuous price changes.

7. The logarithm of the relative price change of a risk factor is a random walk. In light of Sect. 2.3, this implies that the price of a risk factor is lognormally distributed.

8. Interest rates are not stochastic.

   Intuitively, this means that the evolution of interest rates over time does not involve a random component but is completely deterministic. In such a world, future interest rates are known today.

9. Interest rates are constant.
   This means that future interest rates are not only known today, they have today's value as well. Interest rates do not change in the course of time.
10. The volatilities of risk factors are not stochastic.
    Intuitively, this means that the volatilities as a function of time do not involve a random component but are completely deterministic. Future volatilities are known today.
11. The volatilities of risk factors are constant.
    This means that future volatilities are not only known today, they have the same value as today as well. Volatilities do not change in the course of time.
12. Dividend yields of risk factors are constant.
    In many pricing models for derivatives on equities, the dividend (yield) of an underlying is a parameter which is assumed to be constant throughout the term of the derivative contract.
13. Correlations between the risk factors are not stochastic.
    As with volatilities, the evolution of the correlations over time has no random component. Future correlations are known today. As in the case of volatilities, this is not often the case in reality.
14. Dividends are deterministic.
    In reality, dividends depend on the p&l progression of the company and the decision of the board of directors resp. the shareholder assembly and will be fixed at least once a year. Regardless, most models assume that time and amount of all future dividend payments are known in advance until maturity of the derivative.

5

Present Value Methods, Yields and Traditional Risk Measures

Present value methods determine the value of a financial instrument by discounting all future cash flows resulting from the instrument. Applying this method requires few assumptions. Only Assumptions 1, 2, 3, 4 and 5 from Chap. 4 are necessary.

5.1 Present Value and Yield to Maturity

Suppose that between the time t and maturity T of a security, there are n dates t_i on which cash flows $C(t_i)$, $i = 1, \ldots, n$ are due. The *present value* $V(t, T)$ of a security[1] is the sum of all *future* cash flows due between t and T, where each cash flow $C(t_i)$ is discounted at the spot rate belonging to the maturity t_i of the corresponding cash flow:

$$V(t, T) = \sum_{i=m+1}^{n} B(t, t_i)\, C(t_i) \tag{5.1}$$

where

$$\underbrace{T_E = t_0 < t_1 < \cdots < t_{m-1} < t_m}_{\text{Past}} \leq t < \underbrace{t_{m+1} < t_{m+2} < \cdots < t_n = T}_{\text{Future}}.$$

---

[1]In this book, the value of a financial instruments will usually be denoted by V.

Here, T_E denotes the security's date of issue. If all the cash flows are known, the valuation of the security follows from the current discount factors for all future payment days $t_{m+1}, t_{m+2}, \ldots t_n$. Whenever dealing with interest rate instruments, we use the notation $t_0 = T_E$ for the date of issue, t_m for the date of the last cash flow before the present value date t, and $t_n = T$ for the maturity date. Consequently, all future cash flows will occur between times t_{m+1} and t_n while all past cash flows have occurred between times t_0 and t_m.

Since the reader may be unfamiliar with the general notation for the discount factor used throughout this book, examples are provided in this section expressing the present values and sensitivities explicitly for continuous and discrete compounding. This is accomplished by substituting the general expression for the discount factor with the expressions for the desired compounding method found in Table 2.5. Substituting in Eq. 5.1 with continuous compounding yields a present value of

$$V(t, T) = \sum_{i=m+1}^{n} e^{-R(t_i - t)} C(t_i) .$$

Likewise, for annual discrete compounding

$$V(t, T) = \sum_{i=m+1}^{n} \frac{C(t_i)}{(1 + R_A(t, t_i))^{(t_i - t)}}$$

$$= \sum_{i=m+1}^{n} B_A(R_A(t, t_i), t, t_i) C(t_i), \quad \text{with} \quad B_A(v, t_1, t_2) \equiv \frac{1}{(1 + r)^{(t_2 - t_1)}} .$$

An important parameter of a bond is its *yield to maturity*, abbreviated by *YTM*. This is the compounding rate obtained if the bond is held until maturity while reinvesting all cash flows received immediately at the same rate, these also being held until maturity. If the interest rate term structure were perfectly flat until the bond matures (i.e., constant with yield $\bar{R}$) this would be equal to the yield to maturity.[2] Thus, instead of discounting using the current zero coupon bond yields, the cash flows can be discounted using the yield to maturity to obtain the current present value, called the *dirty price* of the bond. Therefore, $\bar{R}$ is given implicitly by Eq. 5.1: if all future cash flows are discounted at the

---

[2]This is generally true if and only if both yields are represented using the same compounding method, e.g. annual discrete compounding.

same yield to maturity $\bar{R}$, the resulting value must be equal to the current present value of the security, i.e.

$$\sum_{i=m+1}^{n} B_A(\bar{R}, t, t_i)\, C(t_i) = \sum_{i=m+1}^{n} B(t, t_i)\, C(t_i) \,. \tag{5.2}$$

This establishes the relationship between the YTM of the instrument and the current spot rates. In general, the yield to maturity cannot be calculated analytically from this equation; $\bar{R}$ must usually be computed using a numerical iteration procedure.

The YTM can be interpreted as the average yield of a security until maturity, but is not an appropriate measure for the performance of the security over a term shorter than the term to maturity (this is already indicated by the name "yield *to maturity*"!). If the intention is to sell the security before maturity, the value at the future settle day as well as the YTM at that time will depend on the future interest rate level and the future YTM will differ from today's YTM.

5.2 Internal Rate of Return and Net Present Value

The *internal rate of return* (*IRR*), denoted by R_E, is defined as the constant annual compounding interest rate which, when applied to *all* (including past) cash flows associated with an instrument or portfolio, makes the of the sum of all these cash flows equal to zero. Here, past cash flows are compounded at the internal rate of return while those in the future are discounted at the same rate, thus

$$0 = \sum_{i=1}^{m} \underbrace{B_A^{-1}(R_E, t_i, t)C(t_i)}_{\text{pas cash flows}} + \sum_{i=m+1}^{n} \underbrace{B_A(R_E, t, t_i)C(t_i)}_{\text{future cash flows}}$$

with $t_1 < \cdots < t_m \leq t \leq t_{m+1} < \cdots < t_n = T$

where $1 \leq m,\ m+1 \leq n$. $\tag{5.3}$

This implicit definition of the internal rate of return can be used to solve for R_E via a numerical iteration procedure. Neither spot rates nor present values appear in the above condition, only the cash flows. For instruments paying a floating rate, the cash flows are affected by the market, but for fixed

rate instruments, all cash flows are determined in advance and thus for such instruments the internal rate of return is completely independent of the current market. It is, however, influenced by the market situation at the time the instrument was purchased since the purchase price $C(t_1)$ of the instrument depends on the market situation and appears in the above equation (with a negative sign since it is a payment).

Consider for example a simple bond with a term of 6.5 years paying an annual rate of 6% on its nominal. Let's assume that the bond paid its last coupon 6 months previous to the present date and was purchased 2 months ago for 102.5% (of its nominal value) plus the accrued interest for the four months (in this example, 2%). All future cash flows (coupons and redemption upon maturity) are incoming and thus positive. In this example, the internal rate of return must be chosen so that the discounted future cash flows are equal to the purchase payment compounded at the same rate.

The internal rate of return can be interpreted as the constant interest rate at which the portfolio yield compounds the investments in the portfolio for the life of the portfolio (i.e., from purchase to sale or to the due date of the last cash flow).[3]

Starting from the above equation defining the internal rate of return, we can define a further quantity, known as the *net present value* or *NPV*. The NPV of an instrument is its current present value plus all the past cash flows compounded at the internal rate of return:

$$NPV(t, T) = \sum_{j=1}^{m} \underbrace{B_{R_E}(t_j, t)^{-1} C(t_j)}_{\text{Past Cash-Flows}} + \sum_{j=m+1}^{n} \underbrace{B_R(t_i, t) C(t_i)}_{\text{Present Value}}$$

$$\text{where } t_i < \cdots < t_m \leq t \leq t_{m+1} < \cdots < t_n = T$$

$$\text{and } 1 \leq m, m+1 \leq n \tag{5.4}$$

The difference between this and Eq. 5.3 is that *future* cash flows are not discounted at the IRR but at the *spot rates* and thus represent the present value. According to Eq. 5.3, the sum of the *past* cash flows compounded at the IRR is exactly the negative value of the sum of the *future* cash flows discounted at

---

[3]Often in definitions of the internal rate of return (and of the NPV) the future cash flows are discounted, while the past cash flows are *not* compounded. This inconsistency makes a reasonable interpretation of the NPV and IRR practically impossible. The definitions of NPV and IRR introduced here are more general and accurate.

the IRR. Thus, the NPV can be interpreted as the difference between the sum of the future cash flows discounted at the IRR on the one hand and at the spot rates on the other:

$$NPV(t, T) = \sum_{i=m+1}^{n} [B(t_i, t) - B_A(R_E, t_i, t)] C(t_i)$$

According to the definition of the YTM given in Eq. 5.2, the sum of future cash flows discounted at the YTM instead of the spot rates also yields the present value. Thus, if the YTM equals the IRR, the NPV will be zero:

> The instrument's internal rate of return equals its yield to maturity if and only if the net present value is zero.

If the IRR is higher (lower) than the YTM, the instrument will have a positive (negative) NPV.

In Table 5.1 (from the sheet IRR in the Excel workbook PLAINVANILLA.xlsx on the accompanying web page), the YTM, IRR and NPV have been calculated for a portfolio. Only the cash flows for the portfolio are required. The actual *instruments* generating these cash flows are irrelevant. In the column "Discounted Using Spot Rate", the *future* cash flows have been discounted at the current spot rates. The sum of these values yields the present value of the portfolio. In the column "Discounted Using YTM", the *future* cash flows are discounted at the YTM. In accordance with Eq. 5.2, the YTM was adjusted until the sum of the discounted cash flows was equal to the present value calculated using the current spot rates (we have utilized the Excel function "Goal Seek" to make the field "PV Difference" equal zero). In the column "Discounted/Compounded Using IRR" *all* cash flows were discounted or compounded, respectively, at the IRR, adjusting the IRR until the sum of all cash flows in this column equaled zero (see Eq. 5.3). Finally, the NPV was determined according to Eq. 5.4 as the sum of the present value and the past cash flows compounded at the IRR. This is the difference between the future cash flows discounted at the IRR and YTM, respectively. The NPV is negative and the IRR of the portfolio lies significantly under the current YTM.

The IRR and NPV are suitable for measuring the *performance* of a portfolio (or a trader), but are by no means parameters relevant to valuation or risk management. This is due to the fact that the *past* cash flows influence these parameters. But the value of an instrument (and thus its risk resulting from a

Table 5.1 Cash flow table of a portfolio

Yield curve			Portfolio cash flows			
Payment date	Time 30E/360	Spot rate	Cash flow	Discounted using spot rates	Discounted using YTM	Compounded/ discounted using IRR
31.10.99	−1,142					0
30.06.00	−0.642		−323,784			−333,439
31.12.00	−0.142		50,000			50,325
21.02.01			−189,774			−189,774
30.06.01	0.358	2.20%	50,000	49,612	48,769	49,186
31.12.01	0.861	3.20%	50,000	48,662	47,093	48,067
30.06.02	1.358	4.50%	60,000	56,518	54,590	56,382
31.12.02	1.861	5.50%	60,000	54,310	52,714	55,099
30.06.03	2.358	6.30%	60,000	51,949	50,922	53,858
31.12.03	2.861	7.10%	70,000	57,526	57,367	61,404
30.06.04	3.358	7.90%	70,000	54,225	55,417	60,022
31.12.04	3.861	8.50%	70,000	51,086	53,513	58,656
30.06.05	4.358	9.00%	−15,000	−10,303	−11,077	−12,286
31.12.05	4.861	9.30%	0	0	0	
30.06.06	5.358	9.50%	40,000	24,596	27,554	31,297
31.12.06	5.861	9.60%	0	0	0	
30.06.07	6.358	9.70%	0	0	0	
31.12.07	6.861	9.80%	30,000	15,796	18,614	21,912
30.06.08	7.358	9.90%	−15,000	−7,489	−8,991	−10,709
Values as of	YTM and IRR				7.20%	4.69%
21.02.01	Past cash flows					−472,887
	Future cash flows			446,487	446,487	472,887

potential change in its value) is independent of its history. The market price of an instrument is completely independent of the past cash flows that may have been generated by the instrument or portfolio previously. Since the risk of an instrument is by definition a reflection of the potential *future* change in the its price, the same argument holds for risk management.

For example, a single instrument could have different NPVs and IRRs, if a trader bought the instrument in the past at a cheaper price than another trader at the same time, or if two traders bought the same instrument at the same price but on different dates. It is in fact quite unlikely that the same instrument in different portfolios will have exactly the same NPV and IRR since for this to happen, the instruments must be purchased at exactly the same time and price (or the price difference must correspond exactly to compounding at the IRR for the period between the differing purchase dates).

5.3 Accrued Interest, Residual Debt and Par Rates

The *residual debt* of an instrument is the amount of outstanding debt on which interest must be paid, i.e., the open claims against the issuer of the instrument. The *accrued interest* is the outstanding interest having accumulated at a contractually agreed fixed interest rate K in the time between the last payment date t_m and t (=today). Accrued interest only makes sense when dealing with fixed rate instruments or for those instruments whose interest rates are established at the beginning of each interest period. In spite of that, it is a common requirement to estimate for accounting purposes accrued interest for variable interest coupons, which need to be adjusted with every change of the interest rate term structure. Especially, the total amount of accrued interest might even be diminished from one day to the next, though no cash flow was paid.

In our general notation

$$\text{Accrued interest} = K N(t_m)(t_m - t) \text{ with } t_m \leq t < t_{m+1} \,,$$

with $N(t_m)$ denoting the residual debt as of the last coupon payment.[4] For instruments which have no amortization payments in the time between today and maturity, the residual debt immediately after a payment date is always equal to the face value N. However, if repayments have already been made by time t_m, the residual debt is accordingly smaller.

Since a fixed interest rate K results in a debt at time t consisting of the residual debt as of the preceding payment date t_m plus the interest accrued since that time, the residual debt at time t is the residual debt from the last payment date compounded up to today at the rate K:

$$\underbrace{N(t)}_{\substack{\text{Residual Debt} \\ \text{at Time } t}} = \underbrace{N(t_m)}_{\substack{\text{Residual Debt} \\ \text{at Time } t_m}} + \underbrace{K N(t_m)(t - t_m)}_{\text{Accrued Interest}} = N(t_m)\left(1 + K(t_m - t)\right) \,.$$

---

[4]For interest rate periods shorter than one year, interest is accrued linear over the interest period, in line with the linear compounding method. For longer interest periods, interest accrued is often calculated by means of annual compounding. For very long periods (e.g. long term zero bonds), the difference can be very significant.

The *clean price* $V_{\text{clean}}(t, T)$ of a financial instrument is defined as the present value less the accrued interest based on a face value of 100, that is

$$V_{\text{clean}}(t, T) = \sum_{i=m+1}^{n} B(t, t_i)C(t_i)\frac{100}{N(t_m)} - 100K(t_m - t) .$$

An interest rate instrument (or a portfolio consisting of interest rate instruments) is quoted *at par* if its present value is equal to the residual debt of the instrument, i.e.,

$$N(t) \stackrel{!}{=} \sum_{i=m+1}^{n} B(t, t_i)C(t_i) \quad \text{where} \quad t_1 < \cdots < t_m \leq t \leq t_{m+1} < \cdots < t_n = T$$

(5.5)

or equivalently, when the clean price is equal to the residual debt at the previous payment date $N(t_m) = V_{\text{clean}}(t, T)$. This is a generalization of the usual definition for bonds and applies to all interest rate instruments and even to portfolios with arbitrary cash flows. A simple bond (with the redemption of the full nominal being made at maturity) is quoted at par if the clean price is equal to its face value.

Prices of bonds traded as securities at an exchange are in general quoted as *Clean Prices*, i.e. accrued interest needs to be added to the quoted price to get the actual price to be paid (i.e., the *Dirty Price*. To avoid any ambiguity with respect to the compounding method used, accrued interest is always linearly accrued, or, which is the same, calculated by simple compounding.

The *par rate* of a fixed rate instrument (for example, bonds, fixed rate mortgages, etc.) is defined as the rate $\bar{K}$ which the instrument should have as its fixed rate in order for the instrument to be quoted at par. Par rates are only defined for fixed rate instruments. For such instruments, the residual debt and the cash flows are dependent on K. The at-par condition given by Eq. 5.5 for $N(t)$) solved for $K = \bar{K}$ yields the par rate of the instrument as a function of the current spot rate. Of particular importance are the par rates of coupon bonds paying an annual coupon calculated using discrete compounding. Graphing these par rates as a function of time to the maturity of the coupon bonds gives a curve called the *par yield curve*.

To summarize, Table 5.2 comprises a list of all interest rates introduced in this section. The table indicates how each of the three interest rates just presented, namely the IRR, the par rate and the YTM, can be calculated from those established by the market (spot rate) and the instrument (coupon rate).

Table 5.2 Summary of the different yields

	Replaced rate	Sum of discounted cash flows equals
IRR	Spot rate	Sum of compounded past cash flows
Par rate	Fixed rate	Residual debt
YTM	Spot rate	Present value

For example, replacing the spot rate by the YTM for discounting of the cash flows will result in a value equal to the present value.

From this table it is apparent that the par rate is a function of the spot rates but *not* of the coupon rate (the par rate replaces the coupon rate). Conversely, the YTM is a function of the coupon rate and *not* explicitly of the spot rates (these are replaced by the YTM). Naturally, the YTM depends implicitly on the spot rates since the spot rates affect the present value (whose value should be attained by adding up the future cash flows discounted at the YTM). Thus, the par rate of an instrument changes when the spot rate changes, in contrast to the YTM which changes when either the coupon rate or the spot rates change. From this behavior, it follows that the par rate and the YTM can only be equal for a particular combination of spot rates (determined by the market) and coupon rates (given by the instrument), e.g. for a bond with annual coupon payments it follows:

> Par rate and yield to maturity of an interest rate instrument are equal if and only if the instrument is quoted at par.

We will demonstrate this relation in the following. As a simplification, the nominal is assumed to be constant, however, the general result would still be valid in case of nominal increases or repayments at coupon dates. The par rate $\bar{K}$ of a fixed income instrument is defined as the fixed interest rate the instrument would need to have such that its present value is equal to its residual debt (see Eq. 5.2). This relation is valid at a payment date $t = t_m$, too. On the other hand, the present value could be calculated using Eq. 5.2 by means of the YTM. Therefore, if the fixed coupon rate equals the par rate, we get at payment date $t = t_m$:

$$N = \sum_{i=m+1}^{n} B_A(\bar{R}, t_m, t_i)C(t_i) \ . \tag{5.6}$$

The cash flows $C(t_i)$ depend on the fixed rate (here equal to the par rate):

$$C(t_i) = N\bar{K} \quad \text{for } i < n$$
$$C(t_n) = N(1 + \bar{K}) \, .$$

We have used here that the coupon is paid annually, i.e. $t_i - t_{i-1} = 1$. Further,

$$B_A(\bar{R}, t_m, t_i) = \left(\frac{1}{1 + \bar{R}}\right)^{i-m} \, .$$

Setting $\bar{K} = \bar{R}$ and replacing in Eq. 5.6 yields

$$
\begin{aligned}
N &= \sum_{i=m+1}^{n} \left(\frac{1}{1 + \bar{R}}\right)^{i-m} N\bar{R} + \left(\frac{1}{1 + \bar{R}}\right)^{n-m} N \\
&= \sum_{i=m+1}^{n-1} \left(\frac{1}{1 + \bar{R}}\right)^{i-m} N\bar{R} + \left(\frac{1}{1 + \bar{R}}\right)^{n-m} N(1 + \bar{R}) \\
&= \sum_{i=m+1}^{n-1} \left(\frac{1}{1 + \bar{R}}\right)^{i-m} N\bar{R} + \left(\frac{1}{1 + \bar{R}}\right)^{n-1-m} N \, .
\end{aligned}
\tag{5.7}
$$

Repeating this step $m - n$ times, we get the final result $N = N$. This demonstrates, that Eq. 5.6 is always fulfilled, if $\bar{K} = \bar{R}$.

It should be noted that this relation is not valid, if $t \neq t_m$, in which case the left side of Eq. 5.6 would have an additional term for the accrued interest, which need to be calculated by simple compounding. Since YTM is given in discrete, annual compounding, YTM and par rate would differ slightly. For the same reason it was necessary to restrict the relation to instruments with annual coupon payments, because only for periods of exactly one year both compounding methods, annually and simple compounding, yield identical results.

5.4   Traditional Sensitivities of Interest Rate Instruments

5.4.1   Average Lifetime and Macaulay Duration

The simplest feature characterizing a security whose value derives from future cash flows is its mean (residual) lifetime. The mean is the weighted average taken over the time periods until each future coupon payment. The weight accorded to a particular period is exactly equal to the contribution of the corresponding cash flow to the present value of the security. This mean lifetime is called the *Macaulay duration.*

$$
D_{\text{Macaulay}}(t) = \sum_{i=m+1}^{n} (t_i - t) \frac{B(t, t_i)\, C(t_i)}{V(t)}
$$

$$
= \sum_{i=m+1}^{n} (t_i - t) \left[\frac{B(t, t_i)\, C(t_i)}{\sum\limits_{k=m+1}^{n} B(t, t_k)\, C(t_k)} \right]. \tag{5.8}
$$

The Macaulay duration has an interesting property: If a time interval of length dt (during which no cash flow payment was due) has passed since the last calculation of the Macaulay duration, the new Macaulay duration is simply given by the difference of the old value less the elapsed time dt.

$$
D_{\text{Macaulay}}(t + dt) = D_{\text{Macaulay}}(t) - dt \quad \text{if} \quad t + dt < t_{m+1} .
$$

The derivation of this result is trivial. We only need Eq. 2.7 and to make use of the fact that the sum over all weights equals one:

$$
D_{\text{Macaulay}}(t + dt) = \sum_{i=m+1}^{n} (t_i - t - dt) \left[\frac{B(t + dt, t_i)C(t_i)}{V(t + dt)} \right]
$$

$$
= \sum_{i=m+1}^{n} (t_i - t) \left[\frac{B(t + dt, t_i)C(t_i)}{V(t + dt)} \right]
$$

$$
- dt \underbrace{\left[\frac{\sum_{i=m+1}^{n} B(t + dt, t_i)C(t_i)}{\sum_{k=m+1}^{n} B(t + dt, t_k)C(t_k)} \right]}_{1}
$$

$$= \frac{B(t, t+dt)}{B(t, t+dt)} \underbrace{\sum_{i=m+1}^{n} (t_i - t) \left[\frac{B(t, t_i)C(t_i)}{V(t)} \right]}_{D_{\text{Macaulay}}(t)} - dt$$

$$= D_{\text{Macaulay}}(t) - dt \ .$$

5.4.2 Modified Duration and Convexity

The methods for pricing financial instruments introduced in this book are applied in risk management when determining the value of a portfolio in any given scenario. This can be extremely time consuming. In order to obtain a quick risk assessment without repeating time-consuming pricing calculations for each different scenario, *approximation techniques* have been developed which are all based on the same procedure, i.e. utilization of *Sensitivities*:

Let V be the price of a financial instrument and S the price of the underlying risk factor, i.e. a market parameter which follows some stochastic process (for instance a share price or interest rate index). Here, we will consider a single risk factor only, though in general more than one risk factor is relevant. The most important risk factor is usually the price or value of the underlying and the sensitivity of the price of the instrument with respect to changes in the underlying the corresponding risk measure. The price is generally a complex function of S (calculated by means of discounted cash flow method or option price theory or some combination of these) which, however, admits a *Taylor series* representation.

$$V(S + dS, t) = V(S, t) + \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k V(S, t)}{\partial S^k} (dS)^k \ .$$

In such a Taylor series, the price of the instrument whose underlying has changed slightly changed to $S + dS$ is calculated by adding the infinite series above to the price $V(S, t)$ corresponding to the original value of the underlying S. The terms of the sum involve all powers $(dS)^k$ of the change in the underlying dS. If this change is small (for example, 1 basis point = $0.01\% = 0.0001$), its powers are particularly small for large k (for example, $(0.0001)^2 = 0.00000001$, $(0.0001)^3 = 0.000000000001$, etc.). Thus, for small changes in S, we can obtain a good approximation of the price by

truncating the series after the first few terms.[5] In this way the infinite sum is approximated by

$$dV(S,t) = \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k V(S,t)}{\partial S^k}(dS)^k \approx \frac{\partial V(S,t)}{\partial S} dS + \frac{1}{2}\frac{\partial^2 V(S,t)}{\partial S^2}(dS)^2 \ ,$$

(5.9)

where $dV_S = V_{S+dS} - V_S$ denotes the difference between the old and new price of the instrument. In the traditional terminology of interest rate management and if S represents the yield to maturity, the coefficients of the first two powers of dS (divided by the original price V) are referred to as the *modified duration* and the *convexity*. In modern risk management, these coefficients are called *delta* and *gamma*, where S is any arbitrary underlying.

More precisely, the modified duration is defined as the (negative) derivative of the price with respect to $\bar{R}$ (YTM) divided by the price, i.e., the (negative) relative sensitivity of the security to linear changes in the yield to maturity, see Eq. 5.10. The minus sign reflects the effect that bond prices decline as interest rates rise. Assuming that the cash flows themselves are independent of the interest rate, Eq. 5.2 gives the following expression for the modified duration:

$$D_{\text{mod}} = -\frac{1}{V}\frac{\partial V}{\partial \bar{R}} = -\sum_{i=m+1}^{n} \frac{C(t_i)}{V}\frac{\partial B_{\bar{R}}(t,t_i)}{\partial \bar{R}} \ .$$

(5.10)

This is the general expression for all compounding methods. For example, for continuous compounding it reads explicitly

$$D_{\text{mod}} = -\frac{1}{V}\frac{\partial V}{\partial \bar{R}} = \frac{1}{V}\sum_{i=m+1}^{n}(t_i - t)C(t_i)\exp\left[-\bar{R}(t_i - t)\right]$$

$$= \sum_{i=m+1}^{n}(t_i - t)\left[\frac{C(t_i)\exp\left[-\bar{R}(t_i - t)\right]}{\sum_{j=m+1}^{n} C(t_j)\exp\left[-\bar{R}(t_j - t)\right]}\right] = D_{\text{Macaulay}} \ .$$

---

[5]The method of expanding a function in a Taylor series for small changes in its parameters, neglecting all but the parameter terms up to a certain power (here, the second power) is well established and frequently used in practice with verifiable correctness (assuming the variables under consideration are continuous). This is not simply an academic "mathematical trick" but a method which finds constant application in the financial world. Almost all hedging strategies as well as the variance-covariance methods used in calculating the value at risk, for example, find their basis in this procedure. The same method has already been applied when we presented Ito's lemma in Sect. 2.4.2.

The last expression allows an intuitive interpretation: it is a weighted average of the "future cash flow maturities" $t_i - t$, where the maturity of each cash flow contributes to the average with the weight specified in brackets. This weight is just the proportional contribution of the corresponding cash flow to the present value of the security. The modified duration is thus, for continuous compounding, equal to the Macaulay duration. This result makes it clear that the Macaulay duration (which was defined as the average cash flow maturity) contains information about the sensitivity of the security with respect to yield changes and can thus be used as a risk measure.

For annual compounding, the modified duration given in Eq. 5.10 becomes

$$
D_{\text{mod}} = \frac{1}{V} \frac{\partial V}{\partial R} = -\frac{1}{V} \sum_{i=m+1}^{n} (t_i - t) \frac{C(t_i)}{(1 + \bar{R})^{(t_i - t + 1)}}
$$

$$
= \frac{1}{1 + \bar{R}} \sum_{i=m+1}^{n} (t_i - t) \left[\frac{C(t_i)/(1 + \bar{R})^{(t_i - t)}}{\sum\limits_{i=m+1}^{n} C(t_j)/(1 + \bar{R})^{(t_j - t)}} \right] = \frac{D_{\text{Macaulay}}}{1 + \bar{R}}
$$

Thus, for annual compounding the modified duration is *not* equal to the Macaulay duration (the weighted average of the cash flow maturities). Instead, the Macaulay duration must be modified by a factor $1/(1 + \bar{R})$. This is the reason for the name *modified* duration. Note that a modification is unnecessary for continuous compounding. It is thus solely an effect of the compounding method applied.

Regardless of whether the compounding is computed continuously or annually, the unit of both the Macaulay duration and the modified duration is "years" since both deal with an average duration multiplied, in the case of the modified duration and annual compounding, by the dimensionless factor[6] $1/(1 + \bar{R})$.

---

[6]Actually, the interest rates should be measured in units of inverse time, since they describe a relative gain per time unit. In the case of annual discrete compounding, the divisor, i.e. the number of compounding periods per year, is usually suppressed because of its value of 1. With this divisor, which has the same unit as the interest rate, to term becomes dimensionless.

According to Eq. 5.9, the modified duration is defined precisely so that it gives a linear approximation of the relative price change when multiplied by the change in the yield to maturity:

$$dV \approx \frac{\partial V}{\partial \bar{R}} d\bar{R} = -V D_{\text{mod}} d\bar{R} \quad \Longrightarrow \quad \frac{dV}{V} = -D_{\text{mod}} d\bar{R} \ . \qquad (5.11)$$

Setting $d\bar{R} = 1\%$ *per year* makes the numerical value of relative price dV/V (measured in percent) equal to the *numeric value* of the modified duration. This provides an intuitive interpretation of the modified duration: the numeric value of the modified duration is equal to the percentage change in the price when the yield changes by 1% per year. This interpretation of the numeric value is often (incorrectly) used as the definition of the modified duration itself, though this holds only in linear approximation.

Substituting one basis point for the change in yield, i.e., $d\bar{R} = 0,01\%$ per year instead of 1% per year, we obtain the *absolute* price change $dV = -V D_{\text{mod}} dR$ referred to as the *basis point value*.

In order to obtain a more precise measure of the sensitivity of a security with respect to its yield to maturity, the linear approximation can be improved by including the second order term of the above mentioned Taylor series of the price function (Eq. 5.9) in our considerations. This term measures the sensitivity of the present value with respect to the square of the change in yield. Intuitively, it describes the curvature of the price curve and is therefore referred to as the *convexity*. Assuming that the cash flows themselves are independent of the YTM and making use of Eq. 5.2, the convexity of the present value is given by

$$\frac{1}{V} \frac{\partial^2 V}{\partial \bar{R}^2} = \sum_{i=m+1}^{n} \frac{C(t_i)}{V} \frac{\partial^2 B_A(\bar{R}, t, t_i)}{\partial \bar{R}^2} \ . \qquad (5.12)$$

This general expression holds for all compounding methods. When the interest is compounded continuously, it reads explicitly

$$\frac{1}{V} \frac{\partial^2 V}{\partial \bar{R}^2} = \frac{1}{V} \sum_{i=m+1}^{n} (t_i - t)^2 C(t_i) e^{-\bar{R}(t_i - t)}$$

$$= \sum_{i=m+1}^{n} (t_i - t)^2 \left[\frac{C(t_i) \exp\left[-\bar{R}(t_i - t)\right]}{\sum\limits_{j=m+1}^{n} C(t_j) \exp\left[-\bar{R}(t_j - t)\right]} \right] \ .$$

For annual compounding, the explicit expression is given by

$$
\frac{1}{V}\frac{\partial^2 V}{\partial \bar{R}^2} = -\frac{1}{V}\sum_{i=m+1}^{n}\frac{\partial}{\partial \bar{R}}\frac{(t_i - t)C(t_i)}{(1+\bar{R})^{(t_i - t + 1)}}
$$

$$
= \frac{1}{(1+\bar{R})^2}\sum_{i=m+1}^{n}(t_i - t)(t_i - t + 1)\left[\frac{C(t_i)/(1+\bar{R})^{(t_i - t)}}{\sum_{j=m+1}^{n} C(t_j)/(1+\bar{R})^{(t_j - t)}}\right].
$$

As is the case for the duration, the cash flow weights appear in the brackets. The convexity can thus be interpreted as the weighted average of the *squared* cash flow maturities.

As yet, the sensitivities have only been considered with respect to the yield to maturity. We could also calculate the sensitivities with respect to the current spot rates. To do so, the original form of the present value given by Eq. 5.1 is differentiated with respect to the spot rates valid for the cash flow maturities, called the *key rates*. In general, the derivative of a price with respect to the current spot rate is referred to as the *yield delta*. In analogy to the modified duration, the *key rate durations* are defined as the negative yield deltas with respect to the key rates divided by the price:

$$
D_i^{\text{key}} = -\frac{1}{V}\underbrace{\frac{\partial V}{\partial R_i}}_{\text{Yield Delta}} = -\frac{1}{V}\underbrace{C(t_i)\frac{\partial B(R_i, t, t_i)}{\partial R_i}}_{\text{Yield Delta}}. \tag{5.13}
$$

A linear approximation of the proportional change in the present value of the instrument resulting from a change in the term structure of interest rates is thus

$$
\frac{dV}{V} = \frac{1}{V}\sum_{i}\frac{\partial V}{\partial R_i}dR_i = -\sum_{i}D_i^{\text{key}}dR_i. \tag{5.14}
$$

Note that the change differs for each key rate $R_i = (t_i)$; the yield curve could be "bent". This method allows risk calculations for far more complicated scenarios in comparison to just using modified duration which, as we recall, is based on a flat term structure of interest rates at the level of the yield to maturity. The modified duration can therefore only show the effect of a parallel shift of the term structure.

5.4.3 Summation of Traditional Sensitivities

In contrast to the "modern" sensitivities (delta, gamma, vega, etc.) of options etc., which give the *absolute* sensitivity (the amount of *money* by which the value changes when a parameter is modified), the traditional sensitivities indicate a *relative* sensitivity (the *percentage* change in the value of an instrument when a parameter is modified). In consequence, modern sensitivities of instruments can simply be added up in order (obeying the sign) to obtain the total sensitivity of all instruments in a given portfolio with respect to a particular parameter (net delta, net gamma, etc.). This is not the case for the traditional sensitivities.

The second and more important difference between traditional and modern sensitivities is that the modern sensitivities refer to *market* parameters such as spot rates which are relevant for most instruments. The traditional sensitivities, however, most often refer to *instrument-specific* properties, such as yield to maturity. For example, how should we interpret the modified duration of a portfolio consisting of a 5-year bond and a 10-year bond? The sensitivity with respect to the 5-year YMT or the 10-year YMT? Despite these difficulties, it is possible to define total traditional sensitivities for an entire portfolio. To do so, we need to perform the following four steps:

1. First, a mark-to-market valuation of the entire portfolio is performed in accordance with 5.1 by discounting all future cash flows at the current spot rates.
2. The YTM of the entire portfolio is determined according to Eq. 5.2. All cash flows are discounted at the same interest rate. This rate is varied until the present value calculated in Step 1 is attained.
3. In the equations for the sensitivities of the individual instruments, the YTM of each instrument is replaced by the YTM of the entire portfolio. For example, consider the modified duration of a zero coupon bond at time t with maturity T which, in discrete compounding, is given by, "$(T-t)/(1+R)$", where R denotes the YTM of the zero coupon bond. This YTM is replaced by the YTM of the entire portfolio yielding the sensitivity of the zero bond with respect to the YTM of the portfolio.
4. Each sensitivity of a particular instrument (with respect to the YTM of the portfolio) is multiplied by the present value of the corresponding instrument. The sum of these products divided by the present value of the portfolio is the desired sensitivity of the portfolio.

The last two steps can be combined in one equation. The modified duration and convexity of a portfolio consisting of two instruments A and B with respect to the YTM $\bar{R}_{A+B}$ of the portfolio, and the key rate duration of this portfolio with respect to a spot rate R (which is the same for all instruments in the portfolio) are[7]:

$$D_{\text{mod}}(A+B)_{\bar{R}_{A+B}} = \frac{1}{A+B}\left(A\,D_{\text{mod}}(A)_{\bar{R}_{A+B}} + B\,D_{\text{mod}}(B)_{\bar{R}_{A+B}}\right)$$

$$\text{Convexity}(A+B)_{\bar{R}_{A+B}} = \frac{1}{A+B}\left(A\,\text{Convexity}(A)_{\bar{R}_{A+B}} + B\,\text{Convexity}(B)_{\bar{R}_{A+B}}\right)$$

$$D_{\text{key}}(A+B)_R = \frac{1}{A+B}\left(A\,D_{\text{key}}(A)_R + B\,D_{\text{key}}(B)_R\right)\,. \tag{5.15}$$

We have denoted the present value of instrument A by simply A, and analogously for instrument B. The situation for the key rate duration is somewhat simpler since this is a sensitivity with respect to the spot rates. Spot rates are not instrument-specific parameters and consequently the first two steps need not be performed.

Traditional sensitivities can be calculated in this way even for modern instruments such as futures contracts. A futures contract corresponds to a portfolio consisting of the cash flows of the underlying occurring subsequent to the futures contract's maturity plus the cash flow resulting from the futures contract's delivery price upon its maturity. The YTM is calculated for a portfolio with these cash flows. Each individual cash flow is regarded as a zero bond and the modified durations and/or the convexities of these zero bonds are added up according to the above equations.

---

[7]The derivation of these equations is trivial and will only be shown for the modified duration:

$$D_{\text{mod}}(A+B)_{\bar{R}_{A+B}} = -\frac{1}{A+B}\frac{\partial(A+B)}{\partial\bar{R}_{A+B}} = \frac{1}{A+B}\left(-\frac{\partial A}{\partial\bar{R}_{A+B}} - \frac{\partial B}{\partial\bar{R}_{A+B}}\right)$$

$$= \frac{1}{A+B}\left(\underbrace{-\frac{1}{A}\frac{\partial A}{\partial\bar{R}_{A+B}}}_{D_{mod}(A)}A + \underbrace{-\frac{1}{B}\frac{\partial B}{\partial\bar{R}_{A+B}}}_{D_{mod}(B)}B\right)\,.$$

6

Check for
updates

Arbitrage

Arbitrage considerations alone are sufficient for deriving relations such as the put-call parity or determining forward prices. Such arguments require only very few assumptions; we need only Assumptions 1, 2, 3, and 5 from Chap. 4 to be satisfied. For simplicity, we will further assume the absence of credit default risk, i.e. require that Assumption 4 is also satisfied.

6.1 Forward and Futures Contracts

Forward and *futures contracts* are transactions in which the purchase or sale of an underlying S at a later date T for a fixed price, called the *delivery price K*, is agreed upon as of the current value date t. In this book, the value of such a contract is denoted by $f_S(t, T, K)$ or, for futures contracts (see below), by $F_S(t, T, K)$. If the delivery price at time t is chosen such that the value of the forward or futures contract is zero, then the delivery price is also referred to as the *forward price*. Throughout this book, the forward price of an underlying with a price $S(t)$ at the current time t will be denoted by $S(t, T)$ where the later time T is the maturity of the corresponding forward contract.

6.1.1 Forward Price and Cash and Carry Arbitrage

The following *Cash & Carry Arbitrage* can be used to determine the value of a forward contract $f_S(t, T, K)$ and subsequently the forward price

$S(t, T)$ of an underlying with a spot price $S(t)$. Consider the following two portfolios:

- Portfolio structure
  - Portfolio **A** consists of a forward contract on an underlying for which the delivery price K is to be paid at time T and a cash sum invested at an interest rate R to yield precisely the delivery price at time T.
  - Portfolio **B** consists of the underlying and a loan in the amount of the present value of the dividends to be received at time t' between the present time and time T.

- Value of the portfolio at time t
  - Portfolio **A** at time t is worth $f_S(t, T, K)$, the value of the forward contract plus the value $B(t, T)K$ of the cash.
  - Portfolio **B** at time t is worth $S(t)$, the value of the underlying, plus the negative (since it is a loan) present value of the dividend payment to be made at the (later) time t', i.e., $-B(t, t')D(t')$. The total value of the portfolio is thus the dividend-adjusted spot price as given in Eq. 5.3. (The argument is completely analogous for the case of a dividend yield).

- Value of the Portfolio at time T
  - Portfolio **A** at time T is worth $S(T) - K$, the value of the forward contract, plus $B(t, T)KB^{-1}(t, T) = K$, from the cash, thus having in total a value of $S(T)$. In Portfolio **A**, the delivery price of the forward contract is paid using the cash and the accrued interest earned. The portfolio now consists of the underlying at its current spot price $S(T)$.
  - Portfolio **B** at time T is worth $+D(t')B^{-1}(t', T)$, the value of the underlying plus the dividend paid at t' and then compounded up to time T and, in addition, the amount $-B(t, t')D(t')B^{-1}(t, T) = -D(t')B^{-1}(t', T)$ from the loan. Thus the whole portfolio is worth in total $S(T)$. In Portfolio **B**, the loan is repaid using the dividend payment received at time t' and invested up to time T. As for Portfolio **A**, Portfolio **B** contains exactly the underlying at time T.

Thus, both portfolios have the same value at time T. Their value at earlier times must therefore also be equal since otherwise the investor would be presented with the following arbitrage opportunity: if Portfolio **A,** for example, was worth more than Portfolio **B** at time t, an investor could sell **A** ("go short") and buy **B** with the proceeds keeping the difference. At time T, the value of both portfolios must be the same, as demonstrated above. Thus the short

position can be offset exactly by the long position without any additional payment. Thus, a profit can be made without risk or capital by investing according to this strategy at time t. Analogously, if **B** was worth more than **A,** an investor could (short) sell **B** and buy **A.** In order to eliminate such arbitrage opportunities, the value of both portfolios at time t must be equal:

$$f_S(t, T, K) + K B(t, T) = \overbrace{S(t) - B(t, t')D(t')}^{\tilde{S}(t,T)}$$

$$\Rightarrow \quad f_S(t, T, K) = \tilde{S}(t, T) - K B(t, T) .$$

The *forward price* $S(t, T)$ of the underlying is now defined as the delivery price K for which the forward contract is worth nothing at time t. This means that for $K = S(t, T)$ we have $f_S = 0$ in the above equation. Thus we obtain for the forward price the explicit expression:

$$S(t, T) = \frac{\tilde{S}(t, T)}{B(t, T)} = \begin{cases} \dfrac{S(t)}{B(t, T)} & \text{No Dividends} \\[2ex] \dfrac{S(t)}{B(t, T)} - \dfrac{D(t')}{B(t', T|t)} & \text{Dividend Payment } D \text{ z. Zt. } t' \\[2ex] \dfrac{e^{-q(T-t)}}{B(t, T)}S(t) & \text{Dividend Yield } q \\[2ex] \dfrac{1 - D(t')}{B(t, T)}S(t) & \text{rel. Dividend } D(t') \text{ at time } t' \end{cases} ,$$

(6.1)

where in the case of dividend payments Eq. 2.7 is used. In this case the forward discount factor for the period between the dividend date t' and the maturity T is needed. For compounding the relative discrete dividend payments $D(t')S(t, t')$ we have used the simple expression

$$\frac{D(t')S(t, t')}{B(t', T|t)} = \frac{D(t')S(t)}{B(t, t')B(t', T|t)} = \frac{D(t')S(t)}{B(t, T)} .$$

In any case, the forward price depends on the spot price, the interest rates and the dividend only, but does not depend on how the spot price evolves until T. Instead, the forward price just reflects the costs for replicating the future cash flows, i.e. the refinancing costs of the contract.

Conversely, using the forward and spot prices the present value $d(t, T)$ at time t of dividend payments or dividend yields accrued in the time period

$T - t$ can be determined. These are referred to as *implied dividends* and *implied dividend yields*, respectively:

$$d(t, T) \equiv S(t) - \tilde{S}(t, T) = S(t) - B(t, T)S(t, T)$$

$$B_q(t, T) = e^{-q(T-t)} \equiv \frac{\tilde{S}(t, T)}{S(t)} = B(t, T)\frac{S(t, T)}{S(t)} \; . \tag{6.2}$$

6.1.2 The Stochastic Process for the Forward Price

Having established the forward price as a function of the risk factor S as given by Eq. 6.1, we make the Random-Walk-Assumption 2.17 and proceed as in Sect. 2.4.2, applying Ito's lemma (see Eq. 2.21) to obtain a stochastic process for the forward price $S(t, T)$. Making use of Eq. 6.1 we define in the case of a dividend yield, here in terms of continuous compounding

$$f(y, t) := S(t, T) = \frac{B_q(t, T)}{B(t, T)}e^{y(S)}, \qquad \text{with} \quad y(S) = \ln(S) \; .$$

The derivatives needed for the application of Ito's lemma are given by

$$\frac{\partial f}{\partial y} = f = S(t, T), \quad \frac{\partial^2 f}{\partial y^2} = f = S(t, T)$$

$$\frac{\partial f}{\partial t} = e^{y(S)}\frac{\partial}{\partial t}\left[\frac{B_q(t, T)}{B(t, T)}\right] = S(t)\frac{\partial}{\partial t}e^{(r-q)(T-t)} = -(r - q)S(t, T) \; .$$

Substituting this in the Ito formula yields the process for the forward price

$$dS(t, T) = \left(S(t, T)\mu + S(t)\frac{\partial}{\partial t}\left[\frac{B(q, t, T)}{B(t, T)}\right] + \frac{1}{2}\sigma^2 S(t, T)\right)dt + S(t, T)\sigma \, dW$$

$$= S(t, T)\left(\mu - r + q + \frac{1}{2}\sigma^2\right)dt + S(t, T)\sigma \, dW \; . \tag{6.3}$$

In anticipation of later chapters we note the following: we will later show that the freedom of arbitrage condition enforces for the valuation of derivatives a particular choice for the drift μ. This choice will be referred to as the "risk-neutral" value (see for example Eq. 9.25). With this choice for the drift, the process for the forward price is reduced to

$$dS(t, T) = S(t, T)\sigma \, dW \; . \tag{6.4}$$

This indicates that the forward price process is drift-free in a risk-neutral world. It is worth mentioning that stochastic term of the forward price process is identical to the stochastic term of the spot price process. This important result will be used frequently in the following.

6.1.3 Forward Positions

As we have seen, $S(t, T)$ is the delivery price of a forward contract agreed upon at time t with maturity T and is referred to as the forward price. This is *not* to be confused with the value of a forward contract (which upon conclusion of the contract is equal to zero, for example). The value at time t of a forward contract entered into at some earlier time $t_0 < t$ with a forward price $K = S(t_0, T)$, valid at that time, can be obtained by considering the following argument:

   Entering into a second forward contract (in addition to the first forward contract as described above) at time t for the *sale* of the underlying at maturity T at the current forward price $S(t, T)$ has the effect that at maturity T the underlying is bought for the price K (on the basis of the first forward contract) and immediately sold at the price $S(t, T)$ (on the basis of the second contract). The cash flow at maturity is thus $S(t, T) - K$. This value discounted back to time t gives the value of the portfolio consisting of both forward contracts at time t. Since the value at time t of a forward contract with a delivery price $S(t, T)$ is zero, the value of the portfolio is identical to the value of the forward contract with a delivery price K. Upon consideration of Eq. 6.1, we conclude that

$$f_S(t, T, K) = [S(t, T) - K] B(t, T) = \tilde{S}(t, T) - K B(t, T) \qquad (6.5)$$

6.1.4 Futures Positions and Basis Risk

For valuation purposes, the essential difference between a forward and a *futures* contract is that futures are instruments which are traded on an exchange. The exchange, taking on the role of a clearing house, demands that margins be paid. The *variation margin* has the effect that the daily fluctuations in the value of a futures position (unrealized profit and loss) are credited or debited to a *margin account* which must be settled at the end of the trading day. Fluctuations in the forward prices therefore result in an *immediate* cash flow rather than one paid at maturity T. For this reason, the cash flow—in contrast to that of a forward contract—is *not* discounted in the valuation of a futures contract. The value

of a futures position with the same data as the above defined forward contract is simply:

$$F_S(t, T, K) = S(t, T) - K = B^{-1}(t, T)\tilde{S}(t, T) - K \ . \qquad (6.6)$$

The forward price $S(t, T)$ is, strictly speaking, the delivery price of a forward contract concluded at time t with maturity T. It can be shown that delivery prices of futures are equal to those of forwards if interest rates are assumed to be deterministic. In this book, it will almost[1] always be assumed that this (i.e., Assumption 8 from Chap. 4) is the case when dealing with forwards and futures, i.e., the delivery price for both contracts will be given by Eq. 6.1.

The difference between the spot and delivery price is called the *basis*.[2] The basis $b(t, T)$, as can be seen from Eq. 6.1, depends on the interest rate and the dividends of the underlying:

$$b(t, T) = S(t, T) - S(t) = \frac{\tilde{S}(t, T)}{B(t, T)} - S(t) \ .$$

Basis trading is the attempt to make a profit from the fluctuations in the basis. The value of a portfolio consisting, for example, of a long futures contract with a delivery price K and a short position in the underlying itself is given by

$$F_S(t, T, K) - S(t) = S(t, T) - K - S(t) = b(t, T) - K \ .$$

Since K is constant, the change in the value of this portfolio result solely from the change in the basis. The only risk involved in holding such a portfolio results directly from the basis alone and as such is referred to as the *basis risk.*

6.2   Options

6.2.1 Upper and Lower Bounds for Option Prices

A *plain vanilla European option* gives the holder the same rights as a forward contract but *without* the obligations of the forward contract. For a long

---

[1]Except for the chapters dealing with stochastic interest rates, i.e. Chaps. 13 and 14.

[2]In general, *basis* denotes a most often very "small" difference between two similar prices, indices or other key values and is used in many different situations.

forward position, for example, the underlying *must* be bought at the price K agreed to in the contract even if it is available at a lower spot price $S(T)$ at maturity T. The associated call option, however, will be exercised by the holder if doing so does not result in a loss, i.e., the option will be exercised only when $S(T) > K$. This is why the price of a plain vanilla European option is at least as high as that of the corresponding forward position. The price of an *American option*, owing to its additional right allowing the holder to exercise at any time up to maturity, is at least as high as its *intrinsic value* (proceeds obtained by exercising, i.e., payoff profile). Besides, the right to exercise the option early can only increase its value in comparison to its European counterpart. These relationships are summarized in Table 6.1, using Eq. 6.5 for the value of a forward position.

Throughout this book we will denote the value at time t of a European call and put on an underlying S with strike price K and maturity T by $c_S(t, T, K)$ and $p_S(t, T, K)$, respectively, as in Table 6.1. For American options, capital letters will be used in the notation, i.e., $C_S(t, T, K)$ and $P_S(t, T, K)$.

Any violation of the relations listed in Table 6.1 results in an arbitrage opportunity which can be realized by selling the over-valued instrument and buying the other instrument with the proceeds. If, for example, the price of a call were higher than that of the underlying (the last relation in the first line), an investor could sell the call and buy the underlying with the proceeds, keeping the difference. If the call is exercised, the investor delivers the underlying and receives (in addition to the money already earned from the difference) the strike price K in exchange. If the call is not exercised, the investor receives (again, in addition to the money already earned from the difference) the amount earned from the sale of the underlying at the spot price at maturity.

Because of the last relation in the second row in Table 6.1, a put can be worth no more than the strike price K (this occurs when the value of the underlying is zero). This fact will prove to be important when deciding whether to exercise an American option before maturity.

Table 6.1 Boundaries for the values of plain vanilla options

	European
Call	$\max\left\{0, \tilde{S}(t, T) - K\, B(t, T)\right\} \leq c_S(t, T, K) \leq C_S(t, T, K)$
Put	$\max\left\{0, K\, B(t, T) - \tilde{S}(t, T)\right\} \leq p_S(t, T, K) \leq P_S(t, T, K)$
	American
Call	$\max\{0, S(t) - K\} \leq C_S(t, T, K) \leq S(t)$
Put	$\max\{0, K - S(t)\} \leq P_S(t, T, K) \leq K$

6.2.2 Early Exercise of American Options

For American options, the right to exercise can be exercised over the entire term of the option. In the following, we would like to answer the question when it is optimal to exercise the option before option maturity.

An American option must always be at least as worth as the corresponding Forward contract (the option holder has more rights, but no obligations). Therefore, it follows with Eq. 6.5 that the following relations hold for American Call resp. Put options:

$$C_S(t, T, K) \geq f_S(t, T, K) = \tilde{S}(t, T) - B(t, T)K$$

$$P_S(t, T, K) \geq -f_S(t, T, K) = -\tilde{S}(t, T) + B(t, T)K .$$

The criterion for the early exercise of American options is that the proceeds gained by exercising be greater than the value of the option, i.e., $S(t) - K > C_S(t, T, K)$ and $K - S(t) > P_S(t, T, K)$ for a call and put, respectively. Inserting this condition into the above relation and using the spot price adjusted for dividend payments Eq. 6.2, the early exercise condition for the call is

$$S(t) - K > \tilde{S}(t, T) - B(t, T)K$$
$$= \underbrace{S(t) - d(t, T)}_{\tilde{S}(t,T)} - B(t, T)K .$$

The procedure is analogous for the put. Canceling the spot price on both sides, we get the following constrained on the dividend of the underlying for optimal early exercise:

$$d(t, T) > K [1 - B(t, T)] \quad \text{for Calls} \tag{6.7}$$
$$d(t, T) < K [1 - B(t, T)] \quad \text{for Puts} .$$

The strike price K, invested in risk-free securities from t until maturity T, earns an interest yield of $K \left[B^{-1}(t, T) - 1 \right]$. The present value of this amount at time t is given by

$$B(t, T)K \left[B^{-1}(t, T) - 1 \right] = K [1 - B(t, T)] .$$

In view of these considerations, Eq. 6.7 can be intuitively interpreted as follows:

- American puts should be exercised prior to maturity if the present value $d(t, T)$ of the dividend (which one forgoes upon early exercise) is less than the interest yield earned on the strike price (obtained when exercising early).
- American calls should only be exercised if the present value of the dividends (which are gained by exercising early) is greater than the interest yield earned on the strike price (which is lost upon early exercise). Since the interest yield on the strike price is always greater than or equal to zero, **calls on underlyings which pay no dividends should never be exercised before maturity**.

6.2.3 Relationships Between Puts and Calls

As an example of the effect an arbitrage-free market has on options, consider the following two portfolios:

- Portfolio structure
  - Portfolio **A** consists of a European call (long) with strike price K and maturity T, and a European put (short) also with strike price K and maturity T, both on the same underlying S.
  - Portfolio **B** consists of a forward contract (long) to buy the underlying upon maturity T at the delivery price K.
- Value of the portfolio at time t
  - Portfolio **A** has at time t a value of $c_S(t, T, K) - p_S(t, T, K)$ from the long call and the short put.
  - Portfolio **B** has at time t a value of $f_S(t, T, K)$ from the long forward.
- Portfolio value at time T
  - Portfolio **A** has at time T a value of $S(T) - K$ from exercising the call if $S(T) > K$ or $-(K - S(T))$ from exercising the put if $S(T) < K$. In either case the portfolio is worth $S(T) - K$.
  - Portfolio **B** has at time T a value of $S(T) - K$ from the maturity of the forward.

If both portfolios have the same value at time T, they must be worth the same at earlier times t as well to prevent arbitrage opportunities. Setting the

values of the two portfolios equal yields the famous *put-call parity* for European options:

One interpretation of this relation is: a European call is the same as a forward contract on the underlying plus a guarantee (the put) that the underlying is worth no less than the agreed upon delivery price at maturity. If, for example, an investor is certain that the underlying price will rise and wishes to make the most out of this information, he or she should buy forward contracts and not calls since by doing so, the (in the investor's opinion unnecessary) insurance premium guarding against a falling price is saved.

If the strike price K is equal to the current delivery price $S(t, T)$, the value of the forward equals zero. From the put-call parity we can immediately conclude that in this case the price of the put must equal the price of the call.

A put-call parity for *American* options can be derived using similar arbitrage arguments. The possibility of exercising early has the consequence that it can be expressed only in terms of an inequality. Using Eq. 6.5 for the value of a forward contract, the parities can be expressed as follows:

$$
\begin{aligned}
\text{European} \quad & c_S(t, T, K) - p_S(t, T, K) = \tilde{S}(t, T) - K\, B(t, T) = f_S(t, T, K) \\
\text{American} \quad & \tilde{S}(t, T) - K \leq C_S(t, T, K) - P_S(t, T, K) \leq S(t) - K\, B(t, T) \,.
\end{aligned}
\tag{6.8}
$$

A further relation between European puts and calls is known as the *call-put symmetry*:

$$
c_S(t, T, K) = \frac{K}{S(T)} p_S\left(t, T, \frac{S(T)^2}{K}\right),
$$

As the put-call parity, this relation is easily obtained by observing that the payments at maturity on the left and right-hand side of the above equation are the same. This symmetry can in theory be used, for example, to hedge exotic options (barriers and look-backs). Note however, that determining the strike price of the put and the number of puts required in the portfolio in the above equation requires that the future *spot* price $S(T)$ of the underlying at maturity (not the forward price $S(t, T)$!) must be known. This makes the call-put symmetry difficult to apply in practice.

7

The Black-Scholes Differential Equation

Having used arbitrage considerations to derive various properties of derivatives, in particular of option prices (upper and lower bounds, parities, etc.), we now demonstrate how such arbitrage arguments, with the help of results from stochastic analysis, namely Ito's formula 2.22, can be used to derive the famous Black-Scholes equation. Along with the Assumptions 1, 2, 3, and 5 from Chap. 4, the additional assumption that continuous trading is possible is essential to establishing the equation, i.e., in the following we assume that Assumption 6 from Chap. 4 holds. For simplification, we begin with the case without credit default risk, i.e. we make the additional Assumption 4 from Chap. 4. The Black-Scholes equation is a partial differential equation which must be satisfied by *every* price function of path-independent European derivatives on a single underlying.[1] Consequently, one method of pricing derivatives consists in solving this differential equation satisfying the boundary conditions corresponding to the situation being investigated. In fact, even quite a number of path-*dependent* options obey this differential equation. A prominent example is the barrier option. In general however, the price of path-dependent options cannot be represented as a solution to the Black-Scholes equation. It is possible to surmount these difficulties by imbedding the state space in a higher dimensional space defining one or several additional variables in an appropriate manner to represent the different paths.

---

[1]Path independence is necessary for the derivative's value V to be only a function of t and the current underlying price S. Path-dependent prices V are functions of further variables which themselves are dependent on the history of the underlying. The total differential of V is then no longer given by Eq. 2.22; additional terms must be included in the expression.

This method is demonstrated explicitly by Wilmott for Asian options with arithmetic means [192]. As we will see below, the valuation of American options can also be accomplished via the Black-Scholes equation (with free boundary conditions).

The Black-Scholes equation can be derived in several different ways. The derivation presented here is the most intuitive and, being based on arbitrage arguments, is somewhat less technical (with the exception of Ito's lemma) since the arguments are largely economical in nature. In addition to presenting its derivation, we will establish its relationship to an equation called the diffusion equation or heat equation and then illustrate a mathematically deeper derivation of the Black-Scholes equation using the backward equation 2.35. Another approach to the Black-Scholes equation which does not involve an application of Ito's lemma will later be presented in Sect. 12.1 for the derivation of Eq. 12.5. For the sake of consistency with the situation considered there, we will here investigate derivatives on underlyings with a dividend yield q.

7.1   Derivation of the Black-Scholes Equation from Arbitrage Arguments

We denote the value at time t of a derivative on an underlying S by $V(S, t)$. Here, as is usually the case in this book, the letter V is used as the general notation for the price (*value*) of a derivative without further specifying the derivative. The underlying is governed by a stochastic process satisfying Eq. 2.19. Without loss of generality, we write the mean return and the volatility in terms of relative (with respect to S) parameters $\widetilde{\mu}$ and σ, i.e.,

$$a(S, t) = S \, \widetilde{\mu}(S, t)$$
$$b(S, t) = S \, \sigma(S, t) \, , \qquad\qquad (7.1)$$

so that the stochastic process has the form

$$dS = \widetilde{\mu}(S, t)S \, dt + \sigma(S, t)S \, dW \, . \qquad\qquad (7.2)$$

The well-known process given by Eq. 2.24 is a special case of the one introduced here where the price yield (drift) $\widetilde{\mu}$ and volatility σ are constant. The process under investigation here is just a reformulation of the Ito Process Eq. 2.19 and is thus just as general as 2.19.

7.1.1 The Black-Scholes Equation for European Options

The value V of a derivative on this underlying is a function of a stochastic variable and is therefore also stochastic. In accordance with Ito's lemma 2.22, V satisfies the following equation in which the term containing dS represents the stochastic component of the change in the value of the derivative:

$$dV(S,t) = \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2(S,t)S^2 \frac{\partial^2 V}{\partial S^2} \right) dt + \frac{\partial V}{\partial S} dS \ . \tag{7.3}$$

A portfolio can now be constructed consisting of the derivative $V(S,t)$ and underlyings where the number of underlyings in the portfolio is denoted by Δ. We denote the value of this portfolio by Π

$$\Pi = V(S,t) + \Delta S \ . \tag{7.4}$$

The change in the underlying during a small time interval of length dt (analogous to Eq. 12.3) results from the stochastic change dS and the dividend yield accumulated at a rate q during this time interval:

$$d(\Delta S) \approx \Delta dS + \Delta S \, q(t) dt \ ,$$

where the linear approximation of the continuous compounding dividend yield was used. The total differential of the entire portfolio Π is thus

$$d\Pi = dV + d(\Delta S) = \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2(S,t)S^2 \frac{\partial^2 V}{\partial S^2} + \Delta Sq(t) \right) dt + \left(\frac{\partial V}{\partial S} + \Delta \right) dS \ . \tag{7.5}$$

Choosing the number of underlyings in the portfolio such that

$$\Delta \overset{!}{=} -\partial V/\partial S \ , \tag{7.6}$$

has the effect that the coefficient of the stochastic term dS vanishes in the above expression for the change in the portfolio process. Thus, a portfolio can be constructed in such a way that the change in its value is generated solely by

a progression dt in time. This is called a *delta hedge.* The total differential of the corresponding portfolio is now given by

$$d\Pi = \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2(S,t)S^2\frac{\partial^2 V}{\partial S^2} - q(t)S\frac{\partial V}{\partial S} \right) dt \ .$$

This no longer contains a stochastic term. It follows that Π is risk-free and as such increases at the same rate as given by the risk free interest rate $r(t)$, otherwise arbitrage would be possible. Thus,

$$d\Pi \overset{!}{=} \left[B^{-1}(t, t+dt) - 1 \right] \Pi \, dt \approx r(t)\Pi \, dt = r(t) \left[V(S,t) - \frac{\partial V}{\partial S}S \right] dt \ ,$$

where the definition of the portfolio's value, Eq. 7.4, is substituted in the last step. Equating the two expressions for $d\Pi$ immediately yields the partial differential equation which must be satisfied by $V(S,t)$:

$$\underbrace{\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2(S,t)S^2\frac{\partial^2 V}{\partial S^2} - q(t)S\frac{\partial V}{\partial S}}_{\text{Yield from a }\Delta\text{-Hedged Portfolio}} = \underbrace{r(t) \left[V - S\frac{\partial V}{\partial S} \right]}_{\text{Yield from a Bank Account}} . \qquad (7.7)$$

This is the famous *Black-Scholes differential equation.* From the arbitrage argument used in its derivation, we immediately see the intuitive meaning of this differential equation, namely that the yield (per time) earned from a delta-hedged portfolio must be equal to the yield earned from a risk-free investment (*bank account*) in the amount of the value of the portfolio. This interpretation is very useful in understanding and working with the Black-Scholes equation. In the literature, the equation is usually written in the following form:

$$\frac{\partial V}{\partial t} + [r(t) - q(t)]\, S\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2(S,t)S^2\frac{\partial^2 V}{\partial S^2} - r(t)V = 0 \ . \qquad (7.8)$$

Note that no reference has been made to the nature of the derivative in the above derivation (with the exception of path independence and the European mode of exercise). An assumption about the underlying, however, has been made, namely Eq. 7.2. Consequently, *all* European, path-independent derivatives on underlyings given by the general stochastic process 7.2 satisfy the above differential equation. The coefficients r, q and σ are functions of t or S or of both.

In general, differential equations have an infinite number of quite different solutions. Thus the prices of quite different derivatives can satisfy the same differential equation. Introducing the additional requirement that *boundary conditions* be satisfied as well as the differential equation, reduces the number of solutions from an infinite number to exactly one. Seen from this perspective, different derivatives (on the same underlying) are distinguished only by requiring different boundary conditions to be satisfied. As the name suggests, boundary conditions are conditions which the solution of the differential equation must satisfy at the boundary of the set on which it solves the differential equation. This set is a subset of the solution's domain of definition. In our case, the arguments of the solution are the variables S and t and the associated set is a subset of the plane spanned by S and t. For example, for a plain vanilla call or put, the set lies between $S = 0$ and $S = \infty$ and between $t =$ today and $t = T$.

The boundary condition in the time coordinate is called the *initial condition*. The initial condition for an option is that the value $V(S, T)$ of the derivative at maturity T must be given by its payoff P (which is a function of the price of the underlying)[2]:

$$V(S, t = T) = P(S) . \tag{7.9}$$

The payoff profile $P(S)$ is established in the derivatives contract.

The boundary conditions in the S-coordinate are often obtained from asymptotic considerations (for example, the value of a plain vanilla call is zero at $S = 0$). For some exotic options, boundary conditions in the S-coordinate are explicitly specified in the option contract. For example, a knock-out option contract specifies that the value of the option is zero if S attains a barrier H, in other words when $S = H$.

7.1.2 The Black-Scholes *In*equality for American Options

American options differ from European options in that the holder has the right to exercise the option *before* maturity. This additional right, under certain circumstances, makes the American option more valuable than its European

---

[2]The fact that the "initial condition" in our case belongs really to the *end* of the time period under consideration is just a question of vocabulary and should not confuse the reader. Alternatively one can introduce a new time variable $\widetilde{t} := T - t$ (sometimes called *term to maturity*). Then the initial condition belongs to the lowest value of $\widetilde{t}$, namely $\widetilde{t} = 0$.

counterpart. We will now show the variant of the Black-Scholes differential equation that American options must satisfy.

The change in the option's price is again given by the total differential equation 7.3. The stochastic component is contained in the dS-term only. We again construct a portfolio Π consisting of the derivative $V(S, t)$ and Δ underlyings:

$$\Pi = V(S, t) + \Delta S \, .$$

The total differential of this portfolio is again of the form 7.5 and the stochastic component can, as before, be eliminated by choosing the number of underlyings in the portfolio as $\Delta = -\partial V / \partial S$. The portfolio thus constructed is risk-free. From arbitrage considerations, we conclude that a risk-free portfolio can yield *no more* than the risk-free interest rate. But it is certainly possible that it may earn less than the risk-free rate if the holder of the option chooses to exercise the option at a less than optimal time, in other words if the option is exercised when it should have been held or held when it should have been exercised (the holder of a European option cannot make this mistake as European options do not entail the right to early exercise). Consequently, a risk-free portfolio containing American options can at best earn a yield equal to the risk-free rate if the option is optimally exercised, otherwise less is earned. The relation for American options corresponding to the Black-Scholes equation.7.7 is thus

$$\underbrace{\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2(S, t)S^2\frac{\partial^2 V}{\partial S^2} - q(t)S\frac{\partial V}{\partial S}}_{\text{Yield from a } \Delta\text{-Hedged portfolio}} \leq \underbrace{r(t)\left[V - S\frac{\partial V}{\partial S}\right]}_{\text{Yield from a Bank Account}}$$

or, written in the form corresponding to Eq. 7.8:

$$\frac{\partial V}{\partial t} + [r(t) - q(t)]\, S\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2(S, t)S^2\frac{\partial^2 V}{\partial S^2} - r(t)V \leq 0 \, . \qquad (7.10)$$

Since the option can be exercised at any time, it is always worth at least the amount given by its payoff profile. In addition to the initial condition, Eq. 7.9 (which also holds for American options), i.e.

$$V(S, t = T) = P(S) \, , \qquad (7.11)$$

the following inequality must therefore always hold

$$V(S, t) \geq P(S) \quad \forall t . \tag{7.12}$$

Intuitively, the two inequalities 7.10 and 7.12 imply the two possible cases:

- Either the value of the option is, on a certain parameter set of S and t, greater than the payoff profile (i.e., Eq. 7.12 is a strict *in*equality) in which case the right to early exercise is worthless. Thus the option behaves as a European option on this parameter set and therefore Eq. 7.10 holds as an *equality.*
- On the set of S and t where the option's value is not greater than its payoff profile (i.e., Eq. 7.12 holds as an equality), it would not be reasonable to hold the option. In this case, the holder should exercise early. Therefore the "risk-free" portfolio including the option would earn less than the risk-free rate since the risk-free rate can only be obtained by exercising *optimally*. Thus, in this case Eq. 7.10 holds as a strict *in*equality.

In summary: there exist two different parameter sets (regions in the (S, t)-plane). In one region we have

$$\frac{\partial V}{\partial t} + [r(t) - q(t)] S \frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2(S, t)S^2 \frac{\partial^2 V}{\partial S^2} - r(t)V = 0 \quad \text{and} V(S, t) > P(S). \tag{7.13}$$

and in the other region we have

$$\frac{\partial V}{\partial t} + [r(t) - q(t)] S \frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2(S, t)S^2 \frac{\partial^2 V}{\partial S^2} - r(t)V < 0 \quad \text{and} V(S, t) = P(S) . \tag{7.14}$$

The boundary between these parameter sets is a curve denoted by $S^*(t)$, called the *optimal exercise boundary* on which the option is exercised optimally. For values of $S(t)$ on one side of this curve, the option should be held, on the other side, the option should have been exercised earlier. Optimal exercise occurs directly on the curve. On the set where Eq. 7.12 holds as an equality, the value of the option need not be computed since it is simply given by the payoff profile. It remains to calculate its value on the set where its behavior is European. We must therefore find a solution to Eq. 7.10 as an *equality* on a parameter set whose boundary is given by the optimal exercise boundary

$S^*(t)$ and which in addition satisfies the initial condition given by Eq. 7.11. The difficulty is that the optimal exercise boundary is not known a priori. This type of problem is known as a *free* boundary condition.[3] For values on the boundary $S^*(t)$, the value of the option (as a solution of 7.13) must be equal to the payoff profile. If it were greater, the option holder would not exercise (the boundary would thus lie elsewhere), whereas a smaller value would yield an arbitrage opportunity (buying the option and exercising immediately). Thus, the option price must be continuous on the boundary:

$$V(S = S^*, t) = P(S = S^*) \quad \forall t .$$

The boundary curve $S^*(t)$ is positioned to give the option the greatest possible value at the boundary for all t (this is precisely the meaning of *optimal* exercise). For most options (but not, for example, digitals), this means that the first derivative of the option price with respect to S is continuous on the boundary:

$$\left. \frac{\partial V(S, t)}{\partial S} \right|_{S=S^*} = \left. \frac{\partial P(S)}{\partial S} \right|_{S=S^*} .$$

These two conditions (the continuity of the option price function and its first derivative) in general enable the determination of the free boundary $S^*(t)$.

7.1.3 A First Contact with the Risk-Neutral World

The mean return (drift) $\widetilde{\mu}$ of the underlying does *not* appear in the general differential equation 7.8 for European derivatives nor in the corresponding (in)equalities 7.13 and 7.14 for American derivatives. If a parameter does not appear in the differential equation nor the boundary condition, it is completely irrelevant for the problem under consideration. The drift of the underlying therefore has absolutely no influence on the value of a derivative on the underlying concerned! It was somehow "lost" in the derivation. This occurred exactly at that point where the delta hedge, Eq. 7.6, was introduced into the equation to eliminate the stochastic component. Here, $\widetilde{\mu}$ eliminated as well. The volatility σ also appears in the stochastic component dS. Contrary to $\widetilde{\mu}$, the volatility σ remains a parameter in the hedged portfolio since, as a

---

[3]In contrast, for European options, the Black-Scholes differential equation is always solved on the set $S = 0$ to $S = \infty$. The boundary is thus always fixed and known for European options.

consequence of the Ito lemma, it has entered into the deterministic component of Eq. 7.3, i.e., into the dt-term.

Apart from this purely technical explanation, from an economical point of view it is at first glance quite extraordinary that a parameter as important as the mean return of an underlying should have absolutely no effect on the value of derivatives on that underlying. In fact, most investors expect a mean rate of return as compensation for accepting the risk of investing at all. And they expect the compensation to be higher the greater the risk. The fact that the mean return has no influence on the price of a derivative (no matter how large the risk—represented by the volatility σ—is!) can be interpreted in the following way: the valuation occurs in a world in which investors are indifferent to the risk involved in investing. This is referred to as *risk neutrality* or a *risk-neutral valuation*.

The valuation of derivatives takes place in a risk-neutral world with objective parameters. This, however, is not the real world. In reality, comparing a risky with a less risky stock and all other conditions identical, an investor would always prefer to buy the less risky stock, unless the higher risk is compensated by an appropriately higher mean return. The mean return is determined by the real world drift $\widetilde{\mu}$.

Contrary, a bank trading derivatives will generally hedge itself against all market (and to some extend credit) risk of its total derivatives position. As shown above, the hedging procedure will eliminate all dependence of the real world drift $\widetilde{\mu}$, which is replaced by the risk neutral drift, which is identical to the risk free interest rate. An investor buying a derivative as an asset for speculation is not hedged and is only interested in the performance of the asset in the real world. From the investor's perspective, the Equation derived above would not make any sense. Therefore, considering the valuation of derivatives, it is always necessary to keep in mind in which world the valuation should be performed, the real or the risk neutral world.

7.2 The Black-Scholes Equation and the Backward Equation

Having completed a derivation of the Black-Scholes equation providing the advantage of an intuitive interpretation[4] in accordance with Eq. 7.7, which can be simply extended to apply to American options, we now proceed from

---

[4]The Black-Scholes equation simply means that the yield from a delta-hedged portfolio must equal the yield from a risk-free investment.

the fundamental principles of stochastic analysis introduced in Sect. 2.4 to establish a connection between the equations derived there (in particular the backward equation) and the Black-Scholes equation and thus the pricing of derivatives. The transition probabilities $p(S', t' | S, t)$ for the stochastic processes satisfying Eq. 2.19 will prove to be very useful for this purpose.

As previously mentioned, the transition probabilities $p(S', t' | S, t)$ contain all available information about the stochastic process. All pertinent values can be established when this function is known. A value of particular interest is the *expectation* of a function $f(S', t')$ at a future date $t' = T$ based on information available today.[5] We denote the expectation by the letter E, thus

$$E(S, t, T) = \mathrm{E}\left[f(S', t' = T) \,\middle|\, \text{at time } t \text{ the process } S' \text{ is equal to } S \right]$$
$$= \mathrm{E}\left[f(S', T) \,\middle|\, S'(t) = S \right] \ .$$

This expectation can be expressed in terms of the transition probabilities of the stochastic process starting today from the point (S, t) and traveling to the future point $(S', t' = T)$ as

$$E(S, t, T) = \int_{-\infty}^{\infty} f(S', T) p(S', T | S, t) dS' \ .$$

Taking the derivative of this expression with respect to t, taking note of the fact that f depends on the time T, but not on t, we obtain an equation for the change in this expectation with respect to a change in time

$$\frac{\partial E(S, t, T)}{\partial t} = \int_{-\infty}^{\infty} f(S', T) \frac{\partial p(S', T | S, t)}{\partial t} dS' \ .$$

We now immediately see that the change in this expectation is completely governed by the time derivative of the transition probability p. Since we wish to establish values at an *earlier* time point proceeding backwards from later information, we substitute the *backward* equation 2.35 for the time derivative

---

[5]In applications, this function $f(S, T)$ is often the payoff profile of a derivative.

of p in the above integral. This allows us to express the $\partial p/\partial t$ in terms of derivatives with respect to S:

$$
\begin{aligned}
\frac{\partial E(S,t,T)}{\partial t} &= \int_{-\infty}^{\infty} f(S',T) \left[-\frac{1}{2}b^2(S,t)\frac{\partial^2 p}{\partial S^2} - a(S,t)\frac{\partial p}{\partial S} \right] dS' \\
&= -\frac{1}{2}b^2(S,t) \int_{-\infty}^{\infty} f(S',T)\frac{\partial^2 p}{\partial S^2} dS' - a(S,t) \int_{-\infty}^{\infty} f(S',T)\frac{\partial p}{\partial S} dS' \\
&= -\frac{1}{2}b^2(S,t)\frac{\partial^2}{\partial S^2} \underbrace{\int_{-\infty}^{\infty} f(S',T)p\, dS'}_{E(S,t,T)} - a(S,t)\frac{\partial}{\partial S} \underbrace{\int_{-\infty}^{\infty} f(S',T)p\, dS'}_{E(S,t,T)} .
\end{aligned}
$$

In the last step, we have made use of the fact that $f(S',T)$ depends on S', but not on S and that the derivatives with respect to S thus have no effect on f. We conclude that the expectation E itself satisfies the backward equation

$$
\frac{\partial E}{\partial t} + \frac{1}{2}b(S,t)^2\frac{\partial^2 E}{\partial S^2} + a(S,t)\frac{\partial E}{\partial S} = 0 , \tag{7.15}
$$

with a somewhat different boundary condition than p, namely

$$
E(S, t=T, T) = f(S,T) .
$$

Writing the parameters a and b as relative parameters as in Eq. 7.1, we see that the differential equation for E is quite similar to the Black-Scholes equation

$$
\frac{\partial E}{\partial t} + \frac{1}{2}\sigma(S,t)^2 S^2 \frac{\partial^2 E}{\partial S^2} + \widetilde{\mu}(S,t)S\frac{\partial E}{\partial S} = 0 .
$$

As in Eq. 7.8 the terms $\sim \partial/\partial t$, $\sim S\,\partial/\partial S$ and $\sim S^2\partial^2/\partial S^2$ appear. Only the term not involving derivatives is missing, i.e., that corresponding to "$r(t)V$" in Eq. 7.8.

We consider therefore the *discounted* future expectation (denoted by V below) of an arbitrary function $f(S',T)$

$$
\begin{aligned}
V(S,t,T) &= B(t,T)E(S,t,T) \\
&= B(t,T)\int_{-\infty}^{\infty} f(S',T)p(S',T\,|S,t)dS' . \tag{7.16}
\end{aligned}
$$

The derivative of this expression with respect to time t is

$$\frac{\partial V}{\partial t} = \frac{\partial B(t, T)}{\partial t} \underbrace{\int_{-\infty}^{\infty} f(S', T) p(S', T \mid S, t) dS'}_{B(t,T)^{-1} V}$$

$$+ B(t, T) \int_{-\infty}^{\infty} f(S', T) \frac{\partial p(S', T \mid S, t)}{\partial t} dS' \ .$$

Using an argument analogous to the one presented above, the substitution of the backward equation for $\partial p/\partial t$ yields[6]

$$\frac{\partial V}{\partial t} + \frac{1}{2} b^2(S, t) \frac{\partial^2 V}{\partial S^2} + a(S, t) \frac{\partial V}{\partial S} = V \frac{\partial}{\partial t} \ln(B(t, T)) \qquad (7.17)$$

with the same boundary condition as for E, namely

$$V(S, t = T, T) = f(S, T) \ .$$

This corresponds exactly to Eq. 7.15 with an addition "source" term appearing as a consequence of discounting. Again writing the parameters a and b as relative parameters as in Eq. 7.1, the equation takes the form

$$\frac{\partial V}{\partial t} + \frac{1}{2} \sigma^2(S, t) S^2 \frac{\partial^2 V}{\partial S^2} + \tilde{\mu}(S, t) S \frac{\partial V}{\partial S} = V \frac{\partial}{\partial t} \ln(B(t, T)) \ . \qquad (7.18)$$

For continuous discounting we have

$$\frac{\partial}{\partial t} \ln(B(t, T)) = r(t) \ .$$

7.2.1 A Second Contact with the Risk-Neutral World

Thus, the differential equation 7.18 for the (until now) abstract value of a "discounted expectation"—including the boundary condition—has the same

---

[6]We have also used the following general property of the logarithm:

$$\frac{\partial}{\partial t} \ln(B) = \frac{1}{B} \frac{\partial B}{\partial t} \ .$$

structure as the Black-Scholes differential equation for pricing derivatives, 7.8. In fact, we only need to choose the parameter $\widetilde{\mu}$ so that $\widetilde{\mu}(S, t) = r(t) - q(t)$ and the differential equation for the discounted expectation is *identical* to the Black-Scholes equation! This then has the following interpretation: the price of a derivative with a payoff profile $f(S, T)$ is the same as the discounted expectation of this payoff profile *if and only if* the drift of the underlying is given by[7]

$$\widetilde{\mu} = r - q \quad \text{or equivalently}$$

$$\mu = r - q - \frac{\sigma^2}{2} \ . \tag{7.19}$$

Here, we also see that the actual drift (mean return) of the underlying has absolutely no effect on the *valuation* of the derivative. We again encounter *risk neutrality*, as in Sect. 7.1.3. The approach via the backward equation however, provides a much clearer interpretation of risk neutrality:

- The price of a derivative can be expressed as the discounted expectation of its payoff profile.
- To do so, the drift of the underlying must be precisely specified as in Eq. 7.19. This is called the *risk-neutral drift*.

As will be seen later, choosing a drift corresponds to choosing a particular probability measure with respect to which the expectation is to be taken. The choice in Eq. 7.19 is therefore referred to as the choice of a *risk-neutral probability measure*.

7.3   The Relationship to the Heat Equation

The Black-Scholes equation is a *linear, parabolic partial differential equation of second order*. The simplest parabolic partial differential equation (PPDE) of second order is the *heat equation* or *diffusion equation*. This equation has the simple form

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial \tau} \tag{7.20}$$

---

[7]In the second equation, Eq. 2.26 was used in the conversion to the drift $\mu = \widetilde{\mu} - \sigma^2/2$ of the stochastic process for the returns $\ln(S)$.

with x being the space variable and τ the time variable. It is one of the most thoroughly investigated equations in physics and applied mathematics. The Black-Scholes equation for constant yields (interest rates and dividends) and constant volatility can, as will be shown below, be transformed into the heat equation with appropriate changes in variables. The Black-Scholes equation and the heat equation have several interesting properties in common. For example, the time derivative in both equations is of first order. The solution to the equation thus has a time orientation. This corresponds to the property that heat flow is not reversible: if, for example, a point on an otherwise cold metal rod is heated, the passage of time leads to a uniform distribution of the heat over the entire rod. The singularity, represented by the application of the heat at a single point, is quickly smoothed out over the entire piece of metal. This property applies to the Black-Scholes equation as well. The time orientation of the Black-Scholes equation, however, is such that time flows from the future into the past. Non-smooth payoff functions (the payoffs of plain vanilla puts and calls have a "corner" at the strike price) become smooth immediately as time begins to move backwards from T. The analogy to the heat equation is, however, not only very useful for obtaining an intuitive understanding of the Black-Scholes equation. The biggest advantage of comparing the two is that the heat equation is one of the best-understood partial differential equations and many results holding for it may be extended to the Black-Scholes equation. We have the results of nearly two hundred years of research at our disposal. For example, the general solution to Eq. 7.20 with initial condition

$$u(x, \tau = 0) = u_0(x)$$

is given by

$$u(x, \tau) = \frac{1}{2\sqrt{\pi \tau}} \int_{-\infty}^{\infty} e^{-(x-y)^2/4\tau} \, u_0(y) \, dy \; . \tag{7.21}$$

Thus, it is worthwhile to transform the Black-Scholes equation into the heat equation in order to make use of the results already known for the heat equation and extend them to Eq. 7.8. We will therefore now explicitly present the changes of variables required for this transformation.

First, we write V in a more convenient form in order to eliminate the term not involving derivatives, i.e., to eliminate the rV in Eq. 7.8:

$$V =: e^{-r(T-t)}u \implies \frac{\partial V}{\partial t} = rV + e^{-r(T-t)}\frac{\partial u}{\partial t}$$

$$\frac{\partial V}{\partial S} = e^{-r(T-t)}\frac{\partial u}{\partial S}, \quad \frac{\partial^2 V}{\partial S^2} = e^{-r(T-t)}\frac{\partial^2 u}{\partial S^2}.$$

This transforms the equation and boundary condition into

$$\frac{\partial u}{\partial t} + (r-q)S\frac{\partial u}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 u}{\partial S^2} = 0, \quad u(S, t=T) = P(S).$$

We now replace the time t by a dimensionless variable which measures the derivative's time to maturity. This also effects the reversal of the time orientation in the differential equation[8]:

$$z := \frac{\sigma^2}{2}(T-t) \implies \frac{\partial}{\partial t} = \frac{\partial z}{\partial t}\frac{\partial}{\partial z} = -\frac{\sigma^2}{2}\frac{\partial}{\partial z}.$$

At this stage we have made use of the fact that the volatility σ does not depend explicitly on the time t. This transforms the differential equation and the initial condition into

$$-\frac{\sigma^2}{2}\frac{\partial u}{\partial z} + (r-q)S\frac{\partial u}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 u}{\partial S^2} = 0, \quad u(S, z=0) = P(S).$$

In order to eliminate the S dependence of the coefficients, we replace S with the logarithm of S. The argument of the logarithm must, however, be dimensionless. We thus divide by a constant K which has the same dimension as S. A natural choice for this constant would be the strike price of the option which appears in the payoff profile.

$$y := \ln(\frac{S}{K}) \implies S = Ke^y, \quad S\frac{\partial}{\partial S} = S\frac{\partial y}{\partial S}\frac{\partial}{\partial y} = S\frac{K}{S}\frac{1}{K}\frac{\partial}{\partial y} = \frac{\partial}{\partial y}$$

$$\frac{\partial^2}{\partial y^2} = S\frac{\partial}{\partial S}(S\frac{\partial}{\partial S}) = S\underbrace{\frac{\partial S}{\partial S}}_{1}\frac{\partial}{\partial S} + S^2\frac{\partial^2}{\partial S^2} \implies S^2\frac{\partial^2}{\partial S^2} = \frac{\partial^2}{\partial y^2} - \frac{\partial}{\partial y}.$$

---

[8]Note that by definition the volatility has the dimension $1/\sqrt{\text{time}}$, i.e. σ^2 has the dimension $1/\text{time}$, thus making the variable z dimensionless.

The resulting differential equation and boundary condition is then given by

$$-\frac{\partial u}{\partial z} + \left[\frac{2(r-q)}{\sigma^2} - 1\right]\frac{\partial u}{\partial y} + \frac{\partial^2 u}{\partial y^2} = 0\,, \quad u(y, z = 0) = P(Ke^y)\,.$$

The term involving the first derivative with respect to y can also be eliminated. For this, a new variable x is needed which depends on both y and z. The change in variable will involve both "old" variables. In consequence, *two* "new" variables x and τ need to be introduced. The transformation of the derivatives is accomplished using the *Jacobian determinant* which can be found in any introductory textbook on analysis:

$$x := y + (\frac{2(r-q)}{\sigma^2} - 1)z\,, \quad \tau := z \Longrightarrow$$

$$\frac{\partial}{\partial y} = \frac{\partial x}{\partial y}\frac{\partial}{\partial x} + \frac{\partial \tau}{\partial y}\frac{\partial}{\partial \tau} = 1\frac{\partial}{\partial x} + 0$$

$$\frac{\partial^2}{\partial y^2} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial y}\right) = \frac{\partial}{\partial y}\frac{\partial}{\partial x} = \frac{\partial^2}{\partial x^2}$$

$$\frac{\partial}{\partial z} = \frac{\partial x}{\partial z}\frac{\partial}{\partial x} + \frac{\partial \tau}{\partial z}\frac{\partial}{\partial \tau} = (\frac{2(r-q)}{\sigma^2} - 1)\frac{\partial}{\partial x} + 1\frac{\partial}{\partial \tau}\,.$$

Here we have made use of the fact that r, q and σ do not depend on y (and thus on S) nor on z (thus t). This completes the transformation of the differential equation and initial condition to

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial \tau}\,, \quad u_0(x) \equiv u(x, \tau = 0) = P(Ke^x)\,. \tag{7.22}$$

This is simply the heat equation. The strategy for calculating option prices is therefore to find a solution u of the heat equation in terms of the variables x and τ and then to substitute the transformed variables by the original variables as follows:

$$\tau = z = \frac{\sigma^2}{2}(T - t)$$

$$x = y + (\frac{2(r-q)}{\sigma^2} - 1)z = \ln(\frac{S}{K}) + (r - q - \frac{\sigma^2}{2})(T - t)$$

$$u = e^{r(T-t)}V \tag{7.23}$$

$$u_0(x) = P(Ke^x, K)\,,$$

thereby expressing the solution in terms of the original financial variables.

8

Integral Forms and Analytic Solutions in the Black-Scholes World

In addition to Assumptions 1, 2, 3, 4, 5 and 6 from Chap. 4 required to set up the differential equation in Chap. 7, we will now further simplify our model by assuming that the parameters involved (interest rates, dividend yields, volatility) are constant (Assumptions 9, 11 and thus 7 from Chap. 4) despite the fact that these assumptions are quite unrealistic. These were the assumptions for which Fischer Black and Myron Scholes derived their famous analytic expression for the price of a plain vanilla option, the Black-Scholes option pricing formula.[1]

For this reason, we often speak of the *Black-Scholes world* when working with these assumptions. In the Black-Scholes world, solutions of the Black-Scholes differential equation (i.e., option prices) for some payoff profiles (for example for plain vanilla calls and puts) can be given in closed form. We will now present two elegant methods to derive such closed form solutions.

---

[1]The mathematician Louis Bachelier was actually the first to derive analytical expressions for the valuation of options in 1900 [7]. However, Bachelier's derivation is based on other assumptions and his work has been forgotten for a long time. Only through the work of Black and Scholes and nonetheless because of the availability of computers, the use of mathematical formulas and methods has become a market standard for the valuation of derivatives.

8.1   Option Prices as Solutions of the Heat Equation

The first and perhaps most natural approach would be to take advantage of the constant parameter assumptions to transform the Black-Scholes equation into the heat equation 7.22 as presented above. Since the solution to the heat equation is known and given by Eq. 7.21 we simply need to write the initial condition corresponding to the desired financial instrument in terms of the variables x and τ, and, in accordance with Eq. 7.23, transform the solution u back in terms of the original financial variables. We now demonstrate this technique using a plain vanilla call as an example.

Expressing the payoff profile of the call in the variables of the heat equation gives

$$P(S) = \max(S - K, 0)$$
$$\implies \quad u_0(x) = P(Ke^x) = \max(Ke^x - K, 0) = K \max(e^x - 1, 0) \ .$$

Substituting this into Eq. 7.21 immediately yields the solution in integral form for this initial condition

$$u(x, \tau) = \frac{1}{2\sqrt{\pi \tau}} \int_{-\infty}^{\infty} e^{-(x-y)^2/4\tau} \, u_0(y) \, dy \ . \tag{8.1}$$

This *integral form* is not only valid for plain vanilla calls but for arbitrary European payoff profiles and the resulting initial conditions u_0. The integral can be computed numerically using, for example, the Monte Carlo method. Decades of research on numerical methods for computing integrals can be taken advantage of here.

In the case of the plain vanilla call, however, it is in fact possible to obtain a closed *analytical* form of the solution of the above integral. Substituting the initial condition for the call into the above equation yields

$$
\begin{aligned}
u(x, \tau) &= \frac{K}{2\sqrt{\pi \tau}} \int_{-\infty}^{\infty} e^{-(x-y)^2/4\tau} \max(e^y - 1, 0) \, dy \\
&= \frac{K}{2\sqrt{\pi \tau}} \int_{0}^{\infty} e^{-(x-y)^2/4\tau} (e^y - 1) \, dy \\
&= \frac{K}{2\sqrt{\pi \tau}} \int_{0}^{\infty} e^{y-(x-y)^2/4\tau} \, dy - \frac{K}{2\sqrt{\pi \tau}} \int_{0}^{\infty} e^{-(x-y)^2/4\tau} \, dy \ .
\end{aligned}
$$

The second of the two integrals above can be computed after making the change in variable $\widetilde{y} \equiv (x - y)/\sqrt{2\tau}$ implying $dy = -\sqrt{2\tau}\,d\widetilde{y}$. The integral bounds must then be transformed as follows: $0 \to x/\sqrt{2\tau}$ and $\infty \to -\infty$. The integral has now become an integral over the density function of the standard normal distribution:

$$K\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x/\sqrt{2\tau}} e^{-\widetilde{y}^2/2}\,d\widetilde{y} = K\,\text{N}\left(\frac{x}{\sqrt{2\tau}}\right),$$

where $\text{N}(x)$ denotes the cumulative normal distribution given by Eq. A.49. The first of the above two integrals can be calculated by completing the squares as follows

$$y - \frac{(x - y)^2}{4\tau} = \tau + x - \frac{(y - x - 2\tau)^2}{4\tau}$$

This transforms the integral into

$$\frac{K}{\sqrt{2\pi}\sqrt{2\tau}} e^{\tau+x} \int_0^\infty e^{-\frac{(y-x-2\tau)^2}{2\sqrt{2\tau}^2}}\,dy$$

The remaining integral can now again be expressed in terms of the standard normal distribution. The necessary change in variable is $\widetilde{y} \equiv -(y - x - 2\tau)/\sqrt{2\tau}$. Combining all these results the solution becomes

$$u(x, \tau) = K\,e^{x+\tau}\,\text{N}\left(\frac{x}{\sqrt{2\tau}} + \sqrt{2\tau}\right) - K\,\text{N}\left(\frac{x}{\sqrt{2\tau}}\right).$$

Substituting now for the original variables using Eq. 7.23 gives

$$e^{r(T-t)}V(S, t) = K\,e^{\ln(\frac{S}{K})+(r-q)(T-t)}$$

$$\times \text{N}\left(\frac{\ln(\frac{S}{K}) + (r - q - \frac{\sigma^2}{2})(T - t)}{\sqrt{\sigma^2(T - t)}} + \sqrt{\sigma^2(T - t)}\right)$$

$$- K\,\text{N}\left(\frac{\ln(\frac{S}{K}) + (r - q - \frac{\sigma^2}{2})(T - t)}{\sqrt{\sigma^2(T - t)}}\right).$$

After multiplying by $e^{-r(T-t)}$, we finally obtain the famous Black-Scholes formula for the price of a European call:

$$V(S,t) = e^{-q(T-t)} S \, \mathrm{N}\,(x) - e^{-r(T-t)} K \, \mathrm{N}(x - \sigma \sqrt{(T-t)}) \qquad (8.2)$$

where

$$x \equiv \frac{\ln(\frac{S}{K}) + (r-q)(T-t)}{\sqrt{\sigma^2(T-t)}} + \frac{1}{2}\sigma\sqrt{(T-t)}\;.$$

This formula is still generally valid, if interest rate and volatility are time-dependent, but deterministic. In this case, the constant interest rate and the constant volatility just need to be replaced by their average values $\tilde{r}$ and $\tilde{\sigma}$ with

$$\tilde{r} = \frac{1}{T-t}\int_t^T r(s)\,ds \qquad \text{and}$$

$$\tilde{\sigma} = \sqrt{\frac{1}{T-t}\int_t^T \sigma^2(s)\,ds}\;.$$

8.2   Option Prices and Transition Probabilities

We will now show how the foundations of stochastic analysis laid in Sect. 2.4 can be used to price options. In Sect. 7.2, we have seen that with the risk-neutral choice of drift, the prices of derivatives are given by the discounted expectation of the payoff profile, Eq. 7.16. This expectation is determined using the transition probabilities $p(S', t'|S, t)$. *If* these are known, calculating the price of the option reduces to simply calculating the integral. In the Black-Scholes world, i.e., for the simple process 2.23, the transition probabilities are given explicitly by Eq. 2.38 with $\widetilde{\mu} = \mu + \sigma^2/2$. Thus, Eq. 7.16 becomes the integral form for the price of an arbitrary derivative with an associated payoff profile $f(S, T)$:

$$V(S,t,T) = B(t,T)\int_{-\infty}^{\infty} f(S', T) p(S', T|S, t)dS'$$

$$= \frac{B(t,T)}{\sqrt{2\pi\sigma^2(T-t)}}\int_{-\infty}^{\infty} f(S', T)\exp\left\{\frac{-\left[\ln(S'/S) - \mu(T-t)\right]^2}{2\sigma^2(T-t)}\right\}\frac{dS'}{S'}\;.$$

$$(8.3)$$

This integral can be computed numerically for arbitrary payoff profiles $f(S, T)$ and is equal to the price of the derivative for the risk-neutral choice of the drift as specified in Eq. 7.19.

For some special payoff profiles, the integral can even be solved analytically, or reduced to an expression in terms of known functions. We demonstrate this using the concrete example of a plain vanilla call option with payoff profile $f(S', T) = \max(S' - K, 0)$. For this payoff profile, the integral can be written as

$$V(S, t, T) = \frac{B(t, T)}{\sqrt{2\pi\sigma^2(T-t)}} \int_K^\infty (S' - K) \exp\left\{\frac{-\left[\ln(S'/S) - \mu(T-t)\right]^2}{2\sigma^2(T-t)}\right\} \frac{dS'}{S'} \,.$$

The substitution $u := \ln(S'/S)$ simplifies the integral to

$$V(S, t, T) = \frac{B(t, T)}{\sqrt{2\pi\sigma^2(T-t)}} \int_{\ln(K/S)}^\infty (Se^u - K) \exp\left\{\frac{-[u - \mu(T-t)]^2}{2\sigma^2(T-t)}\right\} du$$

$$\tag{8.4}$$

$$=: B(t, T)SI_1 - B(t, T)KI_2 \,.$$

Both integrals I_1 and I_2 can be easily calculated. In the first integral we complete the square in the argument of the exp-function:

$$u - \frac{[u - \mu(T-t)]^2}{2\sigma^2(T-t)} = -\frac{u^2 - 2\mu(T-t)u + \mu^2(T-t)^2 - 2\sigma^2(T-t)u}{2\sigma^2(T-t)}$$

$$= -\frac{\left[u - \left(\mu + \sigma^2\right)(T-t)\right]^2 - \left(\mu + \sigma^2\right)^2(T-t)^2 + \mu^2(T-t)^2}{2\sigma^2(T-t)}$$

$$= -\frac{\left[u - \left(\mu + \sigma^2\right)(T-t)\right]^2 - \left(\sigma^4 + 2\mu\sigma^2\right)(T-t)^2}{2\sigma^2(T-t)}$$

$$= \frac{-\left[u - \left(\mu + \sigma^2\right)(T-t)\right]^2}{2\sigma^2(T-t)} + \left(\mu + \frac{\sigma^2}{2}\right)(T-t) \,.$$

Thus, the first integral becomes

$$
\begin{aligned}
I_1 &\equiv \frac{1}{\sqrt{2\pi\sigma^2(T-t)}} \int_{\ln(K/S)}^{\infty} \exp\left\{ u - \frac{[u - \mu(T-t)]^2}{2\sigma^2(T-t)} \right\} du \\
&= \frac{e^{(\mu+\sigma^2/2)(T-t)}}{\sqrt{2\pi\sigma^2(T-t)}} \int_{\ln(K/S)}^{\infty} \exp\left\{ \frac{-\left[u - \left(\mu+\sigma^2\right)(T-t)\right]^2}{2\sigma^2(T-t)} \right\} du \ .
\end{aligned}
$$

With the substitution

$$
y := -\frac{u - \left(\mu+\sigma^2\right)(T-t)}{\sqrt{\sigma^2(T-t)}} \Rightarrow
$$

$$
\frac{dy}{du} = -1/\sqrt{\sigma^2(T-t)} \Rightarrow du = -\sqrt{\sigma^2(T-t)}\,dy
$$

the upper and lower limits of integration become

$$
y_{\text{upper}} = -\frac{\infty - \left(\mu+\sigma^2\right)(T-t)}{\sqrt{\sigma^2(T-t)}} = -\infty
$$

$$
y_{\text{lower}} = -\frac{\ln(K/S) - \left(\mu+\sigma^2\right)(T-t)}{\sqrt{\sigma^2(T-t)}} = \frac{\ln(S/K) + \left(\mu+\sigma^2\right)(T-t)}{\sqrt{\sigma^2(T-t)}} \ .
$$

Exchanging the upper and lower limits results in a change in the sign of the integral. This is compensated for by the sign of du. Combining the above results, I_1 becomes after this substitution

$$
\begin{aligned}
I_1 &= e^{(\mu+\sigma^2/2)(T-t)} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\ln(S/K)+(\mu+\sigma^2)(T-t)}{\sqrt{\sigma^2(T-t)}}} \exp\left\{ \frac{-y^2}{2} \right\} dy \\
&= e^{(\mu+\sigma^2/2)(T-t)} \mathrm{N}\left(\frac{\ln(S/K) + \left(\mu+\sigma^2\right)(T-t)}{\sqrt{\sigma^2(T-t)}} \right) ,
\end{aligned}
$$

where, as usual, N denotes the cumulative standard normal distribution.

The second integral can be computed after making the substitution $y := -(u - \mu(T - t))/\sqrt{\sigma^2(T - t)}$:

$$I_2 \equiv \frac{1}{\sqrt{2\pi\sigma^2(T-t)}} \int_{\ln(K/S)}^{\infty} \exp\left\{\frac{1}{2}\frac{-[u-\mu(T-t)]^2}{\sigma^2(T-t)}\right\} du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\ln(S/K)+\mu(T-t)}{\sqrt{\sigma^2(T-t)}}} \exp\left\{-\frac{y^2}{2}\right\} dy$$

$$= N\left(\frac{\ln(S/K) + \mu(T-t)}{\sqrt{\sigma^2(T-t)}}\right)$$

The generalization of Eq. 7.19 for the risk-neutral choice of drift in arbitrary compounding methods is (see Eq. 9.25)

$$\mu(T - t) := \ln\left(\frac{B_q(t, T)}{B(t, T)}\right) - \frac{\sigma^2}{2}(T - t)\}$$

with $B_q(t, T) = \exp(-q(T - t))$. This simplifies the integrals further to

$$I_1 = \frac{B_q(t, T)}{B(t, T)} N(x) \quad I_2 = N\left(x - \sigma\sqrt{T - t}\right).$$

where x is, as usual, given by

$$x = \frac{\ln(\frac{B_q(t,T)S}{B(t,T)K}) + \frac{1}{2}\sigma^2(T-t)}{\sqrt{\sigma^2(T-t)}}. \tag{8.5}$$

Collecting these results, we obtain the price of a plain vanilla call as

$$V(S, t, T) = B(t, T)SI_1 - B(t, T)KI_2$$

$$= B_q(t, T)S(t) N(x) - B(t, T)K N\left(x - \sigma\sqrt{T - t}\right). \tag{8.6}$$

Again, this is the famous *Black-Scholes option pricing formula* and corresponds exactly to Eq. 8.2 for continuous compounding.

8.3   Compilation of Black-Scholes Option Prices for Different Underlyings

8.3.1   Options on the Spot Price

The above derivation holds for a call on an underlying with a continuous compounding dividend *yield*. However, in reality, dividends are paid as discrete amount at only a few days per year, often once a year only. Such a discrete dividend payment could either be modeled as a fixed absolute amount (absolute discrete dividend) or as an amount relative to the spot price at the ex-dividend date (relative discrete dividend). A simple approach for taking discrete dividends into account is to adjust the spot price of the underlying by subtracting the value of the dividend for the considered time period according to Eq. 2.9. The adjusted spot price can be modeled like an underlying that does not pay any dividends. The payoff profile and the Black-Scholes value are summarized here for puts and calls. The payoff profiles at time T are:

$$c_S(T, T, K) = \max\{0, S(T) - K\}$$
$$p_S(T, T, K) = \max\{0, K - S(T)\} .$$

The Black-Scholes option prices at time t are:

$$c_S(t, T, K) = \tilde{S}(t, T)\mathrm{N}(x) - K\,B(t, T)\mathrm{N}(x - \sigma\sqrt{T-t})$$
$$p_S(t, T, K) = -\tilde{S}(t, T)\mathrm{N}(-x) + K\,B(t, T)\mathrm{N}(-x + \sigma\sqrt{T-t})$$

$$\tag{8.7}$$

where

$$x = \frac{\ln\left(\frac{\tilde{S}(t,T)}{K\,B(t,T)}\right)}{\sigma\sqrt{T-t}} + \frac{1}{2}\sigma\sqrt{T-t} = \frac{\ln\left(\frac{S(t,T)}{K}\right)}{\sigma\sqrt{T-t}} + \frac{1}{2}\sigma\sqrt{T-t} .$$

The application of this famous *Black-Scholes option pricing formula* to option portfolios is demonstrated in detail in the Excel workbook BLACKSCHOLESMODEL.xls from the download section [50]. In this workbook, the derivatives of the option price with respect to its parameters, called the *Greeks*, are also computed in anticipation of Chap. 12. The workbook can be used as a complete option calculator (the fields colored yellow are the input fields).

8.3.2 Options on the Forward Price

The underlying is now not the spot price $S(t)$ but the forward price $S(t, T')$ for a time $T' \geq T$, where T is the maturity date of the option. Options on the forward price refer to either futures or forwards. If physical settlement was agreed, exercising the option yields in both cases to the payment of the value of the underlying forward contract. In addition, the exerciser of the option goes long (in the case of a call) or short (in the case of a put) in the forward contract concerned.[2] Since the value of a futures and forward position according to Eqs. 6.6 and 6.5, respectively, are different, the payoff profile and thus the value of options on these contracts are different as well.

Options on Futures

Upon maturity at time T of the option, the value of the future with a maturity date $T' \geq T$ is paid if this value is positive. The payoff profiles are

$$c_F(T, T, K) = \max \left\{ 0, F_S(T, T', K) \right\} = \max \left\{ 0, S(T, T') - K \right\}$$
$$p_F(T, T, K) = \max \left\{ 0, -F_S(T, T', K) \right\} = \max \left\{ 0, K - S(T, T') \right\} .$$

A method often used to find the Black-Scholes price is to transform the payoff profile into a payoff profile of a known option. For this reason, we write the payoff profile of the call as

$$c_F(T, T, K) = \frac{B_q(T, T')}{B(T, T')} \max \left\{ 0, S(T) - \frac{B(T, T')}{B_q(T, T')} K \right\} .$$

where we have used Eq. 6.1 for the forward price at time T (for the case of a dividend yield q). Thus, the price of a call on a future with strike price K can be written as the price of B_q/B calls on the spot price with strike price $K B/B_q$. The argument for the put is completely analogous. A call on a *future* with strike price K thus has the same payoff profile as B_q/B calls on the *spot* with strike price $K B/B_q$. Thanks to of Eq. 8.7, the price of an option on

---

[2]But of course with the then valid forward price as the delivery price so that the forward contract—as always—has zero value when entered into.

the spot price is known. Substituting KB/B_q for the strike price and using Eq. 2.7 for B and B_q yields:

$$
c_S(t, T, K) = \frac{B_q(T, T')}{B(T, T')} \left[B_q(t, T) S(t) N(x') \right.
$$

$$
\left. - B(t, T) K \frac{B(T, T')}{B_q(T, T')} N(x' - \sigma \sqrt{T - t}) \right]
$$

$$
= \frac{B_q(t, T')}{B(T, T')} S(t) N(x') - B(t, T) K N(x' - \sigma \sqrt{T - t}) .
$$

Here, x' corresponds to the x in Eq. 8.7 with the modified strike price:

$$
x' = \frac{\ln \left(\frac{B_q(t,T) S(t)}{B(t,T) [K \, B(T,T')/B_q(T,T')]} \right)}{\sigma \sqrt{T - t}} + \frac{1}{2} \sigma \sqrt{T - t}
$$

$$
= \frac{\ln \left(\frac{B_q(t,T') S(t)}{B(t,T') K} \right)}{\sigma \sqrt{T - t}} + \frac{1}{2} \sigma \sqrt{T - t} .
$$

Now, again with the help of Eq. 6.1, the spot price $S(t)$ is written in terms of the actual underlying, namely the forward price. Using the expression for B in Eq. 2.7 finally gives the Black-Scholes price for options on futures:

$$
c_F(t, T, K) = B(t, T) \left[S(t, T') N(x') - K \, N(x' - \sigma \sqrt{T - t}) \right]
$$

$$
p_F(t, T, K) = B(t, T) \left[-S(t, T') N(-x') + K \, N(-x' + \sigma \sqrt{T - t}) \right]
$$

$$
\tag{8.8}
$$

where

$$
x' = \frac{\ln \left(\frac{S(t,T')}{K} \right)}{\sigma \sqrt{T - t}} + \frac{1}{2} \sigma \sqrt{T - t} .
$$

Comparing this with the corresponding prices for options on the spot prices, Eq. 8.7, for the special case of an underlying whose dividend yield is exactly

equal to the risk-free interest rate, in other words where (from Eq. 2.9) $\tilde{S}(t, T) = B(t, T)S(t)$, we obtain the following "cookbook" recipe:

> Plain vanilla option on futures can be priced like options on the spot price of an (artificial) underlying whose spot price is equal to $S(t, T')$ and whose dividend yield is equal to the risk free rate r.

If the future matures at the same date as the option, i.e., if $T' = T$, then there is (because of Eq. 6.1) no difference in either the payoff profile or the price of the option on the future and the option on the spot price. In this case Eq. 8.8 (i.e., pricing options using the forward price, even if it is an option on the spot price) is referred to as the *Black-76 model*.

In summary, if either $T = T'$ or $q = r$, there is no difference in the option on a futures contract and the option on a spot price.

Options on Forwards

On the maturity date T of the option, the value of the forward maturing on $T' \geq T$ will be paid to the holder of a call if this value is positive. This is different from the value of the future since in the case of a forward, the value is discounted from maturity T' back to T (see Eq. 6.5). The payoff profiles are thus:

$$c_f(T, T, K) = \max\{0, f_S(T, T', K)\} = B(T, T')\max\{0, S(T, T') - K\}$$
$$p_f(T, T, K) = \max\{0, -f_S(T, T', K)\} = B(T, T')\max\{0, K - S(T, T')\} \ .$$

Comparing this with the payoff profiles for options on futures shows that an option on a forward corresponds to B options on the future. Therefore the Black-Scholes prices can be immediately obtained from Eq. 8.8

$$c_f(t, T, K) = B(t, T')\left[S(t, T')N(x') - K\,N(x' - \sigma\sqrt{T - t})\right]$$
$$p_f(t, T, K) = B(t, T')\left[-S(t, T')N(-x') + K\,N(-x' + \sigma\sqrt{T - t})\right]$$

$$(8.9)$$

with $x\prime$ defined as in Eq. 8.8. The difference between this and Eq. 8.8 is that for options on *forwards* the discounting is done from the maturity of the *forward* contract T', whereas for options on *futures* the discounting is done from the maturity of the *option T*.

If the forward matures on the same date as the option, $T' = T$, there is no difference in the payoff profile or in the Black-Scholes price between an option on the forward and an option on the spot price. In this case, the prices of an option on the spot, an option on a future and an option on a forward are all equal.

8.3.3 Options on Interest Rates

Forward Volatilities

In the derivation of the Black-Scholes equation for options on the forward price, it has been assumed that volatility remained constant throughout. Therefore, in Eq. 8.8 and Eq. 8.9 the volatility of the spot price is used, though the underlying is the forward price. That is because the volatility of the forward price and the spot price are the same if the volatility is constant. This model is commonly referred to as the *Black-76 model*.

The Black-76 model is commonly used especially when the underlying S is an interest rate or an interest rate instrument (like a bond, for instance). It can be shown that the Black-76 model holds even when the Black-Scholes assumptions are weakened somewhat. The underlying process must not necessarily be a random walk with constant volatility. It is sufficient that the logarithm of the underlying $S(T)$ at option maturity is normally distributed. The variance of the distribution of $\ln(S(T))$ will be written as

$$\mathrm{var}\,[\ln S(T)] = \sigma(T)^2 T$$

The parameter $\sigma(T)$ is called the *forward volatility*. It is the volatility of the underlying price $S(T)$ at maturity T. Because the Black-76 model "lives" in the Black-Scholes world, interest rates are assumed to be non-stochastic. Therefore the forward price $S(t, T)$ and the future price $S(T)$ are equal if the underlying S is an interest rate or an interest rate instrument.[3] At an earlier time $t < T$ we can thus use the forward price $S(t, T)$ for $S(T)$ and the current volatility *of this forward price* for the forward volatility.

---

[3]This will be shown explicitly in Sect. 14.3.

Normally Versus Log Normally Distributed Interest Rates

The Black-76 model is applied to options on interest rate instruments such as bonds as well as to options which depend directly on interest rates, such as caps and floors. For $S(t, T)$ we take either the forward price of the underlying instrument (for example, of the bond) or the forward rate of the reference interest rate (3-month LIBOR rate in 6 months, for example). In both cases the application of the Black-76 model implicitly assumes that each underlying (bond price or interest rate) is lognormally distributed. However, both assumptions contradict each other. Yields are the logarithms of relative price changes. If, in the case of a *bond* option, it is assumed that these logarithms are normally distributed, the interest rates (=yields) cannot simultaneously be *log*normally distributed. This is, however, common market practice: bond options are priced under the assumption that bond prices are lognormally distributed, i.e., that interest rates are normally distributed. On the other hand, caps, floors and collars are priced under the assumption that interest rates are lognormally distributed (similarly, swaptions are priced under the assumption of lognormal swap rates).

Though, this inconsistency does not imply that the option prices are also inconsistent, since the used volatilities are different (bond price volatility vs. interest rate volatility). As will be shown in Sect. 30.3.3, price volatility and yield volatility are related to one another through the modified duration, see Eq. 30.18 (in linear approximation).

Now the question arises, which of both assumptions is indeed correct? Actually, both assumptions are neither completely wrong nor fully true. Empirical analysis of interest rate time series shows that the distribution of interest rates depends on the interest rate level (e.g., see [161]). For low or very high interest rate levels (i.e. lower than 1.2% or greater than 5.6%),[4] the distribution of interest rate is lognormal, while within these boundaries the distribution is normal. Common interest rate models, modeling the time evolution of the full interest rate term structure, with consistent modeling of e.g. Caps as well as bond options, usually assume either normally or lognormally distributed interest rates. For a more detailed discussion of the pros and cons of various models see Chap. 14.

---

[4]However, own studies conclude that it is not possible with statistical significance to discriminate between the lognormally and normally distributed interest rate assumption; at least with a advent of negative interest rates the (unmodified) lognormally distribution assumptions has to be ruled out for low interest rate levels.

Here, it should be enough to note that simple interest rate options are prices by means of Black-75 and that the according underlying is assumed to be lognormally distributed. In Case of Caps, Floors and Collars, the underlying is a forward interest rate, in the case of bond options it is the bond price and in the case of swaptions the (forward) swap rate (the assumption of lognormally distributed swap rates, which is a weighted sum of forward rates, is inconsistent with the assumption of lognormally distributed forward rates as well as with the assumption of normally distributed forward rates).

The Black-76 formula could best be understood as a vehicle to price option in an intuitively, simple form (by expressing the price in terms of the Black-76 volatility). Then, the inconsistent assumptions do not cause any trouble, if the markets are kept thoroughly separated and not mixed up.

With negative interest rates in some markets, the Black-76 is increasingly replaced by normal models or shifted log normal models (i.e. assuming that rates plus an offset are lognormally distributed).

Black-76 Model for Interest Rate Options

With the above interpretations of the input parameters and under the assumptions described above, Eq. 8.8 yields the Black-76 model for interest rate options. Explicitly:

$$
\begin{aligned}
c(t, T, K) &= B(t, T) \left[\; S(t, T)\mathrm{N}(\; x) - K\,\mathrm{N}(\; x - \sigma\sqrt{T - t}) \right] \\
p(t, T, K) &= B(t, T) \left[-S(t, T)\mathrm{N}(-x) + K\,\mathrm{N}(-x + \sigma\sqrt{T - t}) \right] ,
\end{aligned}
\tag{8.10}
$$

where

$$
x = \frac{\ln\left(\frac{S(t,T)}{K}\right)}{\sigma\sqrt{T - t}} + \frac{1}{2}\sigma\sqrt{T - t} \; .
$$

Here, $S(t, T)$ is either the forward rate of the underlying interest rate or forward price of an underlying interest rate *instrument* (like a bond, for example). This equation forms the basis for pricing interest rate option *in the Black-Scholes world*. Formally, the difference between this and Eq. 8.8 is that

the forward price with respect to the option's maturity is used, i.e., $T = T'$. As mentioned after Eq. 8.8, there is no difference in this case between an option on a forward price and an option on a spot price. We could just as well work with Eq. 8.7. The only subtlety involved is that the *forward volatility* or the volatility of the forward rate should be used.

9

Binomial and Trinomial Trees

Binomial and *trinomial trees* are very intuitive and comparatively easy to implement tools to calculate prices and sensitivity parameters of derivatives while avoiding direct reference to the fundamental differential equations governing the price of the instrument. In practice, tree methods are applied occasionally only nowadays, since other methods, e.g. the finite difference methods (see Chap. 10), show significantly superior numerical features with respect to stability, accuracy, and flexibility. For pedagogical reasons however, it is useful to learn these tree methods, because of the illustrative and direct approach to the valuation of financial derivatives.

In addition to the usual assumptions when excluding arbitrage opportunities (Assumptions 1, 2, 3, and 5), non-stochastic interest rates and default risk (Assumptions 8 and 4) of Chap. 4 will be also assumed in the subsequent sections. These assumptions allow a general theory of binomial trees to be presented. In order to actually calculate option prices, the underlying must be assumed to behave according to a model. Thus, from Sect. 9.3 onwards, it will be assumed that the underlying can be modeled as a random walk with non-stochastic volatility, i.e., the additional Assumptions 7 and 10 from Chap. 4 will be made. Furthermore, we will assume that the underlying earns a dividend *yield* in accordance with Eq. 2.9 rather than discrete dividend payments.

9.1 General Trees

9.1.1 Evolution of the Underlying and the Replicating Portfolio

Generally, in a tree procedure, the time span in question (the lifetime of the derivative) between t and T is divided into n time intervals of equal length dt:

$$T - t = n\, dt \ . \tag{9.1}$$

In each such time interval the underlying price $S(t)$ may increase in value to $u\,S$ (with $u > 1$) with a probability p', or it may decrease in value to $d\,S$ (with $d < 1$) with a probability $(1 - p')$

$$S(t) \to \begin{cases} S_u(t + dt) = u(t)S(t) \text{ with probability } p' \\ S_d(t + dt) = d(t)S(t) \text{ with probability } 1 - p' \ . \end{cases} \tag{9.2}$$

After three steps, for example, the price can take on $2^3 = 8$ possible values:

$$S(t) \to \begin{cases} u(t)S(t) \begin{cases} u(t+dt)u(t)S(t) \begin{cases} u(t+2dt)u(t+dt)u(t)S(t) \\ d(t+2dt)u(t+dt)u(t)S(t) \end{cases} \\ d(t+dt)u(t)S(t) \begin{cases} u(t+2dt)d(t+dt)u(t)S(t) \\ d(t+2dt)d(t+dt)u(t)S(t) \end{cases} \end{cases} \\ d(t)S(t) \begin{cases} u(t+dt)d(t)S(t) \begin{cases} u(t+2dt)u(t+dt)d(t)S(t) \\ d(t+2dt)u(t+dt)d(t)S(t) \end{cases} \\ d(t+dt)d(t)S(t) \begin{cases} u(t+2dt)d(t+dt)d(t)S(t) \\ d(t+2dt)d(t+dt)d(t)S(t) \ . \end{cases} \end{cases} \end{cases}$$

Consider now a portfolio consisting of Δ underlyings and g monetary units in cash. If the dividend yield q earned in the time interval dt is paid, the cash is compounded at a risk-free rate r and the price of the underlying behaves as described above, the value of the portfolio after dt is given by

$$\Delta(t)S(t) + g(t) \to \begin{cases} \Delta(t)u(t)S(t)B_q^{-1}(t) + g(t)B^{-1}(t) \\ \Delta(t)d(t)S(t)B_q^{-1}(t) + g(t)B^{-1}(t) \ . \end{cases} \tag{9.3}$$

Here, for the sake of simplifying the notation for the discount factors over a small time interval dt we have defined

$$B(t) := B(t, t + dt), \quad B_q(t) := B_q(t, t + dt). \tag{9.4}$$

In the following sections, Δ and g will be chosen so that the price of this portfolio behaves exactly as does the value of the derivative we wish to price. This value of the portfolio will then replicate the value of the derivative at each time point and is thus referred to as the *replicating portfolio.*

9.1.2 Evolution of the Derivative

If the underlying moves in accordance with Eq. 9.2 in the time interval dt, the price V of a derivative on this underlying evolves in accordance with

$$V(S, t) \rightarrow \left\langle \begin{array}{l} V(S_u, t + dt) \\ V(S_d, t + dt), \end{array} \right.$$

where V^u and V^d represent the value of a derivative[1] whose underlying has a price of S_u and S_d, respectively. Setting the value of the portfolio (Eq. 9.3) equal to the value of the derivative at time t as well as after the next binomial step $t + dt$, we obtain the following three equations:

$$V(S, t) = \Delta(t)\, S(t) + g(t) \tag{9.5}$$

$$V(S_u, t + dt) = \Delta(t)\, u(t) S(t) B_q^{-1}(t) + g(t) B^{-1}(t) \tag{9.6}$$

$$V(S_d, t + dt) = \Delta(t)\, d(t) S(t) B_q^{-1}(t) + g(t) B^{-1}(t) \tag{9.7}$$

---

[1]In order to emphasize that this method is valid for all kinds of derivatives, we will continue to denote the value of the derivative with the letter V.

These can be easily rearranged[2] to express the number of underlyings, the cash amount and the value of the derivative at time t in terms of values at time $t+dt$:

$$\Delta(t) = \frac{V(S_u, t + dt) - V(S_d, t + dt)}{[u(t) - d(t)]S(t)/B_q(t)}$$

$$g(t) = \frac{u(t) \, V(S_d, t + dt) - d(t) \, V(S_u, t + dt)}{[u(t) - d(t)]/B(t)}$$

$$V(S, t) = B(t) \, [p(t)V(S_u, t + dt) + (1 - p(t))V(S_d, t + dt)] \qquad (9.8)$$

where

$$p(t) = \frac{B_q(t)/B(t) - d(t)}{u(t) - d(t)} \, . \qquad (9.9)$$

In Eq. 9.8, we have succeeded in expressing the unknown value of the derivative at time t in terms of quantities known at time t, namely $B_q(t)$, $B(t)$, $u(t)$ and $d(t)$, (in Chap. 13, we will show how the values of u and d are determined) and the (likewise unknown) derivative values at time $t + dt$. The reader might ask what this has accomplished. This expression will in deed prove to be useful if the value of the derivative at a future time is known. Such a future time is, for example, the maturity date T of the derivative. At this time, the value of the derivative as a function of the underlying price is given explicitly by its payoff profile, and as such, is known. The strategy is thus to repeat the procedure described above until reaching a time at which the value of the option is known (in most cases, maturity T). This procedure will be demonstrated below.

Equation 9.8 holds for European derivatives since it is implicitly assumed that the option still exists after a time step has been taken. In order to account for the possibility of exercising early as in the case of derivatives with American features, the derivative's value as given in Eq. 9.8 is compared with its intrinsic

---

[2]Subtracting Eq. 9.7 from Eq. 9.6 yields

$$V_S^u - V_S^d = \Delta(t) \, (u - d) S(t) B_q^{-1}(t) \, .$$

This allows us to isolate the $\Delta(t)$ term easily. Multiplying Eq. 9.7 by u, and Eq. 9.6 by d and subtracting the results yields

$$u \, V_S^d - d \, V_S^u = (u - d)g(t)B^{-1}(t) \, .$$

This can be readily solved for $g(t)$. Substituting the expressions thus obtained for $\Delta(t)$ and $g(t)$ into Eq. 9.5 yields the value of the derivative $V_S(t, T, K)$.

value at each node in the tree. Then the larger of the two values is taken as the derivative price at that node. For instance for American calls and puts with payoff profiles $S(t) - K$ and $K - S(t)$, respectively, Eq. 9.8 would be replaced by

$$C_S(t) = \max\left\{B(t)\left[p(t)C_S^u(t+dt) + (1-p(t))C_S^d(t+dt)\right], S(t) - K\right\}$$

$$P_S(t) = \max\left\{B(t)\left[p(t)P_S^u(t+dt) + (1-p(t))P_S^d(t+dt)\right], K - S(t)\right\}.$$

9.1.3 Forward Contracts

The evolution of the replicating portfolio consisting of underlyings and cash in a bank account is described by Eq. 9.3. According to Eq. 6.6, the evolution of a futures position is given by

$$V(S, t) \rightarrow \begin{cases} V(S_u, t+dt) = S_u(t+dt, T) - S(t, T) \\ V(S_d, t+dt) = S_d(t+dt, T) - S(t, T). \end{cases}$$

Here, it is not the value of the future at time t which is unknown (this is equal to zero since $K = S(t, T)$), but the forward price of the underlying $S(t, T)$. Setting the portfolio equal to the future at both time t and at the next time in the binomial tree $t + dt$ yields three equations:

$$0 = V(S, t) = \Delta(t) S(t) + g(t) \tag{9.10}$$

$$S_u(t+dt, T) - S(t, T) = V(S_u, t+dt) = \Delta(t) u(t)S(t)B_q^{-1}(t) + g(t)B^{-1}(t) \tag{9.11}$$

$$S_d(t+dt, T) - S(t, T) = V(S_d, t+dt) = \Delta(t) d(t)S(t)B_q^{-1}(t) + g(t)B^{-1}(t) \tag{9.12}$$

With the help of these three equations, the number of underlyings, the money in the bank account and the forward price at time t can (making use of the expression p defined in Eq. 9.9) be expressed[3] in terms of the forward price

---

[3]Subtracting Eq. 9.11 from Eq. 9.12 yields

$$S^u(t+dt, T) - S^d(t+dt, T) = \Delta(t) (u - d)S(t)B_q^{-1}(t).$$

at time $t + dt$:

$$\Delta(t) = \frac{S_u(t + dt, T) - S_d(t + dt, T)}{[u(t) - d(t)]S(t)B_q(t)} \tag{9.13}$$

$$g(t) = -\frac{S_u(t + dt, T) - S_d(t + dt, T)}{[u(t) - d(t)]B_q(t)}$$

$$S(t, T) = p(t)S_u(t + dt, T) + (1 - p(t))S_d(t + dt, T) \ .$$

9.2   Recombining Trees

9.2.1 The Underlying

If the parameters u and d are independent of time[4]

$$u(t + j \, dt) = u(t) \equiv u \ \forall j \ , \quad d(t + j \, dt) = d(t) \equiv d \ \forall j$$

then obviously $ud\,S(t) = du\,S(t)$ holds, i.e., an upward move followed by a downward move results in the same underlying price as a downward move followed by an upward move. Thus the tree is forced to *recombine*. This significantly reduces the number of possible nodes, making the computation much more efficient. Such a *recombining binomial tree* has the form depicted in Fig. 9.1. The probability for a *single* path ending at $S(T) = u^j d^{n-j} S(t)$ is

$$p'^j (1 - p')^{n-j} \ .$$

The number of all paths ending at $S(T) = u^j d^{n-j} S(t)$ can be deduced from permutation laws and is given by the *binomial coefficient*

$$\binom{n}{j} \equiv \frac{n!}{j!(n - j)!} \ .$$

This allows us to easily isolate $\Delta(t)$. Because of Eq. 9.10, $g(t) = -\Delta(t)S(t)$ holds, which immediately yields $g(t)$ if $\Delta(t)$ is known. Substituting the expressions for $\Delta(t)$ and $g(t)$ into Eq. 9.11 or Eq. 9.12 yields, after a simple calculation, the forward price $S(t, T)$.

[4] It is possible to construct recombining trees with time-dependent u and d, if at the same time some other constraint, e.g. constant time steps, is dropped.
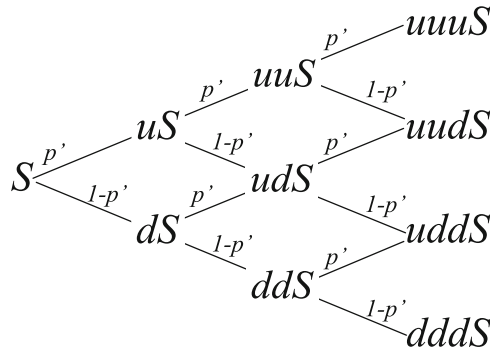
Fig. 9.1   The first steps in a recombining binomial tree

The probability in the above tree of arriving at a value $S(T) = u^j d^{n-j} S(t)$ regardless of the path taken to get there is equal to the number of such paths multiplied by the probability of realizing such a path.

$$P\left[S(T) = u^j d^{n-j} S(t)\right] = \binom{n}{j} p'^j (1 - p')^{n-j} = B_{n,p'}(j) . \qquad (9.14)$$

This is the definition of the probability density function of the *binomial distribution*, see Sect. A.4.2. This is how the binomial distribution enters into the binomial trees.

9.2.2 The Binomial Distribution for European Derivatives

In addition to the assumptions made at the beginning of Chap. 9 and the one just made, namely that the parameters u and d are constant over time, we will henceforth assume that the yields (interest rates and dividends) are constant over time as well, i.e., that Assumptions 9 and 12 from Chap. 4 hold[5]:

$$B(t + j\, dt) = B(t) \quad \forall j , \qquad B_q(t + j\, dt) = B_q(t) \quad \forall j . \qquad (9.15)$$

In consequence, the parameter p defined in Eq. 9.9 is time independent as well:

$$p(t + j\, dt) = p(t) = p \quad \forall j .$$

---

[5]On both sides we use here again the short notation defined in Eq. 9.4.

Equation 9.8 holds not only at time t but for other times as well, for example at time $t + dt$. This is true for both V_S^u and V_S^d:

$$V(S_u, t + dt) = B(t) \left[p \, V(S_{uu}, t + 2dt) + (1 - p) \, V(S_{ud}, t + 2dt) \right]$$
$$V(S_d, t + dt) = B(t) \left[p \, V(S_{ud}, t + 2dt) + (1 - p) \, V(S_{dd}, t + 2dt) \right] \, .$$

Substitution into Eq. 9.8 leads to an expression for $V(S, t)$ as a function of the derivative price at time $t + 2dt$. Analogous expressions can be obtained for V_S^{uu}, V_S^{ud}, etc. This recursive procedure performed iteratively for $n = (T - t)/dt$ binomial steps gives

$$V(S, t) = B(t, T) \sum_{j=0}^{n} \binom{n}{j} p^j (1 - p)^{n-j} V\left(u^j d^{n-j} S(t), T \right)$$

$$= B(t, T) \sum_{j=0}^{n} B_{n,p}(j) V\left(u^j d^{n-j} S(t), T \right) , \qquad (9.16)$$

where the second line was obtained by observing that $\binom{n}{j} p^j (1 - p)^{n-j}$ corresponds to a binomial probability density $B_{n,p}(j)$ but with parameters n and p (*not* with p' as in Eq. 9.14).

Thus the value of the derivative at time t has been expressed as a sum over its values at a later time T. If this time T is chosen to be the maturity then the value of the derivative at time t is written in terms of its payoff profile. This reads explicitly for European calls and puts

$$c_s(t) = B(t, T) \sum_{j=0}^{n} \binom{n}{j} p^j (1 - p)^{n-j} \max \left\{ 0, u^j d^{n-j} S(t) - K \right\}$$

$$p_s(t) = B(t, T) \sum_{j=0}^{n} \binom{n}{j} p^j (1 - p)^{n-j} \max \left\{ 0, K - u^j d^{n-j} S(t) \right\} \, .$$

Because of the maximum function appearing in the summand, the sum for the call is effectively taken over the values of j for which $u^j d^{n-j} S(t)$ is larger than K. This condition can be written as $(u/d)^j > d^{-n} K/S$. Taking the logarithm of both sides yields the equivalent condition

$$j > \ln \left(\frac{K}{S(t) d^n} \right) / \ln \left(\frac{u}{d} \right) \, .$$

The sum is taken over whole numbers j. The smallest whole number greater than the right-hand side in the above inequality is

$$y = 1 + \text{Trunc}\left(\ln(\frac{K}{S(t)d^n})/\ln(\frac{u}{d})\right) , \qquad (9.17)$$

where the function "Trunc" is defined as the greatest whole number smaller than the argument (decimal values are simply truncated and not rounded). The number y defined in Eq. 9.17 is thus the lower limit in the sum for the call (correspondingly, the sum for the put is taken over the whole numbers j ranging from 0 to the upper limit $y - 1$). The value of a call is thus

$$c_S(t) = S(t)B(t,T)\sum_{j=y}^{n}\binom{n}{j}p^j(1-p)^{n-j}u^jd^{n-j} - KB(t,T)\underbrace{\sum_{j=y}^{n}\binom{n}{j}p^j(1-p)^{n-j}}_{B_{n,p}(j\geq y)} .$$

According to Eq. A.42, the last sum is the probability that a binomially distributed random variable (where $B_{n,p}$ denotes the binomial distribution with parameters n and p) is greater than or equal to y. Under Assumption 9.15 that yields are constant, i.e., $B_r(t) = B_r$ independent of t, we can write

$$B(t,T) = \prod_{k=0}^{n-1}B(t+kdt) = B^n , \quad B_q(t,T) = B_q^n .$$

Now the first sum can be represented as a binomial probability as well:

$$c_S(t) = S(t)B_q(t,T)\sum_{j=y}^{n}\binom{n}{j}p^ju^j\frac{B^n}{B_q^n}(1-p)^{n-j}d^{n-j} \qquad (9.18)$$

$$- KB(t,T)\sum_{j=y}^{n}\binom{n}{j}p^j(1-p)^{n-j}$$

$$= S(t)B_q(t,T)\underbrace{\sum_{j=y}^{n}\binom{n}{j}\hat{p}^j(1-\hat{p})^{n-j}}_{B_{n,\hat{p}}(j\geq y)} - KB(t,T)\underbrace{\sum_{j=y}^{n}\binom{n}{j}p^j(1-p)^{n-j}}_{B_{n,p}(j\geq y)}$$

where[6]

$$\hat{p} = u\frac{B}{B_q}p \Rightarrow 1 - \hat{p} = d\frac{B}{B_q}(1 - p) \,.$$

The value of a European put can be determined analogously with the help of the binomial distribution. So the prices of European options expressed in terms of binomial distributions are

$$c_s(t) = B_q(t, T)\, S(t)\, B_{n,\hat{p}}(j \geq y) - B(t, T)K\, B_{n,p}(j \geq y)$$
$$p_s(t) = -B_q(t, T)\, S(t)\, \left[1 - B_{n,\hat{p}}(j \geq y)\right] + B(t, T)\, K\, \left[1 - B_{n,p}(j \geq y)\right] \,.$$
$$(9.19)$$

Note the similarity to the famous Black-Scholes equation (see, for example Eq. 8.6 or Eq. 8.7). The difference is that the *binomial* distribution appears in the above expression in place of the *normal* distribution. In Sect. 9.4, we will see that the binomial distribution for infinitesimally small intervals dt converges towards a normal distribution and thus the binomial model approaches the Black-Scholes model in the limit $dt \to 0$.

As another example of the above procedure we demonstrate how the forward price can be determined by iterating Eq. 9.13 for $n = (T - t)/dt$ binomial steps

$$S(t, T) = \sum_{j=0}^{n} \binom{n}{j} p^j (1 - p)^{n-j} u^j d^{n-j} S(t) \,,$$

---

[6]Writing $S^u = uS$ and $S^d = dS$ in the third equation in 9.13 and using Eq. 6.1 for $S(t, T)$ yields

$$\frac{B_q}{B} S(t) = p\, u\, S(t) + (1 - p)\, d\, S(t) \,.$$

Dividing by the left-hand side gives

$$1 = p\frac{B}{B_q}u + (1 - p)\frac{B}{B_q}d \,.$$

It then follows immediately that

$$1 - \hat{p} \equiv 1 - u\frac{B}{B_q}p = d\frac{B}{B_q}(1 - p) \,.$$

where the spot price $S(T)$ for the forward contract after n steps is substituted into the equation, since the maturity date T will have been reached after this time. Under the assumption of constant interest rates and dividend yields, we obtain

$$S(t, T) = \frac{B_q(t, T)}{B(t, T)} S(t) \sum_{j=0}^{n} \binom{n}{j} p^j u^j \frac{B^n}{B_q^n} (1 - p)^{n-j} d^{n-j}$$

$$= \frac{B_q(t, T)}{B(t, T)} S(t) \sum_{j=0}^{n} \binom{n}{j} \hat{p}^j (1 - \hat{p})^{n-j} = \frac{B_q(t, T)}{B(t, T)} S(t) \underbrace{B_{n,\hat{p}}(j \geq 0)}_{1} ,$$

corresponding to the result in Eq. 6.1 obtained solely on the basis of arbitrage considerations.

9.2.3 A Third Contact with the Risk-Neutral World

Neither for derivatives (see Eq. 9.8) nor for forward prices (see Eq. 9.13) does the probability p' for the underlying S to increase to S_u (see, for example Eq. 9.2) enter into any equation. The *valuation* of derivatives (or forward prices) does not depend on the probability that the underlying rises or falls! Instead, it depends on the value p as defined in Eq. 9.9. If we could interpret p as a kind of "artificial probability", and if for pricing purposes we could put ourselves into an "artificial world" in which, after one step, the price of the underlying is given by S_u with this "probability" p (and not with the probability p' as in the real world), then the probability for the underlying to decrease to S_d over one time step would be $(1 - p)$ in this "artificial world". The expression in brackets in Eq. 9.8 would then just be the *expectation* of the derivative price *one* time step later with respect to the probability p in the artificial world. This holds for *many* time steps as well, since the binomial density $B_{n,p}(j)$ appearing in Eq. 9.16 is the probability in the *artificial* world for the underlying to arrive at the value $S(T) = u^j d^{n-j} S(t)$; just as Eq. 9.14 was this probability in the *real* world. The sum in Eq. 9.16 over all the derivative values is then the *expectation* of the derivative value at the future time T in the artificial world.

We can summarize our observations in the following way: In an artificial world, where the probability of an up-move is p (and not p' as in the real

world), **today's price of a derivative is the expectation of its future price discounted back to today**.

$$V(S, t) = B(t, T) \, \mathrm{E}_p \, [V(S, T)] \text{ Derivat (z. B. Option) auf } S \,. \qquad (9.20)$$

The notation "$\mathrm{E}_p[]$" here means: "expectation with respect to the probability p".

Likewise, the forward price of the underlying in Eq. 9.13 is exactly the expectation of the underlying's price at the future time T with respect to the probability p:

$$S(t, T) = \mathrm{E}_p \, [S(T)] \text{ Forward Price of } S \,. \qquad (9.21)$$

In the case of forward prices (which are not tradable financial *instruments* themselves) the expectation is not discounted.

By substituting Eq. 6.1 for $S(t, T)$ into the above equation it follows that the dividend-adjusted spot price of the underlying (which is a tradable financial instrument) likewise can be expressed as the *discounted* expectation with respect to this probability

$$\tilde{S}(t, T) = B(t, T) \, \mathrm{E}_p \, [S(T)] \text{ Spot price } S \,. \qquad (9.22)$$

As in Sects. 7.1.3 and 7.2.1, it does not matter if the underlying is expected to rise or fall in the real world. This plays no role in the valuation of derivatives on the underlying. The valuation is independent of the expected changes in the underlying. In contrast to the real world where investors are compensated for taking the risk of investing in an underlying by the underlying's mean return, this mean return doesn't play any role at all in the artificial world used for pricing derivatives. This artificial world is thus neutral to the risk inherent in the underlying and is therefore called the *risk-neutral world* and the probability p is called the *risk-neutral probability.* We are again confronted with the *risk neutrality* described in Sects. 7.1.3 and 7.2.1.

This risk neutrality is caused by the fact that the option buyer would hedge himself against the risk of an unfavorable development of the option underlying by entering into a portfolio of shares and cash, which replicates the option pay off at expiry. Therefore, he has eliminated the risk and his total position is risk neutral. The cost for the derivative is identical to the cost of this replication strategy. Would the seller value the derivative differently,

an arbitrage opportunity would arise, since the potential buyer could apply the replication strategy in order to earn a risk-less profit. Likewise, because of this effect, different market participants would be able to agree on the same option price independently of their estimate of the future development of the underlying. Based on these arbitrage considerations making use of the fact that options and futures can be replicated by a portfolio consisting of the underlying and risk-free assets, it follows that a purely objective, risk-neutral probability p for an up-move in a risk-neutral world exists, eliminating any subjectivity (see Eq. 9.9). On the other hand, a derivative buyer, who does plan to hedge his derivative position by means of a strategy replicating the pay off profile (and also does not need this derivative to hedge some other existing positions), would rely on a valuation of the derivative based on his subjective judgment about the future market development. He would indeed seek his advantage in making a profit by means of the difference between risk neutral valuation and the real world development. Of course, this strategy would not be risk-free anymore.

To see how powerful Eqs. 9.20, 9.21 and 9.22 are, we have to be more specific. We now choose a stochastic process for the underlying. In what follows, we will assume that the relative changes of $S(t)$ behave as a random walk as in Eq. 2.17, i.e., we will rely on Assumption 7 from Chap. 4. But Eq. 2.17 was established to model the behavior of S in the *real* world.[7] We will show later,[8] however, that if S performs a random walk in the real world, it *also* performs a random walk in the risk-neutral world. The distribution and first moments of such a random walk are those given in Table 2.7 at the end of Sect. 2.3. Thus, the underlying is lognormally distributed with expectation

$$\langle S(T) \rangle = S(t) e^{(\mu + \sigma^2/2)(T-t)} \ . \tag{9.23}$$

In the risk-neutral world (i.e., in the world we need for pricing) this expectation has to be equal to the expectation $\mathrm{E}_p\left[S(T)\right]$ with respect to the risk-neutral probability p:

$$\langle S(T) \rangle \stackrel{!}{=} \mathrm{E}_p\left[S(T)\right] \ .$$

---

[7] A "risk-neutral" world was never mentioned in the vicinity of Eq. 2.17, nor in the whole of Chap. 2.

[8] We will explicitly show this in great detail and on a much more fundamental basis in Chap. 13 when we discuss the famous Girsanov theorem.

Substituting Eqs. 9.22 and 9.23 into this requirement completely determines the drift μ of the underlying in the risk-neutral world, i.e., the drift to be used for pricing:

$$S(t)e^{(\mu+\sigma^2/2)(T-t)} \stackrel{!}{=} \frac{\tilde{S}(t,T)}{B(t,T)}$$

$$\Leftrightarrow \quad \mu \stackrel{!}{=} \frac{1}{T-t} \ln\left(\frac{\tilde{S}(t,T)}{S(t)B(t,T)}\right) - \frac{\sigma^2}{2}.$$

Using the forward price equation 9.21 instead, we obtain the drift from the ratio of forward price to spot price:

$$S(t)e^{(\mu+\sigma^2/2)(T-t)} \stackrel{!}{=} S(t,T)$$

$$\Leftrightarrow \quad \mu \stackrel{!}{=} \frac{1}{T-t} \ln\left(\frac{S(t,T)}{S(t)}\right) - \frac{\sigma^2}{2}. \tag{9.24}$$

With a dividend yield q and continuous compounding, we obtain, for example,

$$\mu = \frac{1}{T-t} \ln\left(\frac{B_q(t,T)}{B(t,T)}\right) - \frac{\sigma^2}{2} = r - q - \frac{\sigma^2}{2}, \tag{9.25}$$

where the first equality holds as a result of the assumed dividend yield and the second is valid for continuous compounding. But this is exactly Eq. 7.19.

As was pointed out in Eq. 2.30, the drift μ is exactly equal to the expected return of the underlying. This means that the expected return in the risk-neutral world (i.e., with respect to the probability p) is objectively given through the risk-free interest rate and the dividends (or through the ratio of forward price to spot price) and the volatility, independent of an investor's opinion as to whether the price will rise or fall. The parameter μ thus determined is called the *risk-neutral yield* or the *risk-neutral drift*.

We note for later reference that all this holds for any arbitrary time span $T - t$, for instances also for one time step dt in a binomial tree:

$$\mu \, dt = \ln\left(\frac{B_q(t)}{B(t)}\right) - \frac{\sigma^2}{2}dt. \tag{9.26}$$

9.3   Random Walk and Binomial Parameters

Risk neutrality is the essential link connecting the stochastic model describing the underlying with the pricing method used for a derivative. This will be demonstrated for the binomial model. The parameters u and d must first be determined before the binomial model can be applied in pricing derivatives. The choice of these parameters has a significant influence on derivative and forward prices thus calculated. To make a reasonable choice of u and d further assumptions concerning the behavior of the underlying must be made, i.e., a stochastic process for the underlying must be specified. We will again assume that the relative changes of $S(t)$ behave as a random walk as in Eq. 2.17 and are therefore normally distributed with moments of the form given in Table 2.7 at the end of Sect. 2.3. On the other hand we know from the previous sections that in a binomial tree the underlying S is distributed according to the binomial distribution, see Eq. 9.14, where for pricing purposes we have to replace the real world probability p' by the risk-neutral "probability" p. We will now relate the random walk parameters μ and σ to the parameters u and d of the binomial tree by *matching the moments* of the random walk distribution to the moments of the binomial distribution. We will ensure that we work in the risk-neutral world (i.e., that we determine the parameters needed for pricing) by using p as defined in Eq. 9.9 for the binomial tree and by using the risk-neutral drift defined in Eq. 9.25 for the random walk.

In the following derivation, we will assume that the parameters u and d are constant over time until maturity. After j up-moves and $n - j$ down-moves, the final value $S(T)$ and thus the logarithm of the relative price change is

$$S(T) = u^j d^{n-j} S(t) \quad \Rightarrow \quad \ln\left(\frac{S(T)}{S(t)}\right) = j \ln\left(\frac{u}{d}\right) + n \ln(d) \ .$$

$S(T)$ (and thus j) is binomially distributed in our binomial model and from Eq. A.44, it follow that $\langle j \rangle = np$ and $\text{var}(j) = np(1 - p)$.

For the random walk model, on the other hand, the expectation and variance of the logarithmic changes of S are equal to the drift and the square of the volatility, each multiplied by the time difference $T - t$ (see the first column of Table 2.7). Thus matching the first two moments of the random

walk distribution to the distribution induced by the binomial tree yields

$$\mu(T - t) = \left\langle \ln \left(\frac{S(T)}{S(t)} \right) \right\rangle = \underbrace{\langle j \rangle}_{np} \ln \left(\frac{u}{d} \right) + n \ln(d)$$

$$\sigma^2(T - t) = \mathrm{var} \left(\ln \left(\frac{S(T)}{S(t)} \right) \right) = \underbrace{\mathrm{var}\,(j)}_{np(1-p)} \left(\ln \left(\frac{u}{d} \right) \right)^2 . \qquad (9.27)$$

Because $dt = (T - t)/n$ this can be written as

$$\mu dt = p \ln \left(\frac{u}{d} \right) + \ln(d)$$

$$\sigma^2 dt = p(1 - p) \left(\ln \left(\frac{u}{d} \right) \right)^2 .$$

Now we use Eq. 9.9 for the risk-neutral probability p and Eq. 9.26 for the risk-neutral drift μ to establish a system of two (non-linear!) equations for the two unknown binomial parameters u and d:

$$\ln \left(\frac{B_q}{B} \right) - \frac{\sigma^2}{2} dt = \frac{B_q/B - d}{u - d} \ln \left(\frac{u}{d} \right) + \ln(d)$$

$$\sigma^2 dt = \frac{(B_q/B - d)\,(u - B_q/B)}{(u - d)^2} \left(\ln \left(\frac{u}{d} \right) \right)^2 . \qquad (9.28)$$

There exist several closed form solutions to this system which are exact up to linear order in dt. One such solution is given by

$$u = \frac{B_q}{B} e^{-(\sigma^2/2)dt + \sigma \sqrt{dt}} \quad , \quad d = \frac{B_q}{B} e^{-(\sigma^2/2)dt - \sigma \sqrt{dt}} . \qquad (9.29)$$

Inserting this into Eq. 9.9 for the risk-neutral probability yields

$$p = \frac{e^{(\sigma^2/2)dt} - e^{-\sigma \sqrt{dt}}}{e^{+\sigma \sqrt{dt}} - e^{-\sigma \sqrt{dt}}} = \frac{e^{(\sigma^2/2)dt} - e^{-\sigma \sqrt{dt}}}{2 \, \sinh(\sigma \sqrt{dt})} ,$$

where we have used the definition of the hyperbolic sine function in the last step. Using Eq. 9.26 (which is equivalent to $\exp(\mu\,dt) = \exp\left(-\sigma^2 dt/2\right) B_q/B$) we can bring the parameters u and d into a more

intuitive form using the risk-neutral drift:

$$u = e^{\mu\,dt + \sigma\sqrt{dt}} \quad , \quad d = e^{\mu\,dt - \sigma\sqrt{dt}} \; .$$

Thus, in this solution the drift only appears in the parameters u and d, while the probability p is determined solely from knowledge of the volatility. For small values of dt and assuming continuous compounding, the Taylor series representation of the exponential function expanded up to linear terms in dt gives

$$u \approx 1 + \sigma\sqrt{dt} + (q-r)\,dt \; , \quad d \approx 1 - \sigma\sqrt{dt} + (q-r)\,dt \; , \quad p \approx 1/2 \; .$$
(9.30)

Since in this solution both the volatility and the risk neutral drift appear in u and d, we must assume constant volatilities and because of Eq. 9.25 also constant yields and dividends, i.e., Assumptions 9, 11 and 12 from Chap. 4, to ensure that the parameters u and d are constant over time and, in consequence, that the tree recombines.

Another frequently used solution of Eq. 9.28 for which it suffices to assume constant volatilities (Assumption 11) is

$$u = e^{+\sigma\sqrt{dt}} \quad , \quad d = e^{-\sigma\sqrt{dt}} \Rightarrow p = \frac{B_q/B - e^{-\sigma\sqrt{dt}}}{2\;\sinh(\sigma\sqrt{dt})} = \frac{e^{(q-r)dt} - e^{-\sigma\sqrt{dt}}}{2\;\sinh(\sigma\sqrt{dt})} \; ,$$
(9.31)

where the last step is of course only valid for continuous compounding. In this solution, the volatility alone completely determines the parameters u and d. Observe that $u(t) = 1/d(t)$ holds. As long as the volatility is constant (allowing the parameters u and d to remain constant over time), the tree is recombining since the starting price is recovered after an up-move followed by a down-move:

$$u(t)\,d(t+dt) = u(t+dt)d(t) = u(t)/u(t) = 1 \; .$$

The ease in the construction and analysis of binomial trees resulting from this relation prompts us to utilize the solution given by Eq. 9.31 exclusively in the remainder of this book whenever we use binomial trees. For small time intervals dt, the Taylor series representation of the exponential function expanded up to and including terms of linear order in dt yields the following

approximations for u, d and p:

$$u \approx 1 + \sigma\sqrt{dt} + \frac{\sigma^2}{2}dt, \ \ d \approx 1 - \sigma\sqrt{dt} + \frac{\sigma^2}{2}dt \ , \ \ p \approx \frac{1}{2}\left(1 + \frac{\mu}{\sigma}\sqrt{dt}\right) \ ,$$

where again the risk-neutral drift μ is used to simplify the last expression.

A detailed demonstration of the application of binomial trees for the valuation of an option portfolio is provided in the Excel workbook BINOMIAL-TREE.XLS. In anticipation of Chap. 12, the evaluation of the *Greeks* (derivatives of the option price with respect to its parameters) using binomial trees also receives attention. This workbook can be used as a small but fully functioning option calculator (as always, the yellow fields are the input fields).

9.4   The Binomial Model with Infinitesimal Steps

In this section, the Black-Scholes option pricing formula is derived directly from the binomial model for European options as given by Eq. 9.19. A deeper insight into the relationship between these two important methods in option pricing (finding solutions to a differential equation on the one hand and the calculation of (discounted) expectations on the other) can be gained from an understanding of this derivation. The reader less interested in mathematics may choose to continue on to the next section.

A classical result from statistics, the *Moivre-Laplace theorem,* states that the binomial distribution converges towards a normal distribution as the number n of the observed trials approaches infinity. The statement of the theorem in integral form can be expressed as

$$B_{n,p}\left(a \le \frac{j - np}{\sqrt{np(1-p)}} \le b\right) \overset{n\to\infty}{\longrightarrow} N(b) - N(a) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}dz \ .$$

The left-hand side of the equation denotes the probability that the standardized form of a *binomially distributed* random variable j (i.e., j less its expectation divided by its standard deviation) will lie in the interval between a and b, while the right-hand side is simply the probability that a *standard normally distributed* random variable will take on values lying within the same interval.

We can exploit this theorem to see what happens to Eq. 9.19 as the time interval dt converges towards zero, i.e., as the number n of steps in the

binomial tree approaches infinity. This will be demonstrated for a call, the procedure for a put being completely analogous.

We need, for example, to determine the limit of $B_{n,p}(y \leq j)$ in Eq. 9.19 For arbitrary constants f and g with $g > 0$, the probability that $y \leq j$ is of course equal to the probability that $(y - f)/g \leq (j - f)/g$. We take advantage of this fact[9] to manipulate $B_{n,p}(y \leq j)$ into a suitable form to apply the Moivre-Laplace theorem:

$$
\begin{aligned}
B_{n,p}(y \leq j) &= B_{n,p}\left(\frac{y - np}{\sqrt{np(1-p)}} \leq \frac{j - np}{\sqrt{np(1-p)}} \leq \infty \right) \\
&\xrightarrow{n \to \infty} \underbrace{N(\infty)}_{1} - N\left(\frac{y - np}{\sqrt{np(1-p)}} \right) \\
&= N\left(\frac{np - y}{\sqrt{np(1-p)}} \right) ,
\end{aligned}
$$

where in the last step the symmetry property of the normal distribution, Eq. A.54, is used.

Equations 9.27 and 9.17 deliver the necessary elements for computing $(np - y)/\sqrt{np(1-p)}$:

$$
np = \frac{1}{\ln(u/d)} \left(\mu(T - t) - n \ln(d) \right) , \quad \sqrt{np(1-p)} = \frac{1}{\ln(u/d)} \sigma \sqrt{T - t}
$$

and

$$
y = \frac{\ln\left(\frac{K}{S(t)d^n} \right)}{\ln(u/d)} + \varepsilon = \frac{\ln\left(\frac{K}{S(t)} \right) - n \ln(d) + \varepsilon \ln(u/d)}{\ln(u/d)} \quad \text{mit } 0 < \varepsilon \leq 1 .
$$

Here, ε represents the difference between $\ln(K/Sd^n)/\ln(u/d)$ and the smallest whole number greater than this value. We will show immediately that the term $\varepsilon \ln(u/d)$ becomes arbitrarily small. Substituting accordingly yields the argument for the standard normal distribution above:

$$
\frac{np - y}{\sqrt{np(1-p)}} = \frac{\ln(S(t)/K) + \mu(T - t) - \varepsilon \ln(u/d)}{\sigma \sqrt{T - t}} .
$$

---

[9]With the choice $f = np$ und $g = \sqrt{np(1-p)}$.

In both solutions 9.29 and 9.31 the ratio u/d converges towards 1 as $n \to \infty$ (i.e., $dt \to 0$), thus by continuity, $\ln(u/d)$ converges towards zero. Thus, the limit for infinitely many binomial steps becomes

$$\frac{np - y}{\sqrt{np(1 - p)}} \xrightarrow{n \to \infty} \frac{\ln(S(t)/K) + \mu(T - t)}{\sigma\sqrt{T - t}} = x - \sigma\sqrt{T - t} .$$

Using now Eq. 9.25 for the risk-neutral drift we can write

$$\frac{np - y}{\sqrt{np(1 - p)}} \xrightarrow{n \to \infty} x - \sigma\sqrt{T - t}$$

where we have defined the abbreviation x as in Eq. 8.5:

$$x = \frac{1}{\sigma\sqrt{T - t}} \ln\left(\frac{S(t)\, B_q(t, T)}{K\, B(t, T)}\right) + \frac{1}{2}\sigma\sqrt{T - t} .$$

The limit of $B_{n,p}(y \le j)$ is thus established. Proceeding analogously, we can calculate the limit of the other binomial probability in Eq. 9.19. In summary, for an infinite number of binomial steps in a finite time interval, the behavior of the binomial distribution is given by

$$B_{n,p}(y \le j) \xrightarrow{n \to \infty} N\left(x - \sigma\sqrt{T - t}\right) , \quad B_{n,\hat{p}}(y \le j) \xrightarrow{n \to \infty} N(x) .$$

Using these convergence relations, we obtain the value of a call as the number of binomial steps approaches infinity to be

$$c_s(t) \stackrel{n \to \infty}{=} S(t)B_q(t, T)N(x) - K B(t, T)N(x - \sigma\sqrt{T - t}) . \qquad (9.32)$$

This is in complete agreement with Eq. 8.6 and is thus (again!) the famous *Black-Scholes option pricing formula*.

9.4.1 Components of the Black-Scholes Option Pricing Formula

In the above section the Black-Scholes formula was derived from Eqs. 9.18 and 9.19. We can see from this derivation that the cumulative normal distribution found next to the discounted strike price $B(t, T)K$ is the risk-neutral probability for the price of the underlying to be larger than the strike price.

Table 9.1 Interpretation of the various components in the Black-Scholes option pricing formulae for plain vanilla calls and puts. All probabilities mentioned are risk-neutral. The lower four terms are required for replicating (hedge) calls resp. puts

$N(x - \sigma\sqrt{T-t})$	Risk-neutral exercise probability for call
$N(-x + \sigma\sqrt{T-t})$	Risk-neutral exercise probability for put
$KB(t,T)$	Present value (PV) of cash flow at exercise
$B_q(t,T)N(x)$	Number of underlyings to buy for replicating the call
$B_q(t,T)N(-x)$	Number of underlyings to buy for replicating the put
$KB(t,T)N(x - \sigma\sqrt{T-t})$	Amount to be borrowed for call replication
$KB(t,T)N(-x + \sigma\sqrt{T-t})$	Amount to be borrowed for put replication

This is referred to as the *risk-neutral exercise probability*. A comparison with the replicating portfolio in Eq. 9.3 shows that the number $\Delta(t)$ of underlyings needed to replicate the option is given by the factor next to $S(t)$ in Eq. 9.32, namely $B_q(t,T)N(x)$, while the amount $g(t)$ of money in the bank account is given by the second summand in Eq. 9.32. The intuitive interpretations of these values in the Black-Scholes formulae for puts and calls are collected in Table 9.1.

9.5 Trinomial Trees

Trinomial trees present us with an alternative method to binomial trees. The form of a trinomial tree is represented graphically in Fig. 9.2. The jth step of the tree at time t_j is connected, not with two other nodes in the next step (as was the case for the binomial tree), but with three. The price paid for this additional degree of freedom is additional computational effort. The advantage is that a trinomial tree can always be constructed in such a way that it recombines and in addition, achieves the same degree of accuracy as the binomial tree with fewer time steps. The trinomial tree has $2j + 1$ nodes after j steps where the time t is indexed with $j = 0$. The length of the time steps may vary. The value of the underlying at the ith node after j steps is denoted by S_{ji} where

$$i = -j, -j + 1, \ldots, j - 1, j \ .$$

Each node at time step j branches into three nodes at time step $j + 1$, with a probability being associated with each of these branches.[10] Starting from S_{ji},

---

[10]All probabilities appearing in this context are risk-neutral probabilities.
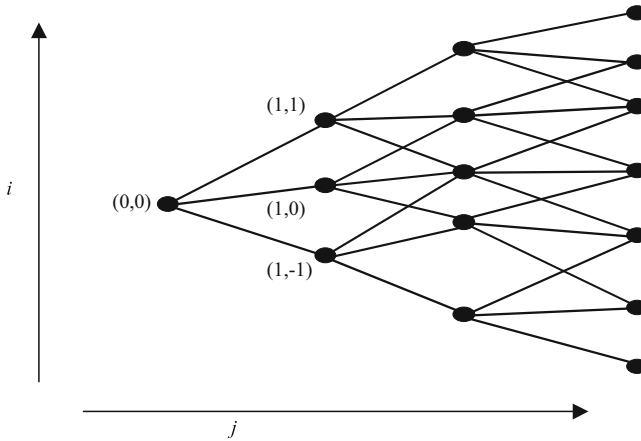
Fig. 9.2 A simple trinomial tree

we denote by p^+ the probability that the underlying will increase from S_{ji} to the value $S_{j+1,i+1}$ at time step $(j + 1)$. Correspondingly, denote by p^0 and p^- the probabilities that the underlying at time step $(j + 1)$ will take on the values $S_{j+1,i}$ and $S_{j+1,i-1}$, respectively. Each of these probabilities must be ≥ 0 and ≤ 1. In addition:

$$p^+ + p^0 + p^- = 1 . \tag{9.33}$$

For the *bi*nomial tree,[11] a portfolio can be constructed consisting of a position Δ in the underlying and money g in a bank account whose value replicates the option price exactly at each time step. This is also possible for trinomial trees. Though, because of the addition of the *third* attainable value for the underlying after one time step, just *two* parameters Δ and g are not sufficient to determine a replicating portfolio that replicates exactly all *three* possible states after one time step. Therefore, we need to make a further choice before all parameters could be determined unequivocally. Because of this additional degree of freedom, trinomial trees can be more flexibly adopted for different purposes.

To do so, it is sufficient to choose the probabilities p^+, p^0 and p^- and the nodes S_{ji} so that the tree reflects the probability distribution of the underlying. Again, we assume that the underlying price is lognormally distributed (this

---

[11]The binomial tree usually assumes a constant time step dt. In the most general case, this assumption is not necessary.

corresponds to Assumption 7 from Chap. 4). The lognormal distribution is completely determined by two parameters, the expectation and the variance of the logarithm. Thus, two conditions are sufficient to adapt the trinomial tree to a lognormal distribution. Like in Eq. 9.23, the expectation for the lognormally distributed random variable after a time step of length dt starting from the node S_{ji} (with the risk-neutral drift Eq. 9.25) is given by

$$E\left[S(t_{j+1})\right] = S_{ji}e^{(r-q)dt} .$$

On the other hand, from the tree we have

$$E\left[S(t_{j+1})\right] = p^+ S_{j+1,i+1} + p^0 S_{j+1,i} + p^- S_{j+1,i-1} .$$

Setting these two expressions equal to one another yields one equation for the determination of the probabilities:

$$S_{ji}e^{(r-q)dt} = p^+ S_{j+1,i+1} + p^0 S_{j+1,i} + p^- S_{j+1,i-1} . \tag{9.34}$$

Analogously, taking the expression for the variance of the lognormal distribution shown in Table 2.7 at the end of Sect. 2.3 we have

$$\mathrm{Var}\left[S(t_{j+1})\right] = S_{ij}^2 e^{2(r-q)dt}\left(e^{\sigma^2 dt} - 1\right) .$$

It is sometimes easier to work with the expectation of $S^2(t_{j+1})$ rather than the variance and such is the case here. With the help of Eq. A.7 we obtain this expectation as

$$E\left[S^2(t_{j+1})\right] = S_{ji}^2 e^{2(r-q)dt}\left(e^{\sigma^2 dt} - 1\right) + S_{ji}^2 e^{2(r-q)dt}$$
$$= S_{ji}^2 e^{2(r-q)dt} e^{\sigma^2 dt} .$$

Expressed in terms of the probabilities for the trinomial tree, the same expectation is given by

$$E\left[S^2(t_{j+1})\right] = p^+ S_{j+1,i+1}^2 + p^0 S_{j+1,i}^2 + p^- S_{j+1,i-1}^2 .$$

Combining the two above expressions gives

$$S_{ji}^2 e^{2(r-q)dt} e^{\sigma^2 dt} = p^+ S_{j+1,i+1}^2 + p^0 S_{j+1,i}^2 + p^- S_{j+1,i-1}^2 . \tag{9.35}$$

Equations 9.33, 9.34 and 9.35 are sufficient to fit the trinomial tree to the lognormal distribution. Of course, only three of the six parameters p^+, p^0, p^-, $S_{j+1,i+1}$, $S_{j+1,i}$ and $S_{j+1,i-1}$ will be determined. In general, these will be the probabilities. The nodes can then be arbitrarily selected.

The actual option pricing now proceeds as for a binomial tree. V_{ji} denotes the price of the option at the ith node of the jth time step. We assume that the tree consists of N time steps with $j = 0, 1, \ldots, N$. To price a European option, the nodes are initialized with the payoff profile of the option at maturity $t_N = T$. In the case of a call option we have:

$$V_{Ni} = \max(S_{Ni} - X, 0) \,.$$

The calculation then *rolls backwards* through the tree. The option value is calculated iteratively for a time step using the values just calculated at the next time step starting with $j = N - 1$ and working back to $j = 0$:

$$V_{ji} = B(t_j, t_{j+1}) \left[p^+ V_{j+1,i+1} + p^0 V_{j+1,i} + p^- V_{j+1,i-1} \right] \,.$$

V_{00} is the present value of the option at time $t_0 = t$ (assuming that S_{00} is the price of the underlying at $t = t_0$). It should be emphasized that the model admits both time-dependent interest rates and volatilities. To take this into consideration either the nodes need to be selected accordingly or the probabilities must be made time-dependent. American options are treated in the same manner as they are treated in binomial trees. Barrier options should be calculated by choosing the nodes such that they lie *directly* on the barrier.

9.5.1 The Trinomial Tree as an Improved Binomial Tree

After *two* time steps a recombining binomial tree has exactly three distinct nodes. This is equal to the number of nodes in the trinomial tree after *one* step. Since the nodes of the trinomial tree can be freely chosen, it is possible to generate a *tri*nomial tree (with an even number of time steps) corresponding to any given recombining *bi*nomial tree. Such a trinomial tree yields the exact same results as the binomial tree, but in only half the time steps. This will be demonstrated for the binomial tree with parameters u, d, and p as given in Eq. 9.31 serving as an example. Starting from the node S_{ij}, we can choose the

nodes at time $t_{j+1} = t_j + 2\,dt$ as follows:

$$S_{j+1,i+1} = u^2 S_{ij}$$
$$S_{j+1,i} = S_{ij}$$
$$S_{j+1,i-1} = d^2 S_{ij} \ .$$

where dt is the length of one time step in the binomial tree and hence $2\,dt$ is the length of one time step in the trinomial tree. The probabilities for the trinomial tree are easily obtained from the probability p in the binomial tree:

$$p^+ = p^2 \ , \quad p^0 = 2p(1-p) \ , \quad p^- = (1-p)^2 \ .$$

The values for the probabilities are consistent with those in Eqs. 9.33, 9.34 and 9.35.

The trinomial tree converges faster than the corresponding binomial tree because it requires only half as many steps. Approximately half of the nodes in the binomial tree need not be computed. This advantage is, however not quite as great as it may seem at first glance. It is known that the results of the binomial tree oscillate strongly when the number of time steps increases by one. The best results are obtained by averaging two calculations with N and $N + 1$ time steps (which doubles the required computation time). This trick cannot be exploited when using trinomial trees. Moreover, the parameters specified above are not an optimal choice for the trinomial tree.

10

Check for updates

Numerical Solutions Using Finite Differences

One of the best known and widely used numerical methods to solve partial differential equations in finance and elsewhere is the *finite difference method*. Finite difference methods are very powerful and flexible as well. They can be applied to a wide variety of various derivatives. Different exercise modes, including European (exercise at expiry only), American (exercise at any time) or Bermudan (exercise at a limited set of exercise dates), could be implemented without major problems. In comparison with the tree methods introduced in Chap. 9, the finite difference method possesses superior convergence features, which justifies the greater initial effort for its implementation. Therefore, we will provide a very detailed discussion of this important method.

Finite difference methods approximate the partial derivatives appearing in partial differential equations like the Black-Scholes equation 7.8 using finite difference quotients. The equation is then solved on a *grid* spanned by the linearly independent variables (for example, time t and price S of the underlying) appearing in the PDE.[1] Doing so, we obtain a solution surface which represents the price on each of the grid points (S, t). In general, a PDE has an unbounded number of solutions. Usually we are only interested in a solution which satisfies specific boundary and/or initial conditions. The finite difference method requires the specification of both boundary and initial conditions.

---

[1]In the following we will often use the abbreviation *PDE* for "partial differential equation", as is common practice in the related literature.

Since finite differences can be applied in quite general settings, we will require only the assumptions needed for arbitrage free trading, i.e., Assumptions 1, 2, 3,5 from Chap. 4, with the additional Assumption 6. Assumption 6 ensures that the variables in the problem are continuous. This is necessary if we want to obtain differentiable solutions. In order to provide a manageable overview of finite differences, we will restrict the treatment in the examples given here to random walks (Assumption 7) with non-stochastic interest rates and volatilities (Assumptions 8 and 10, respectively). In addition, we will not consider counter party default risk (Assumption 4).

10.1 Discretizing the Black-Scholes Equation

Below, the Black-Scholes Differential Eq. 7.8 will be solved numerically with the help of finite difference methods. In the literature, Eq. 7.8 is often transformed into an equation based on a new variable given by $Z = \ln(S)$. Such an equation has the advantage that the coefficients no longer depend explicitly on S. On occasion, it is claimed that a further advantage of this change of variable is that a uniform grid in Z is numerically more efficient than a uniform grid in S. However, the differences are usually negligible and this argument does not hold for barrier options. Furthermore, it is often preferable to use a non-uniform grid. We could distribute the grid points logarithmically, for example so that the non-uniform grid in S corresponds to a uniform grid in Z. In view of these considerations, we will continue to use the Black-Scholes equation in the form given by 7.8. But we will present the finite difference method for general *non-uniform grids*.

The finite difference method now consists in determining the value V of a financial instrument on a grid with coordinates S and t by approximating the partial derivatives with *finite* differences. We will restrict the discussion here to a rectangular grid allowing, however, the distance between grid points to be non-uniform. Such a grid is completely determined by the grid points in the S and t directions denoted by:

$$S_i \; ; \;\; i = 0, 1, 2, \ldots, M \qquad\qquad (10.1)$$
$$t_j \; ; \;\; j = 0, 1, 2, \ldots, N$$

We introduce the notation $W_{i,j} = W(S_i, t_j)$ for the solution's approximation in order to distinguish it from the exact solution $V(S_i, t_j)$ evaluated at points on the grid.

For the sake of clarity we will at several instances start with a grid whose time steps as well as the steps in the direction of the underlying price are uniformly spaced. For such *uniform grids* the spacing between grid points is simply

$$S_i - S_{i-1} = \delta S \; \forall i \implies S_i = S_0 + i \, \delta S \tag{10.2}$$

$$t_j - t_{j-1} = \delta t \; \forall j \implies t_i = t_0 + i \, \delta t \; .$$

We will then, however, always generalize our discussion to non-uniform grids.

The fundamental idea behind the method presented here is to determine the value of the derivative from the values at neighboring time points. Since the value of the derivative (as a function of the underlying's price) is usually known at maturity (it is given a priori by the payoff profile $P(S)$ as a function of S), we proceed using the strategy of starting at maturity $t_N = T$ and calculating *backwards* to time t_{N-1}, from there calculating back to t_{N-2} and so on. To accomplish this, we express the value of the instrument in terms of its Taylor series with respect to time, and express the time derivatives appearing in the Taylor expansion in terms of the derivatives with respect to the underlying price by using the Black-Scholes PDE. Here, the similarity to *backward induction* as described in Sect. 14.6.1 is not accidental.

Using the Black Scholes equation 7.8, the partial derivative of V with respect to time is expressed in terms of derivatives with respecto to S:

$$\frac{\partial V(S,t)}{\partial t} = r(t)V(S,t) \; - [r(t) - q(t)] \, S \frac{\partial V(S,t)}{\partial S} - \frac{1}{2}\sigma^2(S,t)S^2\frac{\partial^2 V(S,t)}{\partial S^2} \; . \tag{10.3}$$

The partial derivatives with respect to S in this equation are approximated by quotients of finite differences, as will be explicitly demonstrated in the next sections.

10.1.1   The Explicit Method

In the explicit method, the Taylor series expansion is used to calculate values at an *earlier* time $t - \delta t$ from the values at time t:

$$V(S, t - \delta t) = V(S, t) - \delta t \frac{\partial V(S,t)}{\partial t} + O(\delta t^2) \; .$$

The last term on the right-hand side of the equation states that no terms will be considered which are of order two or greater with respect to the time difference δt. If the time difference is small enough, we can assume that the time dependence of $V(S, t)$ can be adequately described if we simply neglect terms of order $O(\delta t^2)$. Note that on the left-hand side of the above equation, the derivative of the value evaluated at time $t - \delta t$ appears, whereas the right-hand side consists of terms evaluated at time t. The trick is now *not* to express the right-hand side in terms of a difference quotient in time (we would not have accomplished anything by doing so since it would involve introducing another time point $t - \delta t$ or $t + \delta t$) but rather to express the time derivative in terms of partial derivatives with respect to S obtained from the Black-Scholes equation 7.8. These partial derivatives with respect to S are evaluated at time t so that the value of the derivative at an earlier time $t - \delta t$ can in fact be recovered solely from information available at time t. Here, we *explicitly* calculate earlier values from those (known) values from a later time. This method is thus referred to as the *explicit method.*

10.1.2   The Implicit Method

In the implicit method, we use the Taylor series expansion in the time variable to obtain an expression for values at a *later* time $t + \delta t$ from the values at time t:

$$V(S, t + \delta t) = V(S, t) + \delta t \frac{\partial V(S, t)}{\partial t} + O(\delta t^2) .$$

In this case, later (known) values are expanded in terms of earlier (unknown) ones. This expansion can only be used *implicitly* to calculate the unknown values from the known ones, hence the name *implicit method.*

10.1.3   Combinations of Explicit and Implicit Methods (Crank-Nicolson)

The two methods described above can be combined by taking a linear combination of the two respective Taylor series expansions. To avoid the appearance of three different time points in the resulting expression ($t - \delta t$, t and $t + \delta t$) a change in variable in one of the Taylor series should be made.

For example, the transformation $t \rightarrow t + \delta t$ in the Taylor series expansion for the explicit method (and dividing by δt) yields an equation of the form:

$$\frac{V(S, t + \delta t) - V(S, t)}{\delta t} = \frac{\partial V(S, t + \delta t)}{\partial t} + O(\delta t) \, . \qquad (10.4)$$

Here, we have rearranged the Taylor series so that the left-hand side is written in terms of a *difference quotient* in t, while on the right-hand side a *differential quotient* appears (which will later be replaced by a difference quotient with respect to the underlying price). Note that dividing through by δt has the effect of reducing the order of the error term to a linear order in δt. An analogous procedure for the implicit method allows its respective Taylor series to be rearranged as well (here, a variable transformation is unnecessary since the series is already expressed in terms of t and $t + \delta t$):

$$\frac{V(S, t + \delta t) - V(S, t)}{\delta t} = \frac{\partial V(S, t)}{\partial t} + O(\delta t) \, . \qquad (10.5)$$

The only thing now distinguishing the two expressions is that the differential quotient on the right-hand side (and thus, the difference quotients with respect to S yet to be determined) is written in terms of the time t, whereas in the expression derived from the explicit method, the differential quotient is expressed in terms of the later time point $t + \delta t$. Naturally, the equality holds if we take any linear combination of the two equations:

$$\frac{V(S, t + \delta t) - V(S, t)}{\delta t} = (1 - \theta)\frac{\partial V(S, t + \delta t)}{\partial t} + \theta\frac{\partial V(S, t)}{\partial t} + O(\delta t) , 0 \leq \theta \leq 1 \, .$$

Note the following correspondence[2] between the notation above and the discrete notation introduced in Eq. 10.1:

$$S \mathrel{\widehat{=}} S_i \, , \quad t \mathrel{\widehat{=}} t_j \, , \quad t + \delta t \mathrel{\widehat{=}} t_{j+1}$$
$$W_{i,j} = W(S_i, t_j) \mathrel{\widehat{=}} V(S, t)$$
$$W_{i,j+1} = W(S_i, t_{j+1}) \mathrel{\widehat{=}} V(S, t + \delta t) \, .$$

---

[2]The sign $\widehat{=}$ means "corresponds to".

In the discrete notation the above equation reads

$$\frac{W_{i,j+1} - W_{i,j}}{t_{j+1} - t_j} = (1 - \theta)\frac{\partial V(S_i, t_{j+1})}{\partial t} + \theta\frac{\partial V(S_i, t_j)}{\partial t} + O\left(t_{j+1} - t_j\right), 0 \leq \theta \leq 1 .$$

(10.6)

In this notation, the equation holds for *non*-uniform grids, as well. Setting $\theta = 1$, we obtain the implicit method, with $\theta = 0$ the explicit method. The particular choice of $\theta = 1/2$ has a special name. It is known as the *Crank-Nicolson method.*

Equations 10.4 and 10.5 can also be interpreted as follows: the differen*ce* quotients ("finite differences") on the left-hand side of the equations are approximations of the differen*tial* quotients with respect to time found on the right-hand side of the equations.

The above approximations are exact up to *linear* terms in δt. There are several methods available for approximating partial derivatives using finite differences. A greater accuracy can be obtained if, for instance, all three time points t_{j-1}, t_j and t_{j+1} are included in the finite differences approximating the time derivative. Then the approximation is exact up to second order. This three-time procedure requires that the time derivative be approximated by a carefully selected convex combination of forward, backward and symmetric finite differences. Another possibility is to use symmetric finite differences. This also gives an approximation exact up to second order, however, does not lead to a stable procedure for solving the differential equation. In this book we will not pursue such more precise approximations for the time derivative.

10.1.4 Symmetric Finite Differences of the Underlying Price

For the sake of consistency, the difference quotients with respect to the underlying price S should be exact up to order $O(\delta S^2)$ since it follows from the random walk assumption that $dS \sim \sqrt{dt}$ and thus

$$\delta t \sim \delta S^2 .$$

This means that in order to attain the same degree of accuracy as in the time direction, the approximation in the S direction must be exact up to order $O(\delta S^2)$. To achieve this we will use *symmetric differences* to approximate the

derivatives of first and second order with respect to the underlying needed for Eq. 10.3.

In order to demonstrate the concepts without making the notation unnecessarily complicated, we begin by assuming that the distance between grid points is constant as in Eq. 10.2 and then generalize to non-uniform grids. As above, we expand the value function in its Taylor series, this time in the S dimension:

$$V(S - \delta S, t) = V(S, t) - \delta S \frac{\partial V(S, t)}{\partial S} + \frac{1}{2} \delta S^2 \frac{\partial^2 V(S, t)}{\partial S^2}$$

$$- \frac{1}{6} \delta S^3 \frac{\partial^3 V(S, t)}{\partial S^3} + O(\delta S^4)$$

$$V(S + \delta S, t) = V(S, t) + \delta S \frac{\partial V(S, t)}{\partial S} + \frac{1}{2} \delta S^2 \frac{\partial^2 V(S, t)}{\partial S^2}$$

$$+ \frac{1}{6} \delta S^3 \frac{\partial^3 V(S, t)}{\partial S^3} + O(\delta S^4) \, .$$

Subtracting the first equation from the second and subsequently dividing by δS yields an approximation of the first derivative which is exact up to second order $O(\delta S^2)$. Adding the two equations and then dividing by δS^2 yields an approximation of the second derivative which is also exact up to second order $O(\delta S^2)$:

$$\frac{\partial V(S, t)}{\partial S} = \frac{V(S + \delta S, t) - V(S - \delta S, t)}{2 \, \delta S} + O(\delta S^2) \qquad (10.7)$$

$$\frac{\partial^2 V(S, t)}{\partial S^2} = \frac{V(S + \delta S, t) - 2 \, V(S, t) + V(S - \delta S, t)}{\delta S^2} + O(\delta S^2) \, .$$

For general, *non*-uniforms grids, the above expressions are somewhat more complicated and their derivation is a bit more technical. The principle, however, remains the same: the partial derivatives of V with respect to the underlying evaluated at point S_i can be approximated up to order two with symmetric differences. To this end, V evaluated at S_{i-1} and S_{i+1}, is first expanded in its Taylor series about the points S_i:

$$V(S_{i-1}, t) = V(S_i, t) - \frac{\partial V(S_i, t)}{\partial S}(S_i - S_{i-1})$$

$$+ \frac{1}{2} \frac{\partial^2 V(S_i, t)}{\partial S^2}(S_i - S_{i-1})^2 + O\left((S_i - S_{i-1})^3\right)$$

$$V(S_{i+1}, t) = V(S_i, t) + \frac{\partial V(S_i, t)}{\partial S}(S_{i+1} - S_i)$$

$$+ \frac{1}{2}\frac{\partial^2 V(S_i, t)}{\partial S^2}(S_{i+1} - S_i)^2 + O\left((S_{i+1} - S_i)^3\right) .$$

We need this approximation for the partial derivatives to be exact up to second order. We thus neglect all terms of order $O((\delta S_{max})^3)$ where δS_{max} denotes the greatest distance between two neighboring nodes in the S-grid. Unlike the uniform grid case, mere addition and subtraction of the two equations does not isolate the desired differential quotients since is possible that $(S_i - S_{i-1}) \neq (S_{i+1} - S_i)$. To overcome this inconvenience, we attempt to express the differential quotients as a linear combination of the function evaluated at the points S_i, S_{i-1} and S_{i+1}. We begin by assuming that the first derivative can be written as

$$\frac{\partial V(S_i, t)}{\partial S} = aV(S_{i-1}, t) + bV(S_i, t) + cV(S_{i+1}, t) .$$

Substituting the above Taylor series for $V(S_{i+1}, t)$ and $V(S_{i-1}, t)$ into this equation and using the linear independence of $V(S_i, t)$, $\frac{\partial V(S_i, t)}{\partial S}$ and $\frac{\partial^2 V(S_i, t)}{\partial S^2}$ to compare their coefficients leads to the following system of equations for the unknown coefficients a, b and c

$$a + b + c = 0$$

$$-(S_i - S_{i-1})a + (S_{i+1} - S_i)c = 1$$

$$(S_i - S_{i-1})^2 a + (S_{i+1} - S_i)^2 c = 0$$

which has the solution

$$a = -\frac{S_{i+1} - S_i}{(S_i - S_{i-1})(S_{i+1} - S_{i-1})}$$

$$b = \left(\frac{S_{i+1} - S_i}{S_i - S_{i-1}} - \frac{S_i - S_{i-1}}{S_{i+1} - S_i}\right)\frac{1}{S_{i+1} - S_{i-1}}$$

$$c = +\frac{S_i - S_{i-1}}{(S_{i+1} - S_i)(S_{i+1} - S_{i-1})}$$

The desired approximation of the first partial derivative with respect to S is thus

$$\frac{\partial V(S_i, t)}{\partial S} = \frac{1}{S_{i+1} - S_{i-1}} \left[-\frac{S_{i+1} - S_i}{S_i - S_{i-1}} V(S_{i-1}, t) \right. \tag{10.8}$$
$$+ \left(\frac{S_{i+1} - S_i}{S_i - S_{i-1}} - \frac{S_i - S_{i-1}}{S_{i+1} - S_i} \right) V(S_i, t) + \left. \frac{S_i - S_{i-1}}{S_{i+1} - S_i} V(S_{i+1}, t) \right]$$
$$+ O\left((\delta S_{\max})^2 \right)$$

For the special case of a uniform S-grid with $S_{i+1} - S_i = S_i - S_{i-1} = \delta S$, this reduces to the expression in Eq. 10.7.

The same approach can be taken to isolate the second derivative. This time, we assume that the second derivative can be represented as a linear combination of the value function evaluated at the points S_{i-1}, S_i and S_{i+1}:

$$\frac{\partial^2 V(S_i, t)}{\partial S^2} = a V(S_{i-1}, t) + b V(S_i, t) + c V(S_{i+1}, t) \,.$$

As was done above, the Taylor expansions for $V(S_{i+1}, t)$ and $V(S_{i-1}, t)$ are substituted into this equation. Comparing the coefficients of $V(S_i, t)$, $\frac{\partial V(S_i, t)}{\partial S}$ and $\frac{\partial^2 V(S_i, t)}{\partial S^2}$ again leads to a system of equations for the unknown coefficients a, b and c, now given by

$$a + b + c = 0$$
$$-(S_i - S_{i-1})a + (S_{i+1} - S_i)c = 0$$
$$(S_i - S_{i-1})^2 a + (S_{i+1} - S_i)^2 c = 2$$

which has the solution

$$a = \frac{2}{(S_i - S_{i-1})(S_{i+1} - S_{i-1})}$$
$$b = -\frac{2}{S_{i+1} - S_{i-1}} \left(\frac{1}{S_i - S_{i-1}} + \frac{1}{S_{i+1} - S_i} \right)$$
$$c = \frac{2}{(S_{i+1} - S_i)(S_{i+1} - S_{i-1})} \,.$$

The approximation of the second derivative of V with respect to S is now given by

$$
\frac{\partial^2 V(S_i, t)}{\partial S^2} = \frac{2}{S_{i+1} - S_{i-1}} \Big[\frac{V(S_{i-1}, t)}{S_i - S_{i-1}} \tag{10.9}
$$

$$
- \left(\frac{1}{S_i - S_{i-1}} + \frac{1}{S_{i+1} - S_i} \right) V(S_i, t) + \frac{V(S_{i+1}, t)}{S_{i+1} - S_i} \Big] + O(\delta S_{\max}) \, .
$$

For a uniform S-grid this reduces to the expression in Eq. 10.7.

10.2 Difference Schemes

The approximations for the derivatives with respect to S can now be substituted into Eq. 10.3. For the sake of simplicity, we again consider first the case of a uniform grid as in Eq. 10.2, i.e., we will use the approximation given by Eq. 10.7. Under this assumption $S_i = S_0 + i\,\delta S$ holds and the approximation 10.3 for the differential quotient with respect to t evaluated at the point $t = t_j$ takes the form

$$
\begin{aligned}
\frac{\partial V(S_i, t_j)}{\partial t} &\approx r_j W_{i,j} - (r_j - q_j) S_i \frac{W_{i+1,j} - W_{i-1,j}}{2\delta S} \\
&\quad - \frac{1}{2} \sigma_{i,j}^2 S_i^2 \frac{W_{i+1,j} - 2 W_{i,j} + W_{i-1,j}}{\delta S^2} \\
&= A_{i,j} W_{i-1,j} + B_{i,j} W_{i,j} + C_{i,j} W_{i+1,j}
\end{aligned}
$$

where

$$
\begin{aligned}
A_{i,j} &= \frac{S_i}{2\delta S}(r_j - q_j - \frac{S_i}{\delta S}\sigma_{i,j}^2) \\
B_{i,j} &= r_j + \left(\frac{S_i}{\delta S} \right)^2 \sigma_{i,j}^2 \tag{10.10} \\
C_{i,j} &= -\frac{S_i}{2\delta S}(r_j - q_j + \frac{S_i}{\delta S}\sigma_{i,j}^2) \, .
\end{aligned}
$$

On the right-hand side of the equation, we have again denoted the approximation of the exact solution $V(S_i, t_j)$ evaluated at the grid points by $W_{i,j} = W(S_i, t_j)$. At this stage, we allow the interest rate and the volatility to depend on time. This is the only reason why the coefficients A, B and C have been equipped with the index j. For time-independent interest rates and volatilities, the index j on the coefficients is superfluous.

The corresponding expression for the *non*-uniform grid has exactly the same structure with somewhat more complicated coefficients A, B and C. Replacing the partial derivatives with respect to S with their approximations 10.8 and 10.9 for the non-uniform grid in 10.3 yields the following expression for the differential quotient with respect to t evaluated at $t = t_j$

$$
\begin{aligned}
\frac{\partial V(S_i, t_j)}{\partial t} &\approx r_j \, W_{i,j} - (r_j - q_j) \, S_i \{ \frac{1}{S_{i+1} - S_{i-1}} [- \frac{S_{i+1} - S_i}{S_i - S_{i-1}} W_{i-1,j} \\
&\quad + \left(\frac{S_{i+1} - S_i}{S_i - S_{i-1}} - \frac{S_i - S_{i-1}}{S_{i+1} - S_i} \right) W_{i,j} + \frac{S_i - S_{i-1}}{S_{i+1} - S_i} W_{i+1,j}] \} \\
&\quad - \frac{1}{2} \sigma_{i,j}^2 \, S_i^2 \{ \frac{2}{S_{i+1} - S_{i-1}} [\frac{W_{i-1,j}}{S_i - S_{i-1}} \\
&\quad - \left(\frac{1}{S_i - S_{i-1}} + \frac{1}{S_{i+1} - S_i} \right) W_{i,j} + \frac{W_{i+1,j}}{S_{i+1} - S_i}] \} \\
&= \frac{S_i}{(S_{i+1} - S_{i-1})(S_i - S_{i-1})} \left\{ (r_j - q_j)(S_{i+1} - S_i) - \sigma_{i,j}^2 \, S_i \right\} W_{i-1,j} \\
&\quad + \{ r_j - \frac{(r_j - q_j) \, S_i}{S_{i+1} - S_{i-1}} \left(\frac{S_{i+1} - S_i}{S_i - S_{i-1}} - \frac{S_i - S_{i-1}}{S_{i+1} - S_i} \right) \\
&\quad + \frac{\sigma_{i,j}^2 \, S_i^2}{S_{i+1} - S_{i-1}} \left(\frac{1}{S_i - S_{i-1}} + \frac{1}{S_{i+1} - S_i} \right) \} \, W_{i,j} \\
&\quad - \frac{S_i}{(S_{i+1} - S_{i-1})(S_{i+1} - S_i)} \left\{ (r_j - q_j)(S_i - S_{i-1}) + \sigma_{i,j}^2 \, S_i \right\} W_{i+1,j} \, .
\end{aligned}
$$

Thus, as was the case for the uniform grid

$$
\frac{\partial V(S_i, t_j)}{\partial t} \approx A_{i,j} \, W_{i-1,j} + B_{i,j} \, W_{i,j} + C_{i,j} \, W_{i+1,j} \, , \tag{10.11}
$$

with the slightly more complicated coefficients

$$A_{i,j} = \frac{S_i}{(S_{i+1} - S_{i-1})(S_i - S_{i-1})} \left\{ (r_j - q_j)(S_{i+1} - S_i) - \sigma_{i,j}^2 \, S_i \right\}$$
(10.12)

$$B_{i,j} = r_j + \frac{S_i}{(S_{i+1} - S_{i-1})} \{ -(r_j - q_j) \left(\frac{S_{i+1} - S_i}{S_i - S_{i-1}} - \frac{S_i - S_{i-1}}{S_{i+1} - S_i} \right)$$

$$+ \, \sigma_{i,j}^2 \, S_i \left(\frac{1}{S_i - S_{i-1}} + \frac{1}{S_{i+1} - S_i} \right) \}$$

$$C_{i,j} = -\frac{S_i}{(S_{i+1} - S_{i-1})(S_{i+1} - S_i)} \left\{ (r_j - q_j)(S_i - S_{i-1}) + \sigma_{i,j}^2 \, S_i \right\} \ .$$

For non-uniform grids as well, the time dependence of the coefficients A, B and C is solely a consequence of the time dependence of r and σ. For time-independent interest rates and volatilities, the coefficients need not have an index j. For uniform grids (at least in the S-direction), in other words, for $S_i = S_0 + i \, \delta S$, the above expression for the coefficients reduces to 10.10.

Substituting this approximation 10.11 for the differential quotient with respect to t in the general equation 10.6 finally gives the generalized form of the finite difference scheme for non-uniform grids:

$$\theta \, A_{i,j} \, W_{i-1,j} + (\theta \, B_{i,j} + \frac{1}{t_{j+1} - t_j}) \, W_{i,j} + \theta C_{i,j} \, W_{i+1,j}$$
(10.13)

$$\approx \frac{W_{i,j+1}}{(t_{j+1} - t_j)} - (1 - \theta) \left[A_{i,j+1} \, W_{i-1,j+1} + B_{i,j+1} \, W_{i,j+1} + C_{i,j+1} \, W_{i+1,j+1} \right] \ .$$

As in 10.6, θ can take on arbitrary values between 0 and 1 giving the *implicit* method for $\theta = 1$, the *explicit* method for $\theta = 0$ and the *Crank-Nicolson* method for $\theta = 1/2$. In order to illustrate the importance of this difference equation, note that the values to be calculated on the left-hand side of the equation consist of terms evaluated at time point t_j whereas those on the right-hand side concern only values at time point t_{j+1}, and as such are known, having been calculated in the previous step. Thus, for every (inner) underlying grid point i there is one equation connecting three option values at t_{j+1} with three values at t_j.

The range of the time index j here is always $j = 0, \dots N - 1$ since $j = N$ is already given by the initial condition, i.e., the payoff profile at maturity $T = t_N$. The range of the S-index i (see Eq. 10.1) is at least $i = 1, \dots, M - 1$,

but can take on the values including $i = 0$ and/or $i = M$ (depending on the boundary conditions to be satisfied, see below). The difference equation for $i = 0$ and $i = M$ seems problematic at first glance since values of W are required at points not defined in the grid, for example "$W_{0-1,j}$" or "$W_{M+1,j}$". As will be shown below, these problems can in fact be overcome by consideration of the boundary conditions themselves.

The difference scheme can be written in matrix form (here, for the case where the range of the index i is given by $i = 1, \ldots, M - 1$; other cases can be expressed analogously):

$$
\underbrace{
\begin{pmatrix}
a_1 & b_1 & c_1 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & a_2 & b_2 & c_2 & 0 & & & \vdots \\
\vdots & 0 & a_3 & b_3 & c_3 & 0 & & \vdots \\
\vdots & & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & & \ddots & \ddots & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & \cdots & 0 & a_{M-1} & b_{M-1} & c_{M-1}
\end{pmatrix}
}_{M+1 \text{ columns and } M-1 \text{ rows}}
\underbrace{
\begin{pmatrix}
W_{0,j} \\
W_{1,j} \\
W_{2,j} \\
\vdots \\
\vdots \\
\vdots \\
W_{M,j}
\end{pmatrix}
}_{M+1 \text{ rows}}
=
\underbrace{
\begin{pmatrix}
D_{1,j} \\
D_{2,j} \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
D_{M-1,j}
\end{pmatrix}
}_{M-1 \text{ rows}}
$$

$$(10.14)$$

where

$$
a_i = \theta A_{i,j} , \quad b_i = \theta B_{i,j} + \frac{1}{t_{j+1} - t_j} , \quad c_i = \theta C_{i,j}
$$

$$
D_{i,j} = \frac{W_{i,j+1}}{(t_{j+1} - t_j)} - (1 - \theta) \left[A_{i,\,j+1} \, W_{i-1,j+1} + B_{i,j+1} \, W_{i,j+1} + C_{i,j+1} \, W_{i+1,j+1} \right] .
$$

The $D_{i,j}$ on the right-hand side depend only on the values at time t_{j+1} which have already been calculated in the previous step and are hence completely determined. For each time step j, we have $(M - 1)$ equations for $(M + 1)$ unknown $W_{i,j}$. Other ranges for the index i give the same result: the system of equations is underdetermined; there are two equations fewer than there are unknowns. The two additional equations needed to solve the above system are provided by two boundary conditions.

10.2.1 Initial Conditions

Besides the specification of boundary conditions with respect to the S variable, we also need a boundary condition with respect to the t variable (so called *initial condition*) to be satisfied in order to obtain a unique solution to the above defined difference equation. The initial condition is specified by the *payoff profile $P(S)$* of the derivative concerned, i.e., by the value $V(S, t_N = T)$ given by

$$V(S, T) = P(S)$$

or in the discrete "grid notation"

$$W_{i,N} = P_i .$$

For example, for the payoff profile $P(S)$ for a European call option with strike price K, the initial condition is given by

$$V(S, T) = P(S) = \max(S - K, 0) \implies W_{i,N} = \max(S_i - K, 0) .$$

10.2.2 Dirichlet Boundary Conditions

If either the terms of the option contract (for example, barrier options) or some other information allow us to specify directly the *value* of the option for certain values of S, these values can be used as boundary conditions for the S-grid. Such boundary conditions where the option value itself is given at the boundary are called *Dirichlet boundary conditions*. Let R^U denote these given option values at the *upper boundary S_M*, and R^L denote the given option values at the *lower boundary S_0*, i.e.:

$$V(S_M, t) = R^U(t), \quad V(S_0, t) = R^L(t)$$

or in the discrete "grid notation"

$$W_{M,j} = R_j^U , \quad W_{0,j} = R_j^L .$$

It is often the case that only approximations for R_j^U and R_j^L are known. If this is the case, the difference scheme provides an approximation which is at best as good as this approximation for the boundary conditions. If, for example,

the Crank-Nicolson scheme is applied but the boundary condition can only be approximated in the first order of S, then the solution procedure as a whole is exact up to first order even though the Crank-Nicolson scheme provides an approximation which is exact up to second order.

Two of the option values to be calculated for each time j are thus specified directly if Dirichlet boundary conditions are given. The dimension of the problem (in the sense of dimension equals number of unknowns) is therefore only $(M - 1)$. This is exactly the number of equations in 10.14. Our goal now is to transform Eq. 10.14 into a system consisting of a square matrix, i.e., to reduce the dimension "in each direction" to $(M - 1)$. To keep the boundary conditions separate, we write the vector of option values in the form

$$
\begin{pmatrix}
W_{0,j} \\
W_{1,j} \\
W_{2,j} \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
W_{M,j}
\end{pmatrix}
=
\begin{pmatrix}
0 \\
W_{1,j} \\
W_{2,j} \\
\vdots \\
\vdots \\
\vdots \\
W_{M-1,j} \\
0
\end{pmatrix}
+
\begin{pmatrix}
W_{0,j} \\
0 \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
0 \\
W_{M,j}
\end{pmatrix}.
$$

In the first vector, only the (yet to be determined) option values away from the boundary of the grid appear. We call this the *unknown vector*. The second vector contains only the (already specified) values of the option on the boundary. We refer to this vector as the *known vector*. The matrix in Eq. 10.14 acts on both of these vectors. Let us first consider the unknown vector: the only element in the first column of the matrix, namely a_1, acts only on the first row of the unknown vector. This, however, equals zero. We obtain from the matrix multiplication of the unknown vector the same result as if the first column in the matrix and the first row in the vector were removed. The situation is the same for the last column of the matrix and the last row of the unknown vector. Thus we can simply remove the last column of the matrix and the last row of the unknown vector without changing the result of the matrix multiplication.

Matrix multiplication of the full matrix in Eq. 10.14 with the *known* vector is explicitly performed. Combining everything the system of equations in Eq. 10.14 can be equivalently written using an $(M-1) \times (M-1)$ matrix:

$$
\underbrace{\begin{pmatrix} b_1 & c_1 & 0 & \cdots & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & & \cdots \\ 0 & a_3 & b_3 & c_3 & \ddots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & c_{M-2} \\ 0 & \cdots\cdots & 0 & a_{M-1} & b_{M-1} \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} W_{1,j} \\ W_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ W_{M-1,j} \end{pmatrix}}_{\mathbf{W}_j} + \begin{pmatrix} a_1\, W_{0,j} \\ 0 \\ \vdots \\ \vdots \\ 0 \\ c_{M-1}\, W_{M,j} \end{pmatrix} = \begin{pmatrix} D_{1,j} \\ D_{2,j} \\ \vdots \\ \vdots \\ D_{M-2,j} \\ D_{M-1,j} \end{pmatrix}
$$

$$
\Leftrightarrow \underbrace{\begin{pmatrix} b_1 & c_1 & 0 & \cdots & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & & \cdots \\ 0 & a_3 & b_3 & c_3 & \ddots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & c_{M-2} \\ 0 & \cdots\cdots & 0 & a_{M-1} & b_{M-1} \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} W_{1,j} \\ W_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ W_{M-1,j} \end{pmatrix}}_{\mathbf{W}_j} = \begin{pmatrix} d_{1,j} \\ d_{2,j} \\ \vdots \\ \vdots \\ d_{M-2,j} \\ d_{M-1,j} \end{pmatrix}
$$

(10.15)

with coeffitions

$$
a_i = \theta A_{i,j}\,, \quad b_i = \theta B_{i,j} + \frac{1}{t_{j+1} - t_j}\,, \quad c_i = \theta C_{i,j} \quad \text{for } i = 1, \ldots, M-1
$$

$$
d_{i,j} = \frac{W_{i,j+1}}{(t_{j+1} - t_j)} \qquad \text{for } i = 2, \ldots, M-2
$$
$$
\quad - (1-\theta)\left[A_{i,\,j+1}\, W_{i-1,j+1} + B_{i,j+1}\, W_{i,j+1} + C_{i,j+1}\, W_{i+1,j+1} \right]
$$

$$
d_{1,j} = \frac{W_{1,j+1}}{(t_{j+1} - t_j)} - (1-\theta)\left[B_{1,\,j+1}\, W_{1,j+1} + C_{1,j+1}\, W_{2,j+1} \right] \qquad (10.16)
$$
$$
\quad - \left[\theta A_{1,j}\, R_j^L + (1-\theta) A_{1,j+1}\, R_{j+1}^L \right]
$$

$$
d_{M-1,j} = \frac{W_{M-1,j+1}}{(t_{j+1} - t_j)} - (1-\theta)\left[A_{M-1,j+1}\, W_{M-2,j+1} + B_{M-1,j+1}\, W_{M-1,j+1} \right]
$$
$$
\quad - \left[\theta C_{M-1,j}\, R_j^U + (1-\theta) C_{M-1,j+1}\, R_{j+1}^U \right]\,.
$$

Along the way, the result of the matrix multiplication with the known vector (i.e., with the boundary condition) has been brought over to the right-hand side. This right-hand side (the $d_{i,j}$) only depends on the values at the time points t_{j+1} (which have already been calculated in the previous step) and on the (given) boundary conditions. The $d_{i,j}$ are thus completely determined. The only unknowns in this system are the $(M-1)$ elements of the vector W_j. These can now be calculated *if* it is possible to invert the matrix $\mathbf{A}$. Because of the special form of the matrix (the only non-zero elements are in the diagonal and the two off-diagonals), a calculation-intensive matrix inversion can be avoided. Instead, we employ a very fast procedure, known as the *L-U decomposition*. To this end, we decompose the matrix $\mathbf{A}$ into the product of a matrix $\mathbf{L}$, which contains non-zero elements only in the diagonal and the "lower" off-diagonal and a matrix $\mathbf{U}$ whose only non-zero elements are found in the diagonal and the "upper" off-diagonal:

$$
\mathbf{A} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & \cdots & & \cdots & 0 \\ l_2 & 1 & 0 & & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \vdots \\ \vdots & & \ddots & l_{M-2} & 1 & 0 \\ 0 & \cdots & \cdots & 0 & & l_{M-1} & 1 \end{pmatrix}}_{\mathbf{L}} \underbrace{\begin{pmatrix} h_1 & u_1 & 0 & \cdots & \cdots & & 0 \\ 0 & h_2 & u_2 & 0 & \cdots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & & 0 & \ddots & u_{M-2} \\ 0 & \cdots & \cdots & \cdots & & 0 & h_{M-1} \end{pmatrix}}_{\mathbf{U}}.
$$

From this *ansatz* the elements of the matrices $\mathbf{L}$ and $\mathbf{U}$ are determined to be

$$
\begin{aligned}
u_i &= c_i && \text{for } i = 1, \ldots M - 2 \\
h_1 &= b_1 \\
l_i &= \tfrac{a_i}{h_{i-1}} && \text{for } i = 2, \ldots M - 1 \\
h_i &= b_i - l_i u_{i-1} && \text{for } i = 2, \ldots M - 1 .
\end{aligned}
$$

Calculating the elements in the order indicated above, we can easily compute all the elements appearing in both matrices. Inverting $\mathbf{L}$ and $\mathbf{U}$ is quite simple. We begin by writing

$$
\mathbf{A}\, W_j = \mathbf{L}\, \underbrace{\mathbf{U}\, W_j}_{\mathbf{x}} =: \mathbf{L}\mathbf{x}
$$

and solve the system for our newly defined vector **x**:

$$
\begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & 0 \\
l_2 & 1 & 0 & & & \vdots \\
0 & \ddots & \ddots & \ddots & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & l_{M-2} & 1 & 0 \\
0 & \cdots & \cdots & 0 & l_{M-1} & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_{M-1}
\end{pmatrix}
=
\begin{pmatrix}
d_{1,j} \\ d_{2,j} \\ d_{3,j} \\ \vdots \\ \vdots \\ d_{M-2,j} \\ d_{M-1,j}
\end{pmatrix}.
$$

To solve this system for **x**, we start with the first row of the matrix which contains only one non-zero element and proceed from top to bottom to obtain

$$
x_1 = d_{1,j}
$$
$$
x_i = d_{i,j} - l_i\, x_{i-1} \quad for\, i = 2 \ldots M-1 \,.
$$

Now, with vector **x** is known, the option values $W_{i,j}$ can simply be calculated from the above definition of **x** by solving the equation $\mathbf{U}\; W_j = \mathbf{x}$:

$$
\begin{pmatrix}
h_1 & u_1 & 0 & \cdots & \cdots & 0 \\
0 & h_2 & u_2 & 0 & \cdots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & \ddots & \ddots & 0 \\
\vdots & & & 0 & \ddots & u_{M-2} \\
0 & \cdots & \cdots & \cdots & 0 & h_{M-1}
\end{pmatrix}
\begin{pmatrix}
W_{1,j} \\ W_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ W_{M-1,j}
\end{pmatrix}
=
\begin{pmatrix}
x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_{M-1}
\end{pmatrix}.
$$

To "invert" the matrix **U**, we begin this time with the last row of the matrix and work our way up to obtain

$$
W_{M-1,j} = \frac{x_{M-1}}{h_{M-1}} \tag{10.17}
$$
$$
W_{i,j} = \frac{x_i - u_i\, W_{i+1,j}}{h_i} \quad \text{for} \quad i = 1 \ldots M-2
$$

Thus, we have found an explicit solution for the general finite difference scheme 10.13, expressing the option values at time t_j in terms of the values at a later time point t_{j+1} under consideration of the given values at the boundaries.

We will now consider the plain vanilla put and call as well as the barrier option as examples for the application of this scheme with Dirichlet boundary conditions. While, in principle, we must find the values of the plain vanilla options on the entire set from $S = 0$ to $S = \infty$, the value of the barrier option is specified directly in the option contract on one side of the barrier (and on the barrier itself). The area to be covered by its grid is thus smaller than that for plain vanilla options.

Let us, however, begin by considering plain vanilla options. For these options, the boundary conditions must be specified at "$S = 0$" and "$S = \infty$". We choose a lower bound S_0 small enough, such that the value of the *call* for such values of the underlying price is negligible (the option is well out of the money). Furthermore, we choose an upper bound S_M large enough, such that the value of a *put* is negligible for such values of the underlying price (the option is well out of the money). The put-call parity is then used to establish the other boundary values for each of the two options. Recall that for European options on an underlying paying a dividend yield q the put-call parity, Eq. 6.8, is given by:

$$\text{Price(Call)} - \text{Price(Put)} = \text{Price(Forward)} = Se^{-q(T-t)} - Ke^{-r(T-t)} \ .$$

The right-hand side is the value of a forward contract[3] with strike price K. Since the put at the upper boundary is worthless, the call option has the same value as the forward contract (the option is so far in the money that it will be exercised with certainty). Conversely, the call at the lower boundary is worthless, therefore the put has the same value as a short forward contract (the option is so far in the money that it will be exercised with certainty). Summarizing, we have the following Dirichlet boundary conditions for the plain vanilla options:

	$S_0 \approx 0$	$S_M \approx \infty$
call	$R_j^L = 0$	$R_j^U = S_M e^{-q(T-t_j)} - Ke^{-r(T-t_j)}$
put	$R_j^L = -S_0 e^{-q(T-t_j)} + Ke^{-r(T-t_j)}$	$R_j^U = 0$

---

[3]For the sake of simplicity, the price of the forward contract is given for the case of a flat interest rate term structure, a flat dividend yield curve and no discrete dividend payments. However, this relation also holds for interest rates and dividend yields which are time-dependent.

We now take an up-and-out barrier call option as an example of a somewhat exotic and path dependent option and specify its boundary condition. The initial condition and the boundary condition for $S = 0$ is exactly the same as that of a plain vanilla call but as soon as S attains the upper barrier level H, the option becomes worthless. Often, the option's holder still receives a payment if the barrier is reached, the so-called *rebate*. Let us assume that the rebate R is due at precisely the knock-out time point (i.e., the instant when S touches or breaches the barrier). Then, the option's value is given by the value of the rebate, namely R. It is thus convenient to select the upper boundary of the grid as $S_M = H$. The boundary condition for such a barrier option is then simply $W_{M,j} = R$.

10.2.3 Neumann Boundary Condition

It often occurs that the first derivative of the option price S is specified at the boundary of the grid rather than the option price itself,

$$\left.\frac{\partial V(S,t)}{\partial S}\right|_{S=S_M} = R^U(t), \quad \left.\frac{\partial V(S,t)}{\partial S}\right|_{S=S_0} = R^L(t) .$$

Boundary conditions of this type are called *Neumann boundary conditions.*[4] Since these boundary conditions do not directly specify the *values* of the solution (the option values) on the boundary, all $(M + 1)$ values must be calculated in each time slice. The dimension (the number of values to be calculated) of the problem is thus $(M + 1)$. In order to obtain the same number of equations, the index i in Eq. 10.13 must range from 0 to M. As a result, two "additional unknowns", namely "$W_{-1,j}$" and "$W_{M+1,j}$" appear in the system of equations. The two Neumann boundary conditions will be used to eliminate these additional unknowns. In comparison to the Dirichlet conditions, we will then have system of equations consisting of two more equations but exactly as many equations as unknowns.

Since we use three point finite differences in S-direction (see for example Eq. 10.8) the grid must be extended at both boundaries by one virtual grid point in S-direction for each index j. In other words, we add two additional grid points S_{-1} and S_{M+1} at each time point t_j. This means that the index i in

---

[4]As already pointed out in the section on Dirichlet boundary conditions, the solution is at best as exact as the given boundary conditions. This is important in all cases where there are only approximations to the boundary conditions $R^U(t)$ and $R^L(t)$ available.

the difference equation 10.13 takes on the values $i = 0, \ldots, M$. In the finite difference approximation 10.8 of the first derivative with respect to S, we can substitute the respective Neumann conditions for the cases $i = 0$ and $i = M$ on the left-hand side of the equation and then rearrange the terms so that the option values outside the grid are expressed in terms of those defined within the grid and the boundary conditions. This yields

$$
W_{M+1,j} \approx \left(\frac{S_{M+1} - S_M}{S_M - S_{M-1}} \right)^2 W_{M-1,j} + \left[1 - \left(\frac{S_{M+1} - S_M}{S_M - S_{M-1}} \right)^2 \right] W_{M,j}
$$
$$
+ \frac{S_{M+1} - S_M}{S_M - S_{M-1}} (S_{M+1} - S_{M-1}) R_j^U
$$

at the upper boundary, i.e., for $i = M$. Likewise for $i = 0$ on the lower boundary we have

$$
W_{-1,j} \approx \left[1 - \left(\frac{S_0 - S_{-1}}{S_1 - S_0} \right)^2 \right] W_{0,j} + \left(\frac{S_0 - S_{-1}}{S_1 - S_0} \right)^2 W_{1,j}
$$
$$
- \frac{S_0 - S_{-1}}{S_1 - S_0} (S_1 - S_{-1}) R_j^L \; .
$$

Substituting these expressions into Eq. 10.11, we obtain the approximation for the differential quotient with respect to *time* at the boundary of the S-grid. For the upper boundary, this gives:

$$
\frac{\partial V(S_M, t_j)}{\partial t} \approx A_{M,j} \, W_{M-1,\, j} + B_{M,j} \, W_{M,j}
$$
$$
+ C_{M,j} \left\{ \left(\frac{S_{M+1} - S_M}{S_M - S_{M-1}} \right)^2 W_{M-1,j} + \left[1 - \left(\frac{S_{M+1} - S_M}{S_M - S_{M-1}} \right)^2 \right] W_{M,j} \right.
$$
$$
\left. + \frac{S_{M+1} - S_M}{S_M - S_{M-1}} (S_{M+1} - S_{M-1}) R_j^U \right\}
$$
$$
= A_{M,j}^* \, W_{M-1,j} + B_{M,j}^* \, W_{M,j}
$$
$$
+ \frac{S_{M+1} - S_M}{S_M - S_{M-1}} (S_{M+1} - S_{M-1}) R_j^U \, C_{M,j}
$$

where

$$A^*_{M,j} = A_{M,j} + \left(\frac{S_{M+1} - S_M}{S_M - S_{M-1}}\right)^2 C_{M,j}$$

$$B^*_{M,j} = B_{M,j} + \left[1 - \left(\frac{S_{M+1} - S_M}{S_M - S_{M-1}}\right)^2\right] C_{M,j} .$$

Likewise for the lower boundary:

$$\frac{\partial V(S_0, t_j)}{\partial t} \approx A_{0,j} \left\{\left[1 - \left(\frac{S_0 - S_{-1}}{S_1 - S_0}\right)^2\right] W_{0,\,j} + \left(\frac{S_0 - S_{-1}}{S_1 - S_0}\right)^2 W_{1,j}\right.$$

$$\left. - \frac{S_0 - S_{-1}}{S_1 - S_0}(S_1 - S_{-1})R^L_j\right\} + B_{0,j}\, W_{0,\,j} + C_{0,j}\, W_{1,j}$$

$$= B^*_{0,j}\, W_{0,j} + C^*_{0,j}\, W_{1,j} - \frac{S_0 - S_{-1}}{S_1 - S_0}(S_1 - S_{-1})R^L_j A_{0,j}$$

where

$$B^*_{0,j} = B_{0,j} + \left[1 - \left(\frac{S_0 - S_{-1}}{S_1 - S_0}\right)^2\right] A_{0,j}$$

$$C^*_{0,j} = C_{0,j} + \left(\frac{S_0 - S_{-1}}{S_1 - S_0}\right)^2 A_{0,j} .$$

All preparations have now been made for specifying the difference scheme 10.13 at the boundaries. Replacing the differential quotient with respect to *time* in Eq. 10.6 with the above approximation gives

$$\theta A^*_{M,j}\, W_{M-1,j} + (\theta B^*_{M,j} + \frac{1}{t_{j+1} - t_j})\, W_{M,j}$$

$$\approx \frac{W_{M,j+1}}{(t_{j+1} - t_j)} - (1 - \theta)\left[A^*_{M,j+1}\, W_{M-1,j+1} + B^*_{M,j+1}\, W_{M,j+1}\right]$$

$$- \frac{S_{M+1} - S_M}{S_M - S_{M-1}}(S_{M+1} - S_{M-1})\left[\theta R^U_j\, C_{M,j} + (1 - \theta)R^U_{j+1}\, C_{M,j+1}\right]$$

for the upper boundary. Likewise, the difference scheme for the lower boundary is obtained as

$$(\theta B_{0,j}^* + \frac{1}{t_{j+1} - t_j}) W_{0,j} + \theta C_{0,j}^* W_{1,j}$$

$$\approx \frac{W_{0,j+1}}{(t_{j+1} - t_j)} - (1 - \theta) \left[B_{0,j+1}^* W_{0,j+1} + C_{0,j+1}^* W_{1,j+1} \right]$$

$$+ \frac{S_0 - S_{-1}}{S_1 - S_0} (S_1 - S_{-1}) \left[\theta R_j^L A_{0,j} + (1 - \theta) R_{j+1}^L A_{0,j+1} \right] .$$

Combining the above results, the finite difference scheme 10.13 for Neumann boundary conditions can be written in the following matrix form:

$$\underbrace{\begin{pmatrix} b_0 & c_0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ a_1 & b_1 & c_1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & a_2 & b_2 & c_2 & 0 & 0 & \ldots & 0 \\ 0 & 0 & a_3 & b_3 & c_3 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ddots & \ddots & \ddots & & 0 \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & & 0 & a_{M-1} & b_{M-1} & c_{M-1} \\ 0 & \ldots & \ldots & \ldots & & 0 & a_M & b_M \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} W_{0,j} \\ W_{1,j} \\ W_{2,j} \\ \\ \\ \\ \\ W_{M,j} \end{pmatrix}}_{\mathbf{W}_j} = \underbrace{\begin{pmatrix} d_{0,j} \\ d_{1,j} \\ d_{2,j} \\ \\ \\ \\ \\ d_{M,j} \end{pmatrix}}_{\mathbf{d}_j} \quad (10.18)$$

where

$$a_i = \theta A_{i,j} , \quad b_i = \theta B_{i,j} + \frac{1}{t_{j+1} - t_j} , \quad c_i = \theta C_{i,j} \text{ for } i = 1, \ldots, M - 1$$

$$d_{i,j} = \frac{W_{i,j+1}}{(t_{j+1} - t_j)} \qquad\qquad \text{for } i = 1, \ldots, M - 1$$

$$- (1 - \theta) \left[A_{i,\, j+1} W_{i-1,j+1} + B_{i,j+1} W_{i,j+1} + C_{i,j+1} W_{i+1,j+1} \right]$$

$$a_M = \theta A_{M,j}^* , \quad b_M = \theta B_{M,j}^* + \frac{1}{t_{j+1} - t_j} ,$$

$$d_{M,j} = \frac{W_{M,j+1}}{(t_{j+1} - t_j)} - (1 - \theta) \left[A_{M,\, j+1}^* W_{M-1,j+1} + B_{M,j+1}^* W_{M,j+1} \right]$$

$$- \frac{S_{M+1} - S_M}{S_M - S_{M-1}} (S_{M+1} - S_{M-1}) \left[\theta R_j^U C_{M,j} + (1 - \theta) R_{j+1}^U C_{M,j+1} \right]$$

$$b_0 = \theta B_{0,j}^* + \frac{1}{t_{j+1} - t_j} \ , \quad c_0 = \theta C_{0,j}^*$$

$$d_{0,j} = \frac{W_{0,j+1}}{(t_{j+1} - t_j)} - (1-\theta) \left[B_{0,\ j+1}^* \, W_{0,j+1} + C_{0,j+1}^* \, W_{1,j+1} \right]$$

$$+ \frac{S_0 - S_{-1}}{S_1 - S_0}(S_1 - S_{-1}) \left[\theta R_j^L A_{0,j} + (1-\theta) R_{j+1}^L A_{0,j+1} \right] \ .$$

$$(10.19)$$

This system of equations must be solved for each time step taking the boundary conditions into consideration. The right-hand side of Eq. 10.18 consists only of values evaluated at time point t_{j+1} (which have already been computed in the previous step) and of the (given) boundary conditions. The $d_{i,j}$ are thus completely determined. The system of equations has the same structure as in the corresponding system 10.15 for the Dirichlet problem except that it is two dimensions larger. Here, $\mathbf{A}$ is a $(M+1) \times (M+1)$ matrix. This is explained by the fact that the grid has been extended at the upper and lower boundary in the S-direction. The simple structure of the matrix $\mathbf{A}$ again allows to perform a L-U decomposition. As before, we begin by setting $\mathbf{A} \ W_j = \mathbf{L U} \ W_j =: \mathbf{L x}$ and, exactly as was the case for Dirichlet boundary conditions, we obtain

$$
\begin{aligned}
u_i &= c_i & &\text{for } i = 0, \ldots M-1 \\
h_0 &= b_0 \\
l_i &= \frac{a_i}{h_{i-1}} & &\text{for } i = 1, \ldots M \\
h_i &= b_i - l_i u_{i-1} & &\text{for } i = 1, \ldots M \\
x_0 &= d_{0,j} \\
x_i &= d_{i,j} - l_i x_{i-1} & &\text{for } i = 1 \ldots M \\
W_{M,j} &= \frac{x_M}{h_M} \\
W_{i,j} &= \frac{x_i - u_i W_{i+1,j}}{h_i} & &\text{for } i = 0 \ldots M-1 \ .
\end{aligned}
$$

$$(10.20)$$

The only difference is that the range of the index i has been extended and that a_i, b_i, c_i, d_i are now given by Eq. 10.19.

As an example of the Neumann problem, we consider the plain vanilla put and call. Neumann boundary conditions are usually easier to determine than Dirichlet conditions. For S_0 the call is well out of the money and

the dependence of the price on S can be assumed to be negligible. The corresponding boundary condition is thus $\frac{\partial V(S,t)}{\partial S}\Big|_{S=S_0} = 0$. For extremely large S, the call is way in the money and thus its behavior is approximated by the payoff profile. The derivative in the S direction is thus 1 with the corresponding boundary condition given by $\frac{\partial V(S,t)}{\partial S}\Big|_{S=S_0} = 1$. The behavior of the put is the exact opposite. Summarizing, the Neumann boundary conditions for the plain vanilla put and call are given by:

	$S_0 \approx 0$	$S_M \approx \infty$	$t_N = T$
call	$R_j^L = 0$	$R_j^U = 1$	$P_i = \max(S_i - K, 0)$
put	$R_j^L = -1$	$R_j^U = 0$	$P_i = \max(K - S_i, 0)$

For the sake of completeness, we have also included the initial conditions in the table.

The type of boundary condition selected for the analysis depends on the type of option being priced. For plain vanilla options, Neumann boundary conditions are attractive because they can be easily calculated. The pricing of a forward contract at each time point (which must be done to establish the Dirichlet boundary conditions), on the other hand, can be tedious, particularly when the term structure is not flat and discrete dividends must be taken into consideration.

For Knock-out barrier options, however, the exact value of the option at the barrier is known (0 or the rebate), while the change of its value at the boundary is unknown. In this case, it makes sense to use the Dirichlet boundary conditions. In addition, a combination of Dirichlet and Neumann boundary conditions can be applied to one and the same problem as needed.

10.2.4   Generalized Neumann Boundary Conditions

It is possible to generalize the Neumann boundary condition by a simple method that does not need any explicit information on boundary conditions with respect to S. Valuation of derivatives most often deals with diffusion problems, with the Black Scholes equation being its most prominent example. Diffusion problems share the property that the pay off profile is being smoothed or "smeared out" the more the farther away the expiry is in the future. At the same time, the pay off profile is a simple linear function of S for

very small or very large S. Consequently, at the boundaries, the fair value can be approximated as a linear function of S.[5]

A linear function has the property that the second derivative is zero. Therefore, instead of specifying the value of the first derivative at the boundary, which is a standard Neumann condition, we may specify the value of the second derivative, which is (almost) independent of the option type (i.e. it is equal to zero) as a *generalized* Neumann condition.

We assume a grid in dimension S with $M + 1$ nodes $S_0, S_1, S_2, \ldots, S_M$, with S_0 and S_M being virtual grid nodes. They will be derived from the constraint of zero second order derivatives at the boundaries and thus eliminated.

First consider the lower boundary $i = 1$. We calculate the unknown value $W_{0,j}$ by setting Eq. 10.9 to zero for the second derivative with respect to S at node $i = 1$:

$$\frac{2}{S_2 - S_0} \left[\frac{W_{0,j}}{S_1 - S_0} - \left(\frac{1}{S_1 - S_0} + \frac{1}{S_2 - S_1} \right) W_{1,j} + \frac{W_{2,j}}{S_2 - S_1} \right] = 0 .$$

$$(10.21)$$

The virtual grid node S_0 may be chosen freely. For simplicity, we set $S_1 - S_0 = S_2 - S_1$, i.e. $S_0 = 2S_1 - S_2$. Then, all terms in Eq. 10.21 have the same denominator. If follows:

$$W_{0,j} - 2W_{1,j} + W_{2,j} = 0$$
$$\Leftrightarrow \quad W_{0,j} = 2W_{1,j} - W_{2,j} .$$

An analogous expression can be derived for node M:

$$\Leftrightarrow \quad W_{M,j} = 2W_{M-1,j} - W_{M-2,j} .$$

Now, the problem is reduced to a system of Dirichlet type (Eq. 10.14) with known $W_{0,j}$ and $W_{M,j}$. Alternatively, it is also possible to set a_1 in 10.15 to zero and replace b_1 by $\tilde{b}_1 = b_1 + 2a_1$ and c_1 by $\tilde{c}_1 = c_1 - a_1$. With this, $W_{0,j}$

---

[5]One of the rare exceptions to this rule is a power option without cap, with a pay off profile proportional to S^2. Such options are of minor relevance in praxis.

has been eliminated. Analogous, $W_{M,j}$ can be eliminated as well. This is the same as replacing Eq. 10.15 by

$$
\begin{pmatrix}
b_1 + 2a_2 & c_1 - a_1 & 0 & \cdots & & \cdots & & 0 \\
a_2 & b_2 & c_2 & 0 & & & & \cdots \\
0 & a_3 & b_3 & c_3 & \ddots & & & \cdots \\
\vdots & \ddots & \ddots & \ddots & & \ddots & & 0 \\
\vdots & & & \ddots & \ddots & & \ddots & c_{M-2} \\
0 & \cdots & & \cdots & 0 & a_{M-1} - c_{M-1} & b_{M-1} + 2c_{M-1}
\end{pmatrix}
\begin{pmatrix}
W_{1,j} \\
W_{2,j} \\
\vdots \\
\vdots \\
\vdots \\
W_{M-1,j}
\end{pmatrix}
=
\begin{pmatrix}
D_{1,j} \\
D_{2,j} \\
\vdots \\
\vdots \\
D_{M-2,j} \\
D_{M-1,j}
\end{pmatrix}
$$

with the same coefficients as in Eq. 10.16.

10.2.5   Free Boundary Conditions for American Options

Thus far our discussion has been restricted to European options. Finite difference methods can also be extended to American options. The right to early exercise must, of course, be taken into account in pricing the option. The Black-Scholes equation holds only on a restricted set and must be replaced by the two (in)equalities given by Eqs. 7.13 and 7.14. Equivalently expressed, the Black-Scholes equation 7.13 holds on a parameter set with a free boundary given by $S^*(t)$. This line $S^*(t)$ separates the parameter set where the Black-Scholes *equation* holds from the parameter set where the value of the option is simply given by the payoff profile 7.14. Directly on this curve $S^*(t)$ we have

$$
V(S^*(t), t) = P(S^*(t)) .
$$

This is a Dirichlet boundary condition. The boundary condition is called *free* because $S^*(t)$ is not known a priori.

A simple solution for this problem can be found using the *explicit* finite difference procedure. As for the European option, we start by determining a solution vector $W_{i,j}$, $i = 0, 1, \ldots, M$ for a time step j. An intermediate calculation is performed to determine the solution vector $\widetilde{W}_{i,j}$ for the American option. In this calculation, the European price is determined and compared to the intrinsic value $P(S)$. The value of the American option is then defined as the larger of the two, i.e.,

$$
\widetilde{W}_{i,j} = \max\left[W_{i,j}, P(S_i) \right] .
$$

The procedure continues with the next time step using this solution vector.

If an implicit (or mixed) finite difference procedure is utilized, this will *not* be exact since the vector $W_{i,j}$ will already have been used implicitly in its calculation. Therefore the procedure can in principle not be split into the above two steps. It can be shown, however, that the error in doing so remains small if the grid in t is fine enough. In particular, if the distance between two time points is one day or less, the error is often negligible. An additional trick can be used for call options. It is known that the exercise of a call option can only be optimal (if at all) immediately before the payment of a discrete dividend. We can thus introduce extra time points on and immediately preceding the due dates of the dividend payments with a time difference of, for example, half a day, and thus minimize the error.

In any case, in addition to the free boundary conditions, values for the upper and lower boundaries of the S-grid must also be specified. This is already clear from the explicit procedure described above. We can obviously only compare the solution vector for a European option with the payoff profile if this vector has already been determined. To determine this solution vector for the European option, we need two boundary conditions. They can be either Neumann, Dirichlet, or more general boundary conditions (such as the second derivative equals zero). For an American option all boundary conditions involving derivatives are generally the same as for the corresponding European option. The maximum of the payoff profile and the boundary condition of the corresponding European option is a good candidate for the Dirichlet boundary condition. Thus, for American plain vanilla put and call options:

	Call
$S_0 \approx 0$	$R_j^L = 0$
$S_M \approx \infty$	$R_j^U = \max \left[S_M e^{-q(T-t_j)} - K e^{-r(T-t_j)}, S_M - K \right]$
$t_N = T$	$P_i = \max(S_i - K, 0)$

	Put
$S_0 \approx 0$	$R_j^L = \max \left[-S_0 e^{-q(T-t_j)} + K e^{-r(T-t_j)}, -S_0 + K \right]$
$S_M \approx \infty$	$R_j^U = 0$
$t_N = T$	$P_i = \max(K - S_i, 0)$

Note that for American options this is not enough. We also need to specify the conditions on the free boundary $S^*(t)$. And before we can do this, we need to determine $S^*(t)$ in the first place.

The Lamberton and Lapeyre Procedure

In many cases, the free boundary lies on only one side of the grid. For this case, a suitable procedure for evaluating the American put options has been suggested by *Lamberton* and *Lapeyre* [129]. This procedure does not necessitate any appreciable additional computational effort. It simply varies the order of operations needed to solve the system of equations for one time step. We will now present this procedure.

The Dirichlet boundary condition for the plain vanilla option as given above will be used as the specified boundary condition (other types of boundary conditions could also have been used). The difference scheme then has the form given in Eq. 10.15 with coefficients given by Eq. 10.16. Before introducing the procedure, we consider the following limiting cases: the exercise region for a put option must lie in regions where the value of S is small since for large S (for $S > K$) the intrinsic value of a put is zero; the option only has a time value. For calls (if at all), the exercise region must lie in regions where S is large since for small S (for $S < K$), the intrinsic value of the call is zero.

We now consider the iteration scheme 10.17. In this scheme, we start by determining the value of $W_{M-1,j}$. This is an option value for large S. From the above consideration we know that an optimal early exercise of an American *call* option cannot happen in the small S region, but *could* happen in the large S region. Therefore we have to allow for this possibility at the grid point (S_{M-1}, t_j). Should it indeed be optimal to exercise in this time step then the value $W_{M-1,j}$ must be replaced by the intrinsic value. This fact can be incorporated into our scheme using the expression

$$\widetilde{W}_{M-1,j} = \max\left[\frac{x_{M-1}}{h_{M-1}}, S_{M-1} - K\right]$$

for our calculation of $\widetilde{W}_{M-1,j}$ instead of 10.17. We thus obtain the correct boundary value $\widetilde{W}_{M-1,j}$ for the American call option. In the next step we use

$$\widetilde{W}_{M-2,j} = \max\left[\frac{x_{M-2} - u_{M-2}\widetilde{W}_{M-1,j}}{h_{M-2}}, S_{M-2} - K\right]$$

instead of the corresponding expression 10.17 for the European option. Likewise, this leads to the correct result since the value $\widetilde{W}_{M-1,j}$ for an American option was calculated correctly before. This means, however, that an American

call option is correctly computed even with implicit or mixed methods if we simply substitute

$$\widetilde{W}_{i,j} = \max\left[\frac{x_i - u_i \widetilde{W}_{i+1,j}}{h_i},\ S_i - K\right]$$

for $i = M - 2, M - 3, \ldots, 2, 1$ at every time step. The procedure just described started with large S (Index $i = M - 1$) and computes step-wise to smaller values of S. This makes it only suitable for *call* options.

The procedure must be modified for *puts*. Instead of an **LU** decomposition we now carry out a **UL** decomposition, for which the upper and lower triangular matrices appear in reverse order. The decomposition of the coefficient matrix **A** is then given by:

$$\mathbf{A} = \underbrace{\begin{pmatrix} h_1 & u_1 & 0 & \cdots & \cdots & & 0 \\ 0 & h_2 & u_2 & 0 & \cdots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & 0 \\ \vdots & & & 0 & \ddots & u_{M-2} \\ 0 & \cdots & \cdots & \cdots & 0 & h_{M-1} \end{pmatrix}}_{\mathbf{U}} \underbrace{\begin{pmatrix} 1 & 0 & \cdots & \cdots & & \cdots & 0 \\ l_2 & 1 & 0 & & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & l_{M-2} & 1 & 0 \\ 0 & \cdots & \cdots & 0 & l_{M-1} & 1 \end{pmatrix}}_{\mathbf{L}}$$

The matrix elements differ only slightly for those given by the **LU** decomposition:

$$
\begin{aligned}
u_i &= c_i && \text{for } i = 1, \ldots M - 2 \\
h_{M-1} &= b_{M-1} \\
l_i &= \frac{a_i}{h_i} && \text{for } i = 2, \ldots M - 1 \\
h_i &= b_i - l_{i+1} u_i && \text{for } i = 1, \ldots M - 2
\end{aligned}
$$

The decisive difference is that h_i is iteratively determined from large values down to small ones (instead of small to large). Proceeding in this way, the matrix elements of **U** and **L** can be easily calculated. Analogously to the case of the **LU** decomposition we set

$$\mathbf{A}\mathbf{W}_j = \mathbf{U}\underbrace{\mathbf{L}\mathbf{W}_j}_{\mathbf{x}} =: \mathbf{U}\mathbf{x}$$

and first solve the system of equations for the new vector **x:**

$$
\begin{pmatrix}
h_1 & u_1 & 0 & \cdots \cdots & & 0 \\
0 & h_2 & u_2 & 0 & \cdots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & \ddots & \ddots & 0 \\
\vdots & & & 0 & \ddots & u_{M-2} \\
0 & \cdots \cdots \cdots & & 0 & h_{M-1}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
\vdots \\
\vdots \\
x_{M-1}
\end{pmatrix}
=
\begin{pmatrix}
d_{1,j} \\
d_{2,j} \\
d_{3,j} \\
\vdots \\
\vdots \\
d_{M-2,j} \\
d_{M-1,j}
\end{pmatrix} .
$$

To compute **x** we start with the row in the matrix which contains only one non-zero element, i.e., the last row, and proceed from bottom to top to obtain

$$
x_{M-1} = \frac{d_{M-1,j}}{h_{M-1}}
$$

$$
x_i = \frac{d_{i,j} - u_i \, x_{i+1}}{h_i} \quad \text{for} \quad i = 1 \dots M - 2 .
$$

Now that **x** is known, the desired option values $W_{i,j}$ can be calculated from the definition of **x**, solving the equation **L** $W_j = \mathbf{x}$:

$$
\begin{pmatrix}
1 & 0 & \cdots & \cdots & & \cdots & 0 \\
l_2 & 1 & 0 & & & & \vdots \\
0 & \ddots & \ddots & \ddots & & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\
\vdots & & \ddots & l_{M-2} & 1 & 0 \\
0 & \cdots \cdots & & 0 & l_{M-1} & 1
\end{pmatrix}
\begin{pmatrix}
W_{1,j} \\
W_{2,j} \\
\vdots \\
\vdots \\
\vdots \\
W_{M-1,j}
\end{pmatrix}
=
\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
\vdots \\
\vdots \\
x_{M-1}
\end{pmatrix} .
$$

To "invert" the matrix **L**, we start with the row containing only one non-zero element, working from top to bottom to obtain

$$
W_{1,j} = x_1
$$

$$
W_{i,j} = x_i - l_i \, W_{i-1,j} \quad \text{for} \quad i = 2 \dots M - 1 .
$$

Now, the iteration begins with the small index values, i.e., with small S-values. This is what we need since from the limit considerations above we know that

optimal early exercise of an American put option can only happen in the small S region. Therefore we have to start in this region to "capture" the early exercise possibilities correctly. This is completely analogous to the case of an American call presented above, where we also started in the region where early exercise could be possible (it was the large S region in that case).

Consequently, we now obtain correct values of American put options for implicit and mixed methods by setting:

$$\widetilde{W}_{1,j} = \max\left[x_1, S_1 - K\right]$$
$$\widetilde{W}_{i,j} = \max\left[x_i - l_i\widetilde{W}_{i-1,j}, S_i - K\right] \quad \text{for} \quad i = 2 \ldots M - 1.$$

10.3 Convergence Criteria

As for all numerical methods, the essential question to be answered before implementing finite differences is whether the procedure is stable, i.e., whether the numerical solution in fact converges towards the actual solution. To motivate this complex subject, we first consider as an example the special case $\theta = 0$, i.e., the *explicit* method. As can be seen from Eqs. 10.16 and 10.19, the off-diagonal terms are zero regardless of the type of boundary conditions used. This implies that the system of equations is completely uncoupled. For example, for Eq. 10.16, the system of equations reduces to

$$W_{0,j} = \left(1 - (t_{j+1} - t_j)\left(B_{1,j+1} + 2A_{1,j+1}\right)\right) W_{1,\ j+1}$$
$$\qquad - (t_{j+1} - t_j)\left(C_{1,j+1} - A_{1,j+1}\right) W_{2,j+1}$$
$$W_{i,j} = \left(1 - (t_{j+1} - t_j)B_{i,j+1}\right) W_{i,\ j+1} \qquad \text{for } i = 1 \ldots M - 1$$

(10.22)

$$\qquad - (t_{j+1} - t_j)A_{i,j+1}\ W_{i-1,j+1} - (t_{j+1} - t_j)C_{i,j+1}\ W_{i+1,j+1}$$
$$W_{M-1,j} = \left(1 - (t_{j+1} - t_j)\left(B_{M-1,j+1} + 2C_{M-1,j+1}\right)\right) W_{M-1,\ j+1}$$
$$\qquad - (t_{j+1} - t_j)\left(A_{M-1,j+1} - C_{M,j+1}\right) W_{M-2,j+1}.$$

This means that we obtain the unknown values at time point t_j directly (without carrying out any matrix inversion!) in terms of the values already calculated for time point t_{j+1}. This result can also be obtained much simpler and more directly by starting with Eq. 10.4 for the explicit method rather than from the more general expression 10.6.

Table 10.1 Properties of the three most commonly used finite difference methods

Scheme	θ	Numerical effort	Convergence	Stability
Implicit	1.0	Large	Slow	Unlimited
Explicit	0.0	Minor	Slow	Limited
Crank-Nicolson	0.5	Large	Fast	Unlimited

At this point the reader could ask why we have taken the trouble of introducing the generalized expression involving θ when the explicit method is so simple. For this we have to understand how the parameter θ influences the properties of the finite difference scheme. An important criterion is stability. Table 10.1 contains an overview of the three most commonly used values of θ.

A difference scheme is called *stable* if small deviations from the correct solution do *not* grow arbitrarily fast (no faster than exponentially) with time.[6] If a difference scheme is not stable, errors in the approximation arising in an iteration step are amplified by calculating backwards in the time-grid from one iteration step to the next. Obviously, a difference scheme is only useful if it is stable, since deviations from the correct solution would otherwise grow faster than exponentially when calculating backwards through the grid, finally yielding a result with an arbitrarily large error. It can be shown that for $\theta \geq 0.5$, the finite difference scheme is always stable, independent of the choice of grid.

A difference scheme *converges* towards the correct solution if it is both *stable* and *consistent*. *Consistency* is satisfied if the difference scheme applied to a smooth function at a fixed time point gives an arbitrarily good approximation of the solution to the differential equation by making the grid fine enough. Consistency thus means "convergence for one time step". If the system is consistent and stable, i.e., the growth of the error over one iteration step is bounded, the scheme converges *entirely*, i.e., over all time steps.

There exists a criterion for the stability of an *explicit* difference scheme ($\theta = 0$) for the well-known heat equation 7.22 which, as we have seen, is closely related to the Black-Scholes equation. This criterion is (expressed in the variables from Eq. 7.22):

$$\frac{\delta\tau}{\delta x^2} \leq \frac{1}{2} \, .$$

---

[6]Independent of the number of time steps.

For the Black-Scholes equation, the stability condition is considerably more complicated [181]:

$$\frac{1}{2}\left[(r(t)-q(t))S\frac{\delta t}{\delta S}\right]^2 \leq \frac{1}{2}\sigma(S,t)^2 S^2 \frac{\delta t}{(\delta S)^2} \leq \frac{1}{2}\,. \tag{10.23}$$

We are, in fact, now dealing with two inequalities. The inequality appearing on the left in the above expression is generally satisfied for the usual values for r, q and σ. The second inequality however, poses a strong restriction in the application of the explicit method since it requires the number of time steps to increase quadratically with the number of the S-steps (i.e., as δS becomes smaller). A consequence of the S-dependence of this criterion is that the difference scheme can be stable on one part of the grid and unstable on the other. Such local instabilities often go unnoticed, but lead to incorrect results; for example, negative prices could arise. Such an example can be found in Hull [103]. The criterion 10.23 for stability can be generalized further: the generalized difference scheme for non-uniform grids remains stable for $\theta < 0.5$ when the following holds for the second inequality:

$$\frac{(t_{j+1}-t_j)}{(S_{i+1}-S_i)^2}\sigma_{i,j}^2 S_i^2 \leq \frac{1}{2(1-2\theta)}\,, \quad \theta < \frac{1}{2}\,.$$

The Crank-Nicolson method $\theta = 0.5$ has an advantage over the implicit and explicit methods in that it converges considerably faster. Because of the averaging of the finite differences at times t_j and t_{j+1}, this method is exact up to second order although the time derivative corresponds to only a first order approximation. Achieving the same degree of accuracy using the implicit method would require considerably more time steps. The Crank-Nicolson method loses some of its efficient convergence when the initial and/or boundary conditions become less smooth. In many applications, for instance, the first derivative of the initial condition is not continuous. This is already the case for the plain vanilla option (at the strike). However, the discontinuity at that point leads to a negligible inaccuracy in the option price. A significant error, however, can be found for the value of the delta, while the values for gamma at the point of the discontinuity can be so extreme that the error may exceed 100%. As a result, it may be the case that the gamma for the plain vanilla option cannot be computed with sufficient accuracy when the spot price equals the strike price (the first derivative of the payoff is discontinuous at this point). In such situations the implicit method is often the only suitable method for calculating gamma with sufficient accuracy.

Another possibility would be to use a three-time-level approximation of the time derivative which also has the effect of smoothing the oscillations. A detailed description of this and other procedures can be found, for example, in the works of Smith [176] and Willmott [191].

10.3.1   Improving the Convergence Properties

If only few time steps are chosen, we observe that the results oscillate strongly with the number of time steps used [176]. Nonetheless, as few as 20 time steps are already sufficient to obtain stable results for a plain vanilla option with a time to maturity of six months. But it is usually worthwhile to increase the number of time steps, depending on the accuracy required and the available computation time. Note however, that the improvement in the approximation resulting from increasing the number of time steps is limited. Improving the approximation requires a refinement of both the time-grid as well as the S-grid. The proportion of the number of t-points to the number of S-points in the grid depends on the lifetime of the option, the required width of the S-grid (see below) and the type of difference scheme applied. The Crank-Nicolson scheme, for example, requires a significantly coarser t-grid to attain comparable accuracy to that obtained from using the implicit method with the otherwise same parameters. Experimenting with these parameters is in any case recommended.

In general, it makes sense to start with a *uniform* grid. An analysis of the solution surface, i.e., the price function at each point of the grid quickly reveals information about the time regions at which new grid points should be introduced. Figure 10.1 shows such a solution surface for an American call option with discrete dividend payments (see also the next section) approximately half way through its lifetime. Additional time points have been introduced at and immediately before the due date of the dividend payment. Doing so exactly incorporates the dividend payment into the calculation. We can clearly see where exercising shortly before the dividend payment is optimal (the jump occurs as a result). Likewise, it makes sense to introduce additional grid points in the time region shortly before the maturity of the option since there the curvature of the solution surface is quite large.

At every time point where an external shock such as a dividend payment occurs, the addition of grid points prevents the oscillation of the solution dependent on the number of t-grid points. This also holds for the S-grid at the points where the payoff function is discontinuous. For example, it makes sense to choose the strike price as a grid point, and also the underlying value S
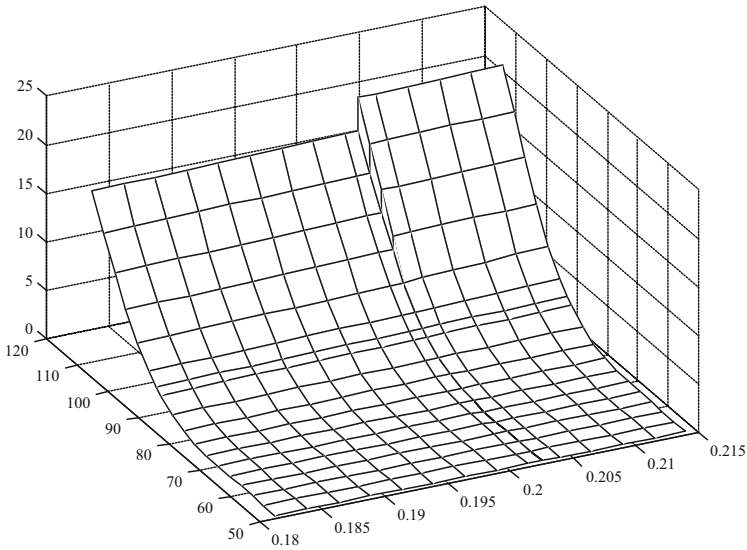
Fig. 10.1 Part of the solution surface of an American plain vanilla call on an underlying with a discrete dividend payment

for which the option price is to be determined. Likewise the barrier of a barrier option should ideally be directly on the grid. However, a non-uniform grid can *reduce* accuracy, since, for example, the Crank-Nicolson scheme is exact up to second order only for a *uniform* grid in t. In practice serious consideration should thus be giving as to where and whether additional grid points should be introduced.[7]

A good choice of whole S-region, i.e., of S_0 and S_{M+1} can substantially increase the accuracy attained. A rule of thumb is commonly used for choosing S_0 and S_{M+1}, such as

$$S_0 = \frac{1}{4} \min(S, K)$$

$$S_{M+1} = 4 \max(S, K) \, ,$$

where here, S is the spot price of the underlying at time $t = 0$ and K the strike price of the option. The S-region thus obtained, however, (depending on the values of other parameters such as the volatility and the interest rate

---

[7]Experience shows that increasing/decreasing the step size in a regular manner works often well.

term structure) seldom leads to an optimally accuracy if the t-grid and the number of S-steps are assumed as given.

Alternatively, we could attempt to find the minimum (and maximum) value the underlying can attain in its lifetime *with a given probability*. This can be done if the underlying's cumulative probability distribution P has already been established. In the simplest cases, even an analytic formula for this probability can be derived (see Sect. A.4). In order to obtain minimal (maximal) S-values for a given probability (confidence) c, the expression $P(x \leq a) = c$ must be solved for a, the *percentile* of the underlying's distribution corresponding to the confidence level c. It is often worth the trouble to make this single calculation for each valuation to establish the S-region since it can result in a more precise solution with substantially fewer S-steps.

10.4 Discrete Dividends

Consideration of discrete dividends represents a further complication. The particular difficulty is that the different methods available rely on different assumptions. One commonly used method is based on the separation of the stochastic and deterministic components of the stochastic process S. The dividend payments correspond to the deterministic component. One assumption is thus that the exact value of the dividend payment is known a priori. The present value D_t of the dividends which are to be paid until maturity of the option is subtracted from the spot rate S of the underlying:

$$\widetilde{S} = S - D_t .$$

Instead of S we now use $\widetilde{S}$ as the new variable in the Black-Scholes equation. D_t is the nominal value of the dividends discounted back to t (where t denotes the time step currently under consideration). This process is consistent with the Black-Scholes formula for the pricing of a plain vanilla European option. There, discrete dividends are generally taken into consideration by subtracting the present value of the dividends from the spot price of the underlying. Consequently, the same volatility holds for both cases. Note however, that the *intrinsic* value of the option is still given by

$$\max(S - X, 0) = \max(\widetilde{S} + D_t - X, 0) .$$

This value is required to determine the free boundary condition for an early exercise of the option.

In the world of finite differences, we have an additional and fundamentally different method at our disposal. We first rewrite Eq. 7.8 replacing the dividend term $q(t)S$ resulting from the dividend rate $q(t)$ with $D(S, t)$:

$$\frac{\partial V}{\partial t} + (r(t)S - D(S, t))\frac{\partial V}{\partial S} + \frac{1}{2}\sigma(S, t)S^2\frac{\partial^2 V}{\partial S^2} = r(t)V .$$

The new term $D(S, t)$ represents the nominal value of the dividend payments at time t. If no dividends are paid at time t, the value of this function $D(S, t)$ is equal to zero. $D(S, t)$ is thus highly discontinuous. At time τ of a dividend payment, the option price must remain continuous, although the spot price of S is reduced by the amount given by the dividend payment D:

$$\lim_{\epsilon \to 0} V(S(\tau - \epsilon), \tau - \epsilon) = \lim_{\epsilon \to 0} V(S(\tau + \epsilon), \tau + \epsilon) \quad \epsilon > 0 ,$$

where

$$\lim_{\epsilon \to 0} S(t - \epsilon) = S$$

$$\lim_{\epsilon \to 0} S(t + \epsilon) = S - D .$$

This jump in S can be simulated with finite difference methods by first determining the solution vector at time τ. Subsequently, the vector is translated by D so that the continuity condition is satisfied. If necessary, missing intermediate values must be determined by interpolation, or in the case of boundary values, by extrapolation.

10.5 Example

The download website [50] accompanying this book contains a complete, executable example program written in Visual Basic (see FINITEDIFFER-ENCEMETHOD.XLSM). The program was not written for optimal performance, but was structured in such a way as to incorporate all the concepts introduced in this section. For example, the S-grid is recalculated for every new time step (which, of course, need not be done if the grid is uniform). The volatility, interest rates and payoff are separated into individual functions to clearly organize the points at which such structures could be loaded externally via interfaces. If, for example, the interest rate or volatility is constant, this effort is superfluous. The program is structured to be, for the most part, self-

explanatory with helpful comments incorporated into the code. It can serve as a starting point for generating optimal-performance variants suitable for specific problems. This Excel workbook also demonstrates the calculation of option sensitivities, known as *Greeks*, which will be introduced in detail later in Chap. 12. Thus, this workbook can serve as a complete little option calculator (the yellow fields are input fields, a standard which applies to all workbooks on the accompanying website [50]).

In addition to the L-U decomposition introduced above, the example includes a further method for solving a system of equations, which is widely used in practice. The method is a very old procedure referred to as *Gaussian elimination*. Gaussian elimination is just as fast as the L-U decomposition and can be applied to both American and European options. The idea is even simpler than that of the L-U decomposition. Instead of decomposing the matrix in Eq. 10.15 into the matrix product of an upper and lower triangular matrix, it is transformed into a single triangular matrix with non-zero terms appearing only in the diagonal and a single off-diagonal. There are two possible ways of doing this: the matrix can be transformed into an upper or lower triangular matrix by eliminating the lower off-diagonal or upper off-diagonal terms respectively.

The first element in the lower off-diagonal, a_2, is eliminated by multiplying the first row of the matrix by a_2/b_1, and subtracting the result from the second row. After doing so, the second row contains only two elements, namely $b_2' = b_1 - c_1 a_2/b_1$ in diagonal and c_2 in the upper off-diagonal. This procedure is repeated analogously by multiplying the second row by a_3/b_2' and subtracting from the third, and so on. In terms of the system of equations, this means that certain equations are multiplied by constants and subtracted from one another. These are the usual operations implemented when solving a system of equations. In performing such operations, it is clear (in contrast to the L-U decomposition) that the right-hand side of the equation also changes. Carrying out this procedure for the entire matrix yields the following system of equations:

$$
\begin{pmatrix}
b_1' & c_1 & 0 & \cdots & \cdots & 0 \\
0 & b_2' & c_2 & 0 & & \cdots \\
\vdots & 0 & b_3' & c_3 & \ddots & \cdots \\
\vdots & & \ddots & \ddots & \ddots & 0 \\
\vdots & & & \ddots & \ddots & c_{M-2} \\
0 & \cdots & \cdots & \cdots & 0 & b_{M-1}'
\end{pmatrix}
\begin{pmatrix}
W_{1,\,j} \\
W_{2,\,j} \\
\vdots \\
\vdots \\
\vdots \\
W_{M-1,\,j}
\end{pmatrix}
=
\begin{pmatrix}
d_{1,\,j}' \\
d_{2,\,j}' \\
\vdots \\
\vdots \\
d_{M-2,\,j}' \\
d_{M-1,\,j}'
\end{pmatrix}
$$

where

$$b'_1 = b_1 , \quad d'_{1,\,j} = d_{1,j}$$

$$b'_i = b_i - c_{i-1}\frac{a_i}{b'_{i-1}} , \quad d_{i,\ j}^{\,\prime} = d_{i,\,j} - d'_{i-1,\,j}\frac{a_i}{b'_{i-1}} , \quad \forall i = 2, \ldots, M - 1 .$$

This system can be solved quite easily by starting with the row containing only one non-zero term, in this case the last row, and working upwards from the bottom. We obtain the option values

$$\widetilde{W}_{M-1,\,j} = \max\left[\frac{d'_{M-1,\,j}}{b'_{M-1}}, P(S_{M-1})\right]$$

$$\widetilde{W}_{i,\,j} = \max\left[\frac{d'_{i,\,j} - c_i\,\widetilde{W}_{i+1,\,j}}{b'_i}, P(S_i)\right] \quad \forall i = 1, \ldots, M - 2 .$$

To this extent, this method offers an alternative to the L-U decomposition for the valuation of European options. The free boundary condition for American options is dealt with, as in the method of Lamberton and Lapeyre, by calculating the value of the option by Gaussian elimination as if it were European and, at each time step, comparing it with the intrinsic value of the option $P(S)$, i.e., the payoff profile. The value of the American option at this time point is then defined as the greater of the two. The components of the solution vector $\widetilde{W}_{i,j}$ are thus given by

$$\widetilde{W}_{M-1,\,j} = \max\left[\frac{d'_{M-1,\,j}}{b'_{M-1}}, P(S_{M-1})\right]$$

$$\widetilde{W}_{i,\,j} = \max\left[\frac{d'_{i,\,j} - c_i\,\widetilde{W}_{i+1,\,j}}{b'_i}, P(S_i)\right] \quad \forall i = 1, \ldots, M - 2 .$$

As was already stressed earlier when we presented the method of *Lamberton and Lapeyre*, this procedure functions only when we calculates along the *S*-grid points starting from the exercise region, where the value of the option is given by its intrinsic value (the payoff profile), toward the region where the Black-Scholes equation holds. Otherwise, we would observe an edge or jump at the transition from one region into the other. The Gaussian elimination procedure described above starts from large *S* (index $i = M-1$) and calculates backwards

step by step towards smaller values of S. This works fine for call options since early exercise of a call can only (if at all) be optimal for large S because for small S (for $S < K$) the intrinsic value of the call is of course zero.

For put options, however, the exercise region can only lie in areas where the value of S is small, since for large S (for $S > K$), the intrinsic value of the put is zero, the option itself having only a time value. Therefore the second possibility for performing a Gaussian elimination has to be used for puts, namely eliminating the upper off-diagonal. This results in a procedure which begins with small S proceeding step by step towards larger S-values. The upper off-diagonal is eliminated by multiplying the last row of the matrix by c_{M-2}/b_{M-1}, subtracting the result from the next to last row. Proceeding analogously, the newly created next to last row is multiplied by c_{M-3}/b'_{M-2} and subtracted from the row lying immediately above it, and so on. Performing this procedure for the entire matrix yields the following system of equations:

$$
\begin{pmatrix}
b'_1 & 0 & \cdots & \cdots & \cdots & 0 \\
a_2 & b'_2 & 0 & & & \vdots \\
0 & a_3 & b'_3 & 0 & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & a_{M-2} & b'_{M-2} & 0 \\
0 & \cdots & \cdots & 0 & a_{M-1} & b'_{M-1}
\end{pmatrix}
\begin{pmatrix}
W_{1,\,j} \\
W_{2,\,j} \\
\vdots \\
\vdots \\
\vdots \\
W_{M-1,\,j}
\end{pmatrix}
=
\begin{pmatrix}
d'_{1,\,j} \\
d'_{2,\,j} \\
\vdots \\
\vdots \\
d'_{M-2,\,j} \\
d'_{M-1,\,j}
\end{pmatrix}
$$

where

$$
b'_{M-1} = b_{M-1}\,, \quad d'_{M-1,\,j} = d_{M-1,\,j}
$$

$$
b'_i = b_i - a_{i+1}\frac{c_i}{b'_{i+1}}\,, \quad d'_{i,\,j} = d_{i,\,j} - d'_{i+1,\,j}\frac{c_i}{b'_{i+1}}\,, \quad \forall i = 1, \ldots, M-2\,.
$$

This system can be solved quite easily by starting with the row containing only one non-zero element, in this case the first row, and proceeding from top to bottom. We thus obtain the option values

$$
W_{1,\,j} = \frac{d'_{1,j}}{b'_1}
$$

$$
W_{i,\,j} = \frac{d'_{i,j} - a_i W_{i-1,j}}{b'_i} \quad \forall i = 2, \ldots, M-1\,.
$$

The components of the solution vector $\widetilde{W}_{i,j}$ for an American option with payoff $P(S)$ is given by

$$\widetilde{W}_{1,\ j} = \max\left[\frac{d'_{1,j}}{b'_1},\ P(S_1)\right]$$

$$\widetilde{W}_{i,\ j} = \max\left[\frac{d'_{i,j} - a_i\ \widetilde{W}_{i-1,\ j}}{b'_i},\ P(S_i)\right] \quad \forall i = 2,\ldots,M-1\ .$$

10.5.1 Relationship Between Explicit Finite Difference and Tree Methods

Numerically, the trinomial tree and the explicit method of finite differences are equivalent, at least in certain special cases. Consider the ith node of the time step j. This corresponds to an underlying price S_{ji}. This node is connected to the nodes $i+1$, i and $i-1$ of the time step $j+1$. The value of the option V_{ji} at node i of the time step j is given by

$$V_{ji} = B(t_j, t_{j+1})\left[p^+ V_{j+1,i+1} + p^0 V_{j+1,i} + p^- V_{j+1,i-1}\right]\ .$$

The analogous expression for the explicit finite difference method is given by Eq. 10.22 with the coefficients given by Eq. 10.12 (or Eq. 10.10 for uniform grids):

$$W_{i,j} = \left(1 - (t_{j+1} - t_j)B_{i,j+1}\right)W_{i,\ j+1} - (t_{j+1} - t_j)A_{i,j+1}\ W_{i-1,j+1}$$
$$- (t_{j+1} - t_j)C_{i,j+1}\ W_{i+1,j+1}\ .$$

We immediately see the similarity in structure. The probabilities can be easily associated with the coefficients of $W_{i+1,j+1}$, $W_{i,j+1}$ and $W_{i-1,j+1}$. This requires the grid points of the finite difference grid to correspond exactly to the nodes of the trinomial tree and thus the restriction $S_{ji} = S_{j+1,i}$ to hold for the nodes of the trinomial tree. The finite difference grid was constructed specifically to satisfy just this very condition (even for non-uniform grids) to prevent the expressions for the finite differences from becoming unnecessarily complicated. The trinomial tree appears more flexible from this point of view. However, it has all the disadvantages associated with the explicit method. Moreover, a procedure corresponding to the implicit or Crank-Nicolson method is not available for trinomial trees.

11

Monte Carlo Simulations

Having recognized the fact that prices of financial instruments can be calculated as discounted future expectations (with respect to a risk-neutral probability measure), the idea of calculating such expectations by simulating the (stochastic) evolution of the underlyings several times and subsequently averaging the results somehow is not far removed. In fact, this relatively simple idea is widely used and is successful even in the valuation of very exotic options for which other methods are either too complicated or completely unsuitable, the only requirement being the availability of sufficient computation time. Before proceeding with financial *applications* of Monte Carlo techniques, we begin with a presentation of the technique itself.

If random events occur often enough, they can be used to answer diverse questions statistically. This has long been common knowledge in science and we have seen a vast increase in applications with the advance of modern computers, since computers suddenly made it possible to generate "random" events cheaply and in large numbers, or in the language of the specialist, to *simulate* them. Ever since, *computer simulations* have been indispensable in science, and since lately also in the modern financial world. Since generating "random" events lies at the core of such simulations, the name *Monte Carlo simulation* has become accepted despite the fact that the method's namesake city in Monaco could never generate as many random events as are sometimes necessary in practice, even if all the casinos in Monte Carlo were open non-stop for business every day for a million years. From this point of view, the computer can far outperform the roulette table.

With simulations, it has become possible to solve problems for which classical solutions fail, e.g. if the dimensionality of the problem is to high (regarding for example most equity basket products with three or more different stocks in the basket), or if the problem at hand could not be formulated in terms of a partial differential equation (e.g. the multi-factor LIBOR market model). Use of Monte-Carlo simulations, on the other hand, allow for the direct simulation of stochastic differentials. Even complex path dependencies with the final pay off depending highly on the evolution path of the underlying over time can be modeled by means of Monte Carlo methods without difficulty, since the various paths are simulated directly anyway. Lattice methods (trees, finite differences) can be used to solve path-dependent problems only in some special, simple cases or with great computational effort.

In addition, computer simulations allow to perform a what-if-analysis, with which in short time many different scenarios could be simulated and analyzed. The scenarios to be simulated could be given as fixed parameter sets (*static simulation*, without any stochastic parameters, and are therefore no Monte Carlo simulations) or randomly synthesized by application of a set of rules (*dynamic Monte Carlo simulation*). The recently becoming increasingly popular stress tests are nothing else then computer simulations, too. Finally, computer simulations are fairly easy to implement and understood.

This advantages do have a price: computer simulations require lots of computation time, and especially for low-dimensional problems, this is a significant disadvantage with respect to lattice methods. Also, it is very difficult and time intensive to calculate sensitivities with required accuracy. Therefore, MC simulations are often the method of last resort, i.e. MC simulations should in general only be used if other alternative methods fail.

Assumptions

A Monte Carlo simulation in its most general form consistent with arbitrage-free pricing requires only the assumptions eliminating arbitrage opportunities, i.e., Assumptions 1, 2, 3, 4 and 5 from Chap. 4. In addition, at first, we constrain our considerations to European options only. As will be shown later, the valuation of American Options with Monte Carlo methods requires significant extra effort.

The Monte Carlo method presented in the following is based on the random walk equation 2.17 with constant drift and volatility. Hence, because of Eq. 9.25 for pricing in a risk-neutral world, constant yields must also be assumed. This means that for the method presented, the additional

Assumptions 7, 9, 11 and 12 from Chap. 4 are also made. The assumptions of constant interest rates and volatilities naturally imply that interest rates and volatilities are non-stochastic, i.e., Assumptions 8 and 10 hold as well. Generalization to time-dependent parameters is possible without problems, though.

These extensive assumptions are made to allow a clear presentation of the material but are, in principle, not necessary for performing Monte Carlo simulations. For example, not only the underlying price, but also the volatility can be simultaneously simulated as stochastic processes (for instance also as a random walk). Since then two of the parameters involved are stochastic processes, the random walk occurs now in two dimensions. Of course, the volatility of the volatility and also the drift of the volatility would be required as parameters in such a situation. The simulation then proceeds as follows: first, a value for the volatility is simulated. Using this simulated volatility as a parameter, the next step for the underlying price is simulated. The two random walks are thus not independent of one another since the random walk describing the volatility affects that of the price. In reality, it can often be observed that the converse holds as well, i.e., the price of the underlying has an influence on the volatility (low volatility for rising prices and high volatility for falling prices). This effect can be incorporated into the simulation using a special form for the drift of the volatility random walk, which depends on the underlying price, for example as "volatility-drift = $rdt/S(t)$".

When simulating interest rates for pricing caps or floors, for example, Eq. 2.17 can naturally be replaced by a more complex process corresponding to a term structure model such as Heath-Jarrow-Morton, Ho-Lee, Hull-White, etc. Mean reversion, for example, intuitively corresponds to a random walk under the influence of an external force which has the effect of "driving" the random walk back to an asymptotic mean value.

11.1 A Simple Example: The Area of a Disk

First, we will make some effort to understand clearly the essence of the Monte Carlo simulation, before the method will be applied to an actual problem. Therefore, at this point we present a very simple example of the idea behind the Monte Carlo method before entering into a discussion of applications in the financial world. As known from elementary mathematics, the area of a disk of radius R is πR^2, the product of the square of the radius with the constant π, already known to the ancient Egyptians: $\pi = 3.14159...$ A disk with a radius of one meter, $R = 1m$, thus has an area of $\pi \times (1m)^2 = 3.14159m^2$.
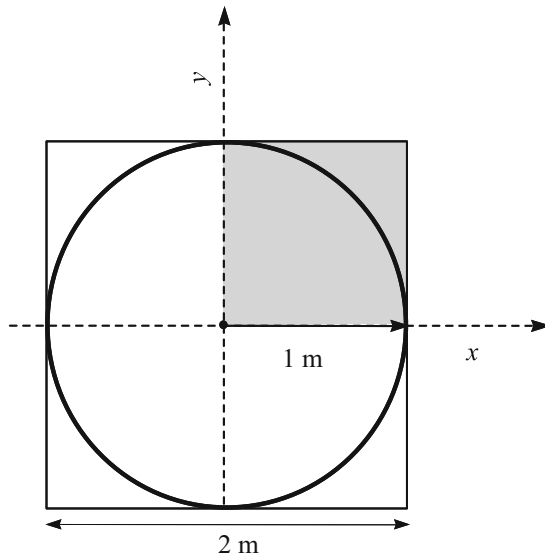
Fig. 11.1 A disk with diameter $2m$ in a square with sides of length $2m$

Supposing we had never heard of the number π, this fact can be ascertained with the help of random events. To do so, we simply place square box with sides of length $2m$ containing a "pie dish" as shown in Fig. 11.1 out in the rain. The pie dish represents the disk of radius $1m$ whose area is to be determined. We know that the square has an area of $2m \times 2m = 4m^2$. The random events are the falling raindrops. Assuming that the raindrops fall evenly on the square, then the area of the disk can be given by

$$\text{disk area} = \frac{\text{number of raindrops falling on the disk}}{\text{number of raindrops falling on the entire square}} \times 4m^2 \ .$$

In this way, the area of a disk can be determined with the help of random events.

The same "Experiment" admits another interpretation. If the equation "disk area $= \pi R^2$" is known, then the value of π can be determined:

$$\pi = \frac{\text{disk area}}{R^2} = \frac{\text{disk area}}{1m^2}$$

$$= 4 \times \frac{\text{number of raindrops falling on the disk}}{\text{number of raindrops falling on the entire square}} \ .$$

Thus, with the help of randomly falling raindrops, the value of the natural constant π has been determined.

The primary application of Monte Carlo methods—in particular in the financial world—is the calculation of integrals. For example, the price of an option can be expressed as the integral of its payoff profile with respect to an appropriate probability measure associated to the price of its underlying at maturity. Making use of our disc-example, we give a third interpretation of the experiment described above to illustrate how random events can be used to calculate an integral: as is known from elementary mathematics, a semi-circle can be represented as the graph of a function. Choosing the x- and y-axis as shown in Fig. 11.1, the ancient *Pythagorean Theorem* says that the points (x, y) on the circle all satisfy $x^2 + y^2 = R^2 = 1$, thus the upper semi-circle is the graph of the function $y = \sqrt{1 - x^2}$. The lower semi-circle is the graph of the function $y = -\sqrt{1 - x^2}$. As is well know from any introduction to mathematical analysis, integrating a function gives the area under the curve given by the graph of that function. Since the upper semi-circle is represented by the graph of the function $y = \sqrt{1 - x^2}$, the area of the half-disk can be determined by integrating:

$$\int_{-1}^{1} \sqrt{1 - x^2}dx = \frac{1}{2}\text{Area of the disk}$$

$$= 2\frac{\text{Number of raindrops falling on the disc}}{\text{Number of raindrops falling on the entire square}} .$$

Thus, we have calculated this integral with the help of randomly falling raindrops.

Note that the quantity we want to calculated using random events is by no means random itself: the area of a disk of radius $1m^2$ never changes nor does the value of π or the value of the above integral. We could ask the question: does rain know anything about geometry and circles? Or about the natural constant π? Or about integration? Probably not. However, this information can be obtained if we are clever enough to ask the right questions! The only condition, the randomly falling raindrops had to meet, was that they fall evenly, or more precisely, that the probability for them to fall on any particular point in the square is exactly the same as to fall on any other point in the square. For example, on a square field with an area of 1000 square miles, it might be raining in some places and sunny in others. Such a field would be completely unsuitable for our experiment.

The "machine" used to generate the random events, whether the rain or a computer, is completely irrelevant. The events merely have to be "sufficiently random" and their probability distribution must be known. If this is the case, calculations can be performed with an accuracy which is limited solely by the machine's capacity to produce the random events: according to the laws of statistics, the accuracy of results improves as the number of random events involved in the simulation increases (see Sect. 31.2).

How would the above experiment have been conducted with a computer? Nowadays, most programming languages such as Pascal, Fortran, C, C++, etc. and the common spread sheet programs, such as Microsoft Excel (or Visual Basic) or Lotus 1-2-3, are equipped with *random number generators*. These are (small) programs or functions which usually generate uniformly distributed *random numbers* between 0 and 1. Simulating a random event such as a falling raindrop with a computer is accomplished by generating *two* such random numbers Z_1 and Z_2 to simulate the coordinates (one x- and one y-coordinate) of the point on which the raindrop falls. Since the simulated random numbers are all between 0 and 1 the generator producing such coordinates simulates raindrops falling only in the shaded area in Fig. 11.1. To simulate raindrops falling on the entire square, the following transformation must be made:

$$x = 2Z_1 - 1 , \quad y = 2Z_2 - 1 .$$

Because this transformation is linear, x and y are uniformly distributed random variables, as are Z_1 and Z_2. Thus, the coordinates of a raindrop in the square have been determined. The simulation of the random event is complete and we can continue with the evaluation. As in the case of raindrops, two events must be counted:

- The total number of "raindrops" falling in the square. This, however, is exactly the number of simulated random events, since no coordinates were generated which lie outside of the square (we make no unnecessary simulations). This means that this counter will be increased by one after each random event has been simulated.
- The number of "raindrops" falling within the circle. According to *Pythagoras*, these are the random events whose coordinates x and y satisfy the inequality $x^2 + y^2 \leq 1$. This means that after each random event, the counter is only increased by one if the generated event satisfies this inequality.

If, for example, 10,000 events were simulated (this requires the generation of 20,000 random numbers, one for each coordinate), of which, for example, 7851 satisfy the above inequality, we obtain

$$\text{Area of the disk} \approx 0,7851 * 4m^2 = 3,1404m^2$$

$$\pi \approx 0,7851 * 4 = 3,1404$$

$$\int_{-1}^{1} \sqrt{1 - x^2}\, dx \approx 0,7851 * 2 = 1,5702 \ .$$

Of course, this result is not exact, but the statistical error involved can be determined as described in Sect. 31.2. In principle, this error can be reduced almost arbitrarily by increasing the number of simulated random events, as long as computer capacity allows it.

11.2 The General Approach to Monte Carlo Simulations

We can extract the general procedure for conducting a Monte Carlo simulation from the above example which will be summarized here to provide the reader with a "recipe" for performing such simulations. Each step will be explained by reference to its corresponding step in the above disk experiment as well as to the simulation of a random walk.

1. **Generate the random numbers required for a Monte Carlo step, usually from a uniform distribution between 0 and 1.**
   In the disk experiment, a Monte Carlo step corresponds to the falling of a simulated "raindrop". Two random numbers were required. To simulate a random walk, a Monte Carlo step represents a step in the random walk. One random number is required for each dimension of the space in which the random walk occurs.
2. **Transform the random numbers to generate numbers according to a desired distribution.**
   In the disk experiment, the uniformly distributed random numbers generated between 0 and 1 were transformed to yield uniformly distributed random numbers between -1 and 1. For a random walk, they are trans-

formed to normally distributed random numbers. Section A.5 presents methods for doing this.

3. **Perform a Monte Carlo step using the random numbers generated.**
   In the case of the disk, this corresponds to the falling of a "raindrop". For the random walk, it consists of adding one time step to the random walk.
4. **Repeat steps 1, 2 and 3 until the system reaches the state required for the proposed investigation.**
   For a random walk consisting of n steps, the required state is attained when n Monte Carlo simulations have been carried out. For the disk, the evaluation can take place after each single Monte Carlo step.
5. **Measurements: measure the system variables of interest.**
   In the case of the disk, we measure the number of "raindrops" falling within the disk. For a random walk, for example, the "end-to-end" distance could be measured.
6. **Repeat steps 1, 2, 3, 4 and 5 until enough systems for a statistical analysis have been generated**.
   For the disk, many "raindrops" must be simulated; for the price evolution of an underlying, many random walks must be generated.
7. **Final Analysis: compute the means of the measured variables and determine the statistical error.**
   For the disk, this was the ratio of the number of raindrops falling within the disk to the total number simulated. For a random walk it could be, for example, the mean length of all measured end-to-end vectors (or the square of their Euclidean norm).

11.3 Monte Carlo Simulation of Risk Factors

11.3.1 Simulation of the Evolution of a *Single* Risk Factor

On the basis of Eq. 2.17, the time interval from t to T is divided into n intervals of length δt where n equals the number of steps. The price of a risk factor is to be simulated as a random walk over this time interval. The random walk equation 2.17 (or its equivalent form 2.23) holds for infinitesimal changes in $\ln(S)$ occurring in an infinitesimally small time span dt. The time step δt used for the simulation is not infinitesimally small, however. Therefore we do not

use the stochastic PDE 2.17 (or 2.23) itself but its solution, Eq. 2.28. Taking the logarithm on both sides of this solution yields[1]:

$$\ln\left(S(t+\delta t)\right) = \ln\left(S(t)\right) + \mu\,\delta t + \sigma X\sqrt{\delta t} \quad \text{with} \quad X \sim N(0,1)\,.$$

This is a *recursion*: Knowledge of $\ln(S)$ at time t enables the calculation of $\ln(S)$ at the next time point $t + \delta t$. To emphasize this point, we enumerate the time points, i.e., we introduce the following notation:

$$t_i = t + i\,\delta t \quad \text{where} \quad i = 0, 1, \ldots, n \quad \text{d. h.} \quad t_0 = t\,, \quad t_n = T\,.$$

Denoting the ith (standard normally distributed) random number by $X_i = X(t_i)$, we obtain the basis of the simulation of a risk factor whose behavior is governed by Eq. 2.17 in terms of the notation just introduced:

$$\ln\left(S(t_i)\right) = \ln\left(S(t_{i-1})\right) + \mu\,\delta t + \sigma X_i\sqrt{\delta t}\,, \quad i = 1, \ldots, n\,. \,. \quad (11.1)$$

This equation is suitable for a direct simulation: $\ln(S(t_{i-1}))$ is the end-point of a random walk after $i-1$ steps. In the next step, the value "$\mu\,\delta t + \sigma X_i\sqrt{\delta t}$" will be added on to this end point, yielding the end point of the random walk after i steps, namely $\ln(S(t_i))$.

A concrete example at this point may clarify this procedure. With a volatility of $\sigma = 20\%$ per year, a drift of $\mu = 6\%$ per year and a time step $\delta t = 1$ day $= 1/365$ years, Eq. 11.1 gives the following simple recursion relation:

$$\ln\left(S(t_i)\right) = \ln\left(S(t_{i-1})\right) + \frac{0,06}{365} + 0,2\sqrt{1/356}\,X_i$$
$$= \ln\left(S(t_{i-1})\right) + 0,0001644 + 0,01046\,X_i\,.$$

The contribution of the drift to each step is approximately one hundred times smaller than that of the volatility. This explains the negligible effect of the drift for small time spans $T - t$. However, over longer time periods, the drift cannot be ignored. Figure 11.3 shows the result of a simulation over a longer

---

[1]To be able to work with standard normally distributed random numbers we also used Eqs. 2.16 and 2.27 here. Those Equations say that the Wiener-Process δW has the same distribution as $\sqrt{\delta t}$ times a standard normally distributed random number:

$$\delta W \sim X\sqrt{\delta t} \quad \text{with} \quad X \sim N(0,1)\,.$$
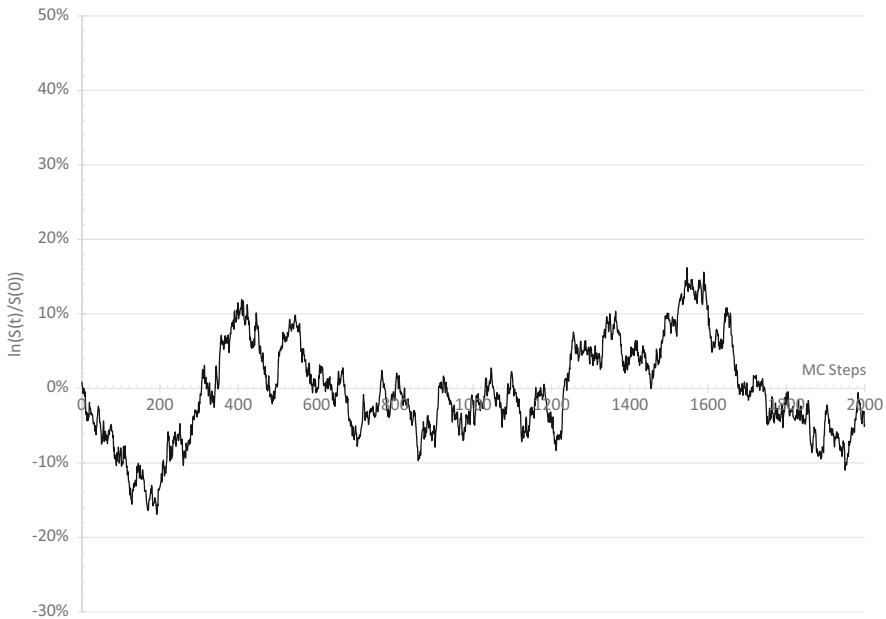
Fig. 11.2 Simulation of $\ln(S(t)/S(0))$ by means of normally distributed random numbers. 1 step corresponds to 1 day, with a total number of 2,000 steps. The annual volatility is 20% and the average drift is 0

time span of 2,000 days with respect to this recursion. To emphasize the effect of the drift, the 2,000 randomly generated values of X_i were saved and used again for the recursion, this time setting $\mu = 0$. The result is presented in Fig. 11.2. The values at the end of the simulation performed with the drift (Fig. 11.3) are approximately twice as large as those for the simulation with zero drift.

Such a curve represents one possible price progression over time, called a *path* of the risk factor. Repeating the simulation, we obtain an additional path. The simulation of many such paths yields the probability distributions for the price at each time point t_i in the simulated time span, in other words the probability distribution of the paths. In particular, we obtain a distribution of the values of the risk factor at the end point $t_n = T$. The simulated price of the risk factor at time $t_n = T$ is obtained by adding up the n steps generated in accordance with Eq. 11.1:

$$\ln\left(S(t_n)\right) = \ln\left(S(t_0)\right) + \mu \sum_{i=1}^{n} \delta t + \sum_{i=1}^{n} X_i \, \sigma \sqrt{\delta t} \ .$$

Fig. 11.3 The same random walk as in Fig. 11.2 but with a drift (mean return) of 6% per year

If the volatility σ is assumed to be constant, it can be factored out of the above sum. The length of the time steps δt are the same for each Monte Carlo step and thus

$$\ln\left(S(t_n)\right) = \ln\left(S(t_0)\right) + \mu\, n\, \delta t + \sigma\, \sqrt{\delta t}\, \sum_{i=1}^{n} X_i\ .$$

The sum of n independent, standard normally distributed random variables is again a normally distributed random variable with expectation 0 and variance n, i.e., the standard deviation of the sum is $\sqrt{n}$. If only the end distribution (the distribution of S_n) is of interest and not the *path* of the underlying (each of the S_i) taken *during* the time span under consideration, the sum of the n standard normally distributed random variables X_i can be replaced by a standard normally distributed random variable X multiplied by $\sqrt{n}$:

$$\sum_{i=1}^{n} X_i \rightarrow \sqrt{n}\, X\ .$$

In doing so, the simulated underlying price at the end of the time period under consideration can be generated with a *single* random number *directly*. Using $n\delta t = T - t$ from the definition of δt we can thus write

$$\ln\left(S(T)\right) = \ln\left(S(t)\right) + (T - t)\mu + \sigma\sqrt{T - t}\, X \ . \qquad (11.2)$$

The path taken by the underlying *during* the simulation is of interest only for certain exotic, *path-dependent* derivatives (see Chap. 19). Otherwise, in particular for pricing and risk management of European-style instruments where the underlying price is relevant only at maturity (or at the end of the *liquidation period* in the case of risk management), the efficiency of a Monte Carlo simulation can be significantly improved by this simplification.

11.3.2 Simulation of *Several* Correlated Risk Factors

We are often interested in the progression of several risk factors rather than just one. The risk management of a portfolio, for example, requires the simulation of all risk factors affecting the portfolio. Since those are usually not statistically independent of one another, we are confronted with the question of how the *correlation* between risk factor processes can be incorporated into the simulation. The general approach for an arbitrarily large number of different securities is presented in Sect. 23.1. Here, we restrict the discussion to the important special case of *two* correlated prices. This can be used, for example, in determining the price of *exchange options*.[2]

The two price processes S_1 and S_2 have drifts μ_1 and μ_2, volatilities σ_1 and σ_2, respectively, and a correlation ρ_{12}. The logarithm of the random walks will again be used to model the time evolution of the risk factors:

$$\delta \ln S_1(t) = \mu_1\,\delta t \,+\, Y_1$$

$$\delta \ln S_2(t) = \mu_2\,\delta t \,+\, Y_2 \text{ with correlated random variables } \quad Y_1,\ Y_2 \ .$$

How should correlated pairs of random variables be constructed? First, the two equations for $\delta \ln(S_i)$ are combined by interpreting the indexed equations as components of a random vector.

$$\begin{pmatrix} \delta\, \ln S_1(t) \\ \delta\, \ln S_2(t) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \delta t \,+\, \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \ .$$

---

[2]See Sect. 19.1.5.

If the prices were uncorrelated, the random variables Y_i would be independent, normally distributed random variables with variance $\sigma_i^2 \delta t$. However, since the prices are correlated, it is not sufficient to simply specify the variance of both variables in order to fully determine the distribution. Instead the *co*variance is needed to describe both the variances *and* the correlations. Since several Y_i (in this case, two) may come into play when constructing the random vector, the covariance is not a single number but a matrix, called the *covariance matrix*. As will be presented in detail in Sect. 21.5, the covariance matrix $\delta \Sigma$ of two random variables is composed of the correlations and the standard deviations of the associated random variables as follows:

$$\delta \mathbf{\Sigma} = \begin{pmatrix} \delta \Sigma_{11} & \delta \Sigma_{12} \\ \delta \Sigma_{21} & \delta \Sigma_{22} \end{pmatrix} \quad \text{where} \quad \delta \Sigma_{ij} = \underbrace{\rho_{ij}}_{\substack{\text{Correlation} \\ \text{von } i \text{ mit } j}} \underbrace{\sigma_i \sqrt{\delta t}}_{\substack{\text{Standard} \\ \text{dev. of } i}} \underbrace{\sigma_j \sqrt{\delta t}}_{\substack{\text{Standard} \\ \text{dev. of } j}} \text{for} \quad i, j = 1, 2 .$$

(11.3)

Correlations are symmetric, i.e., $\rho_{ij} = \rho_{ji}$. Also, $\rho_{ii} = 1$, in other words a risk factor is always fully correlated with itself. With $\rho = \rho_{12} = \rho_{21}$ the covariance matrix of two risk factors becomes

$$\delta \mathbf{\Sigma} = \delta t \begin{pmatrix} \sigma_1^2 & \rho\, \sigma_1 \sigma_2 \\ \rho\, \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} .$$

In order to generate normally distributed random numbers Y_i with correlation matrix $\delta \Sigma$ from independent, standard normally distributed random variables X_i, the "square root" $\mathbf{A}$ of the matrix $\delta \Sigma$ is needed. This matrix satisfies the condition

$$\mathbf{A}\mathbf{A}^T = \delta \Sigma ,$$

where $\mathbf{A}^T$ denotes the *transpose* of the matrix obtained by writing the column vectors of the matrix $\mathbf{A}$ as row vectors. As shown in detail in Sect. 21.5.3, this matrix yields the desired transformation

$\mathbf{Y} = \mathbf{A}\mathbf{X}$ where

$\mathbf{X} =$ vector of standard normally distributed, uncorrelated random variables

$\mathbf{Y} =$ vector of normally distributed, correlated random variables with

   covariance matrix $\delta \Sigma$.

The "square root of a matrix" can be obtained using a procedure from linear algebra called *Cholesky decomposition*. The general form of this decomposition is given in Sect. 21.5.3. Here, we restrict our consideration to 2×2 matrices, carrying out the procedure explicitly for this case. We begin by assuming that the matrix $\mathbf{A}$ has the following form:

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} \implies \mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} \\ 0 & a_{22} \end{pmatrix} .$$

The components a_{ij} can now be determined from the requirement that the equation $\mathbf{A}\mathbf{A}^T = \delta\boldsymbol{\Sigma}$ be satisfied:

$$\mathbf{A}\mathbf{A}^T = \delta\boldsymbol{\Sigma}$$

$$\begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} \\ 0 & a_{22} \end{pmatrix} = \begin{pmatrix} \delta\Sigma_{11} & \delta\Sigma_{12} \\ \delta\Sigma_{21} & \delta\Sigma_{22} \end{pmatrix}$$

$$\begin{pmatrix} a_{11}^2 & a_{11}a_{21} \\ a_{11}a_{21} & a_{21}^2 + a_{22}^2 \end{pmatrix} = \begin{pmatrix} \delta\Sigma_{11} & \delta\Sigma_{12} \\ \delta\Sigma_{21} & \delta\Sigma_{22} \end{pmatrix} .$$

Comparing the components on both sides yields a linear system of equations for the a_{ij}:

$$a_{11}^2 = \delta\Sigma_{11} \implies a_{11} = \sqrt{\delta\Sigma_{11}} = \sigma_1\sqrt{\delta t}$$

$$a_{11}a_{21} = \delta\Sigma_{12} \implies a_{21} = \frac{\delta\Sigma_{12}}{\sqrt{\delta\Sigma_{11}}} = \rho\,\sigma_2\sqrt{\delta t}$$

$$a_{21}^2 + a_{22}^2 = \delta\Sigma_{22} \implies a_{22} = \sqrt{\delta\Sigma_{22} - \frac{\delta\Sigma_{12}^2}{\delta\Sigma_{11}}} = \sqrt{1-\rho^2}\,\sigma_2\sqrt{\delta t} .$$

Where Eq. 11.3 was also used. Now that the matrix elements have been determined, this matrix can be used to generate the correlated random numbers:

$$\mathbf{Y} = \mathbf{A}\mathbf{X}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \sqrt{\delta t} \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sqrt{(1-\rho^2)}\sigma_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} ,$$

and thus

$$Y_1 = \sqrt{\delta t}\, \sigma_1\, X_1$$
$$Y_2 = \sqrt{\delta t}\, \rho\, \sigma_2\, X_1 + \sqrt{\delta t}\, \sqrt{1-\rho^2}\, \sigma_2\, X_2\,. \qquad (11.4)$$

We can now formulate a random walk equation for two *correlated* price processes in terms of two *un*correlated *standard* normally distributed random variables X_1, X_2:

$$\delta \ln S_1(t) = \mu_1\, \delta t\, +\, \sigma_1\, X_1\sqrt{\delta t}$$
$$\delta \ln S_2(t) = \mu_2\, \delta t\, +\, \sigma_2\, (\rho X_1 + \sqrt{1-\rho^2}X_2)\sqrt{\delta t}\,. \qquad (11.5)$$

The first equation has the form of a single random walk. The correlation affects only the second equation. The interpretation of this representation is that the *second* price is correlated with the *first*. Since the correlations are symmetric, this interpretation is irrelevant for the final result. Assuming another form for the matrix **A** (or simply renaming the Y_i) would yield the result that the *first* price is correlated with the *second*. This holds in general: we can select any risk factor (for example, the one we feel most comfortable working with) as the leading factor and simulate it independently. Then all correlations appear in the evolutions of the other risk factors.

Exactly as in the previous section, the random walk equations 11.5 now provide the basis for the recursion which can be programmed in a Monte Carlo simulation:

$$\ln S_1(t_i) = \ln S_1(t_{i-1}) + \mu_1\, \delta t\, +\, \sigma_1\, X_1(t_i)\sqrt{\delta t}\quad,\quad i = 1, \ldots, n$$
$$\ln S_2(t_i) = \ln S_2(t_{i-1}) + \mu_2\, \delta t\, +\, \sigma_2\left[\rho X_1(t_i) + \sqrt{1-\rho^2}X_2(t_i)\right]\sqrt{\delta t}\,.$$
$$\qquad (11.6)$$

Or, if only the values at the end of the time period under consideration are of interest but not the paths of the risk factors:

$$\ln S_1(T) = \ln S_1(t) + \mu_1\,(T-t)\, +\, \sigma_1\sqrt{T-t}X_1$$
$$\ln S_2(T) = \ln S_2(t) + \mu_2\,(T-t)\, +\, \sigma_2\sqrt{T-t}\left[\rho X_1 + \sqrt{1-\rho^2}X_2\right]\,.$$
$$\qquad (11.7)$$

The Excel workbook from the download site [50] MONTECARLODEMO.XLSX includes a demonstration of the Monte Carlo simulation of a risk factor. For demonstration purposes, the simulation is accomplished using Excel cell functions without programming in Visual-Basic. Of course, this would be too slow for use in real applications. We therefore also provide a workbook entitled MONTECARLOSIMULATION.XLSM which contains an executable Visual Basic module for Monte Carlo simulations.

11.4 Pricing

According to Eq. 9.20 the value at time t of a financial instrument can be determined from the expectation of its price at a future time T, if this expectation is taken with respect to the risk-neutral probability distribution of the underlying. A clever choice of T can make pricing the instrument using Monte Carlo simulations quite easy. If T is chosen to be the maturity of the derivative, then the derivative's price at time T is simply given by the payoff profile P. Thus, if the underlying S is simulated according to Eq. 11.1 up to time T we can easily obtain a *simulated* probability distribution for the payoff values. This works also for path dependent instruments if we measure the relevant path-dependent quantities along the way (like for instance average of the simulated underlying values for specified days in the case of Asian options). The *mean* of all these simulated payoff values is then used as an estimator for the *expectation* of the payoff. According to Eq. 9.20, discounting[3] this estimator back to time t yields an estimator for the value of the instrument at time t, *if* the calculation of the mean has been carried out with respect to the risk-neutral probability. To ensure that this is the case, the risk factor must be *simulated* with respect to this probability. This is accomplished by simply choosing the *drift* of the random walk to be risk-neutral in accordance with Eq. 9.25.

Thus, the approach for pricing derivatives using Monte Carlo simulations is basically clear: We simulate the underlying (or more precisely the logarithm of the underlying value) in a risk-neutral world up to the time of maturity in accordance with Eq. 11.1 (for path-*in*dependent derivatives we can even use Eq. 11.2 and save a lot of computing time), and then measure the (discounted) mean of the payoff profile. In order to determine the error involved as discussed

---

[3]For *future styled* instruments, whose value changes are settled by daily adjustments in a margin account (as is the case for futures, for example), today's price is directly related to this expectation without any discounting, see Eqs. 6.6 and 9.21.

in Sect. 31.2, the mean of the square of the payoff profile must be measured, too. That's all. The calculation of option prices using the Monte Carlo method merely involves determining the mean of certain functions during the simulation.

An essential point, however, should not be overlooked. The time of exercise must be known, otherwise it is not clear "up to when" the price is to be simulated. For American or Bermudan Options we do not have this information up front. Section 11.5 describes methods for calculating at least lower bounds for the option values, which could be used to approximate the exact solution quite accurately.

A detailed demonstration of the application of the Monte Carlo method to the valuation of an option portfolio is provided in the Excel workbook MONTECARLOSIMULATION.xlsx from the download site [50]. By making appropriate adjustments in the valuation part of the Visual Basic module, this program can be used to price all sorts of European derivatives[4] in the context of the Black-Scholes world (constant yields and volatility). In anticipation of Chap. 12, the calculation of *Greeks*, i.e., the sensitivities of the derivative's price with respect to its parameters, is also demonstrated in the workbook. The workbook can be used as a small but complete option calculator (as always, the yellow fields are the input fields).

11.5 American Monte Carlo

The Monte Carlo methods can be extended for the valuation for options with a Bermudan or American exercise profile. As first step, American exercise can be approximated by Bermudan exercise. The higher the frequency of Bermudan exercise days (i.e. daily), the better the continuous exercise right will be approximated. For numerical computations, this discretization of continuous time to a set of distinct time points is anyway required. Therefore, it is only a matter of required accuracy and available computer time, how small the time steps must resp. should be chosen. Thus, in the following, we only consider Bermudan exercise rights.

The second problem is more difficult to solve. At each potential exercise day t_e we need to decide, whether the option should be exercised or not. Besides the price of the underlying (directly simulated by the Monte Carlo method) used to calculate the pay off in case of exercise (exercise value), we need to

---

[4]See also Sect. 19.3.1.

compute the price of the option to calculate the value of exercise at some later point of time (hold value). A simple approach would start for simulated path at each exercise date a new Monte Carlo simulation with many paths to simulate the future price of the option. With this approach, computing time increases dramatically with the number of exercise dates of the option and the number of simulated paths. Similar to non-recombining tree, this method is usually not feasible. Therefore, various authors have suggested alternative methods, of which the most widely known (and used) is probably the work of Longstaff ans Schwartz [135]. This method is also known as the Longstaff Schwartz method. Another, alternative approach, will be presented in Sect. 14.10.

The general idea of American Monte Carlo method is to calculate the optimal exercise time based on key figures well known at the time of exercise. This key figures should have the best explanatory power for the considered problem. For a Plain Vanilla American option, this could be the relative moneyness

$$x = \frac{K - S(t_e)}{S(t_e)} \ .$$

Here, $S(t_e)$ is the simulated underlying price at the (potential) exercise time. K is, as usual, the strike of the option. Then, the hold value can be modeled as a function $h(x)$. E.g., a simple approach would be a 2nd degree polynomial.

$$h(x) = a_2 x^2 + a_1 x + a_0 \ .$$

In the beginning, the 3 parameters a_0, a_1, a_2 are unknown and need to be determined. This is down in a simulation run preceding the final simulation for calculating the full option. In this pre-run, only the underlying price will be simulated. For each path $i = 1, 2, \ldots, n$ and each possible exercise date t_j with $t > t_j < T$ and $j = 1, 2, \ldots, m$, the simulated underlying price $S_{i,j}$ is computed and stored. Then, we will go backwards in time through all paths. At the final exercise date t_m, the hold value of the option can be computed (it will be either exercised at this date or expires worthless) as pay off at time T discounted from T to t_m. For discounting, the numeraire simulated for each path can be used. For each of the n paths, the hold value H_m is calculated separately. Then, using an optimization algorithm, the 3 parameters a_0, a_1, a_2 will be determined such that the n values H_m can be approximated by the function $h(x)$ as close as possible. By approximating H_m by $h(x)$, we describe the hold value by means of parameter x which is well known at time t_m without

using the information of the future (unknown) underlying price at time T we needed to calculate H_m.

Next, we can go one step back to date t_{m-1} and calculate the hold values H_{m-1} for each path. Now, we need take into account the possibility of exercise at time t_m. This would be exactly the case if the inner value $c(t_m)$ of the option at date t_m is worth more than the hold value, since the option buyer seeks to maximize the profit. Therefore, for each path, the hold value H_{m-1} is calculated as discounted value of $\max[c(t_m), H_m]$. After the hold values have been determined, we use again an optimizer over all paths to determine another set of three parameters a_0, a_1, a_2 to minimize the differences between $h(x)$ and the H_{m-1}. This procedure is repeated until for each exercise date an optimal set of parameters has been found. Finally, we have n functions $h_i(x)$ determined, each function modeling the hold value of the option at one exercise date.

At last, the simulation to calculate the full Bermudan option could be exercised. For this, we simulate again many Monte Carlo paths (in general more than has been used for optimizing the parameters for approximating the hold value, e.g. $10n$ paths). As with the "normal" Monte Carlo method, we will go forward in time. At each exercise date t_j we check, whether $c(t_j) > h_j(x)$ is fulfilled. If so, the option will be exercised. If not, we continue the simulation to the next exercise date, until either the option is exercised or matured. Then, the price can be calculated the sum of all cash flows per path discounted with the numeraire averaged over all simulated paths. To avoid a numerical bias, it is important to make sure that the random numbers used in the pre-run and those in the second (full) run are independent.

In this way, it is possible to compute a lower bound to the actual value for the Bermudan option. The better the choice of function $h(\cdot)$, the better the lower bound will approximate the real value. This can be seen very easily: In the end, $h(\cdot)$ serves as a mathematically fixed exercise strategy and is only dependent on parameters well known at the exercise date. If we have made a bad choice for $h(\cdot)$, our exercise decision will be suboptimal and we would not be able to receive the full value of the option.

A drawback of this approach is the fact that a good choice for $h(\cdot)$ is often not easily found. The choice of moneyness as explanatory variable in a second order polynomial is probably not the best choice, if the well-known functional form of the intrinsic value of a Amerikan Plain Vanilla option is envisaged. The weighted sum of European Plain Vanilla options for each exercise date might be a better choice. Unless there is no benchmark price, a polynomial is a good starting point. Since we know that we will get only a lower limit, it is useful to experiment with various approaches for $h(\cdot)$ and finally chose that function, for which the simulated American options had maximum value.

12

Hedging

The replication of derivatives with a portfolio consisting of underlyings and a bank account as, for example, in Eq. 9.3, can be used to *hedge* the derivative's risk resulting from the stochastic movement of its underlying (or conversely a derivative could be used to hedge such a portfolio). This is accomplished by going short in the portfolio and long in the derivative or vice versa. This idea can be extended to hedging against influences other than the underlying price, for example, changes in the volatility, interest rate, etc. Such concepts of safeguarding against a risk factor have already made their appearance in arbitrage arguments in previous chapters and will be presented in their general form in this chapter. In addition to the fundamental Assumptions 1, 2, 3, 4 and 5 from Chap. 4, continuous trading will also be assumed below, i.e., Assumption 6. We will allow the underlying to perform a general Ito process[1] of the Form 2.19 and assume that it pays a dividend yield q.

12.1 Hedging Derivatives with Spot Transactions

Consider first the change in the value V of a derivative resulting from the change in the price S of its underlying. V, as a function of the stochastic process S, is also a stochastic process. From Ito's lemma in the form of Eq. 2.22

---

[1]Thus, we are a more general here than Assumption 7 from Chap. 4.

this process is explicitly

$$dV(S, t) = \left[\frac{\partial V}{\partial t} + \frac{1}{2} b^2(S, t) \frac{\partial^2 V}{\partial S^2} \right] dt + \frac{\partial V}{\partial S} dS(t) . \qquad (12.1)$$

In a similar manner as in Sect. 7.1.1, we will now construct a portfolio which replicates (hedges) the derivative in Eq. 12.1. This portfolio will consist of a certain number Δ of underlyings and an amount g of money borrowed from or invested in the capital market. This portfolio is constructed to have the same value as the derivative,

$$V = \Delta S + g . \qquad (12.2)$$

To ensure that the derivative is perfectly hedged for at least a short time span dt, it is required that the change in its value over dt to be the same as that of the derivative. The change in value of the derivative is given by Eq. 12.1. The value of the portfolio consisting of money and the underlying changes as already described in Eq. 9.3, for example. For small time intervals dt, B and $1/B$ can be approximates by $(1 - r\, dt)$ and $(1 + r\, dt)$, respectively, for all compounding conventions. The same approximation holds for B_q. Denoting the change in S by dS, the change in the value of the portfolio can be expressed by

$$d(\Delta S + g) = \underbrace{\Delta(S + dS)(1 + q\, dt) + g(1 + r\, dt)}_{\text{new value}} - \underbrace{(\Delta S + g)}_{\text{old value}}$$

$$= \Delta\, dS + (\Delta q\, S + r\, g) dt + \Delta q\, dS\, dt$$

$$= \Delta\, dS + (\Delta q\, S + r\, g) dt + \cdots \qquad (12.3)$$

where in the last line only the first order terms in dS and dt are considered. To ensure that this change corresponds exactly to a change in the value of the derivative, the coefficients of dt and dS must be the same as those appearing in Eq. 12.1. Equating the coefficients of dS yields the well known result that the number of underlyings required for the hedge, called the *hedge ratio*, must be equal to the sensitivity of the derivative with respect to its underlying. Combined with Eq. 12.2 we can determine the amount of money in the bank account required for the hedge as well:

$$\Delta(t) = \frac{\partial V}{\partial S(t)} \implies g(t) = V - S(t) \frac{\partial V}{\partial S(t)} . \qquad (12.4)$$

Thus, with information on the value V_s and its first partial derivative with respect to the underlying, it is possible to construct a portfolio hedging the derivative *perfectly* over the next time step. The partial derivatives of the price function are often referred to as the *sensitivities* (see Sect. 12.4).

We have yet to exploit the second condition, which is the equality of the coefficients of dt in Eqs. 12.1 and 12.3. This yields

$$\Delta S(t)q + rg = \frac{\partial V}{\partial t} + \frac{1}{2}b^2(S,t)\frac{\partial^2 V}{\partial S^2} \ .$$

Inserting Eq. 12.4 for Δ_S and g, we recognize the famous differential equation of Black and Scholes for derivatives on the spot price S on an underlying:

$$\frac{\partial V}{\partial t} + (r-q)S\frac{\partial V}{\partial S} + \frac{1}{2}b^2(S,t)\frac{\partial^2 V}{\partial S^2} = r\,V \ . \tag{12.5}$$

If we write (without loss of generality) the parameters a and b as in Eq. 7.1 we immediately see that Eq. 12.5 is exactly equal to Eq. 7.8.

As all other derivation of the Black-Scholes PDE, this most simple derivation of the *Black-Scholes differential equation*, does not specify the derivative under consideration in any way. No information about the payoff profile or any other property characterizing the derivative is necessary. Only the properties of the *underlying* are used, these being that the underlying earns a dividend yield q and that its price is governed by a stochastic process of the form 2.19. Thus, there is only *one* equation for the construction of the hedging portfolio and *one* differential equation for the price of *all* derivatives on this underlying. The different derivative instruments can be distinguished from one another solely on the basis of their respective initial and/or boundary conditions (for example, the payoff profile of an option at maturity or the fact that a forward contract is worth nothing at the time when the contract is entered into).

In effect, the value of the option is equivalent to the total cost of the hedge, i.e. the replication of the pay off profile by the carefully chosen portfolio (the replicating portfolio). The whole theory and practice of arbitrage free valuation of derivatives is based on this fundamental assumption that the value of an option is equivalent to the total cost to hedge it.

12.1.1   Hedging of Forwards and Futures

Forward and futures contracts belong to the most simple derivatives. Therefore, they will serve as an example how to apply Eqs. 12.4 and 12.5 to hedge derivatives with spot trades. We adopt here the commonly used convention of continuous compounding.

Consider first a forward contract. The value of a forward with delivery price K, assuming continuous compounding, is given by Eq. 6.5 as

$$f(t, T, K) = [S(t, T) - K] B(t, T) = e^{-q(T-t)} S(t) - e^{-r(T-t)} K .$$

The partial derivatives of this function appearing in Eqs. 12.4 and 12.5 are

$$\frac{\partial f}{\partial t} = q\, e^{-q(T-t)} S - r e^{-r(T-t)} K , \quad \frac{\partial f}{\partial S} = e^{-q(T-t)} , \quad \frac{\partial^2 f}{\partial S^2} = 0 .$$

Thus, Eq. 12.4 yields

$$\Delta(t) = e^{-q(T-t)} , \quad g(t) = -e^{-r(T-t)} K .$$

In order to hedge a (short) forward, $e^{-q(T-t)}$ underlyings must be purchased and an amount $e^{-r(T-t)} K$ of cash must be raised. Substituting the value of the forward and its derivative into Eq. 12.5 shows that the Black-Scholes differential equation is satisfied:

$$q e^{-q(T-t)} S - r e^{-r(T-t)} K + (r - q) S e^{-q(T-t)} = r e^{-q(T-t)} S - r e^{-r(T-t)} K .$$

Now consider a futures contract. From a valuation perspective, the essential difference compared to a forward contract is the obligatory maintenance of a margin account. As a consequence, every price change of the futures contract is reflected by an equivalent change of the margin account balance. The margin account is similar to a bank account. The buyer or seller of the futures contract receives interest on the margin account, but has no direct access to the money in the account. At the trade day of the futures contract, the buyer/seller must transfer an initial margin on the account. If the futures price raises or falls later, the margin account balance raises or falls by the same amount. If the margin account balance falls below a certain level, fresh money must be transferred to the account, otherwise the whole position will be closed out. On the other hand, if the margin account balance raise beyond a certain level, money could be transferred back if the balance exceeds certain trigger levels. This is called

variation margin. Neglecting default risk, it can be assumed that the money for initial and variation margin can be borrowed for the risk-free interest rate. Also, the margin account itself is assumed to pay interest based on the risk-free interest.[2] We will denote the total cash amount (which can be negative or positive) as $g_F(t)$.

Let's consider the evolution of the total value of the futures position over time. This development is completely reflected in the development of the function $g_F(t)$. Contrary to a forward contract, a futures contract does not define a fixed strike price to be paid at maturity in exchange of the underlying. Instead, the buyer of the futures contract will have to pay at maturity the then current spot price, while the difference of forward and future spot price at trade date is at maturity identical to the balance of the variation margin (plus accumulated interest). We need the following derivatives of $g_F(t)$:

$$\frac{\partial g_F}{\partial t} = -(r-q)e^{(r-q)(T-t)}S + rg_F \ , \quad \frac{\partial g_F}{\partial S} = e^{(r-q)(T-t)} \ , \quad \frac{\partial^2 g_F}{\partial S^2} = 0 \ .$$

Inserting these in Eq. 12.4 yields

$$\Delta(t) = e^{(r-q)(T-t)} \ , \quad g(t) = g_F(t) \ .$$

Therefore, hedging a (short) futures contract requires to buy $e^{(r-q)(T-t)}$ underlyings and to borrow an amount of $-g(t) = g_F(T)$ currency units. Inserting the value of the Portfolio $g_F(t)$ into Eq. 12.5 shows that $g_F(t)$ fulfills the Black-Scholes differential equation:

$$-(r-q)e^{(r-q)(T-t)}S + rg_F + (r-q)e^{(r-q)(T-t)}S = rg_F \ .$$

The mindful reader may wonder, in which way $g_F(t)$ differs from the fair value $F(S, t, T, K)$ of the futures contract, which itself does not fulfill the Black-Scholes equation. According to Eq. 6.6 the fair value of a futures contract is

$$F(S, t, T, K) = e^{(r-q)(T-t)}S(t) - K \ ,$$

---

[2]Reality, as often, is more complex. E.g. the actual margin interest rates may differ from the risk-free interest rate or maybe floored at zero, while the risk-free interest rate might be allowed to become negative.

where K is the futures price at trade date. In contrast to $g_F(t)$ the fair value does not contain the accumulated interest, since they have already been realized.

12.2 Hedging Derivatives with Forward Contracts

The portfolio for the synthetic derivative is now to be constructed using *forward* contracts instead of spot transactions. We cannot merely replace the spot price with the forward price in Eq. 12.3 since the forward price itself is not tradable. Forward and futures contracts on the other hand are tradable.

12.2.1 Hedging with Forwards

Since futures satisfy the Black Scholes differential equation only upon conclusion of the contract, we will avoid this difficulty at this point and begin by using forwards to construct the synthetic derivative. The hedging portfolio is then

$$\Delta_f(t) f(t, T', K) + g(t) = \Delta_f(t) \left[S(t, T') - K \right] B(t, T') + g(t) \text{ with } T' \geq T > t ,$$

where the subscript f indicates that the hedge is accomplished with forwards. T denotes the maturity of the derivative and T' the maturity of the forward contract. Note that the condition $T' \geq T$ is not strictly necessary for our considerations, but it is convenient since it ensures that the hedge does not "vanish" before the maturity T of the derivative. If $T' < T$ it would be required to roll over the hedge into another forward contract (with a later maturity) at time T' or earlier.

In the time span dt this portfolio changes its value as a result of changes in the forward price and the discount factor as follows:

$$
\begin{aligned}
d(\Delta_f f + g) &= \underbrace{\Delta_f \left[S(t, T') + dS(t, T') - K \right] B(t, T')(1 + r dt) + g(t)(1 + r dt)}_{\text{new value}} \\
&\quad - \underbrace{\Delta_f \left[S(t, T') - K \right] B(t, T') + g(t)}_{\text{old value}} \\
&= \Delta_f B(t, T') dS(t, T') + \Delta_f \left[S(t, T') - K \right] B(t, T') r \, dt + g r \, dt + \cdots \\
&= \Delta_f B(t, T') \frac{\partial S(t, T')}{\partial S(t)} dS(t) + \Delta_f(t) f(t, T', K) r \, dt + g r \, dt + \cdots ,
\end{aligned}
$$

$$(12.6)$$

where again we consider only changes which are of linear order in dt or dS. In the last equation, the change in the forward price $dS(t, T')$ was rewritten in terms of a change in the spot price to facilitate the comparison of the coefficients with the associated coefficients in Eq. 12.1 for the change in the price of the derivative. Equating the coefficients of dS in the above mentioned expressions now yields:

$$\Delta_f(t) = \frac{1}{B(t, T')} \frac{\partial V / \partial S(t)}{\partial S(t, T') / \partial S(t)} = \frac{1}{B(t, T')} \frac{\partial V}{\partial S(t, T')} .$$

Using Eq. 6.1 for the dividend yield, we can calculate the derivative of the forward price with respect to the spot price and thus obtain the following expression for the hedge ratio:

$$\Delta_f(t) = \frac{1}{B_q(t, T)} \frac{\partial V}{\partial S(t)} = \frac{1}{B(t, T')} \frac{\partial V}{\partial S(t, T')} . \tag{12.7}$$

In other words, the number of forwards required to hedge the derivative is equal to the sensitivity of the derivative with respect to the spot price compounded at the dividend yield up to maturity of the derivative, or equivalently, is equal to the sensitivity with respect to the forward price compounded at the risk-free rate up to maturity of the hedging instrument (the forward).

Since by construction, the value of the replicating portfolio is equal to the value of the derivative we can determine the amount of cash needed for the hedge:

$$g(t) = V - f(S, t, T', K) \, \Delta_f(t) . \tag{12.8}$$

We have thus completely determined the hedging portfolio using solely information derived from the price of the derivative and its sensitivities.

Before we establish a mathematical expression obtained from comparison of the coefficients of dt in Eqs. 12.1 and 12.6, we change the coordinates in Eq. 12.1 to transform the second derivative of the price with respect to *spot* price $S(t)$ into a derivative with respect to the *forward* price $S(t, T')$. This is accomplished using 6.1.

$$\frac{\partial}{\partial S(t)} = \frac{\partial S(t, T')}{\partial S(t)} \frac{\partial}{\partial S(t, T')} = \frac{B_q(t, T')}{B(t, T')} \frac{\partial}{\partial S(t, T')}$$

$$\frac{\partial^2}{\partial S^2(t)} = \frac{\partial}{\partial S(t)} \frac{B_q(t, T')}{B(t, T')} \frac{\partial}{\partial S(t, T')} = \frac{B_q^2(t, T')}{B^2(t, T')} \frac{\partial^2}{\partial S^2(t, T')} . \tag{12.9}$$

Equating now the coefficient of dt and using Eq. 12.8 for g yields

$$\frac{\partial V}{\partial t} + \frac{1}{2} b^2(S, t) \frac{B_q^2(t, T')}{B^2(t, T')} \frac{\partial^2 V}{\partial S^2(t, T')} = rV \; .$$

This is the Black-Scholes differential equation for derivatives on the forward price of an underlying.

   To put this in a more familiar form we rewrite (without loss of generality) the parameter b as in Eq. 7.1 and use Eq. 6.1 to introduce the forward price into the parameter b, i.e.,

$$b(S, t) =: \sigma(S, t) S(t) = \sigma(S, t) \frac{B(t, T')}{B_q(t, T')} S(t, T') \; . \tag{12.10}$$

With this substitution we arrive at the well-known form of the Black-Scholes differential equation for derivative on the forward price of an underlying:

$$\frac{\partial V}{\partial t} + \frac{\sigma^2}{2} S^2(t, T') \frac{\partial^2 V}{\partial S^2(t, T')} = r V \; . \tag{12.11}$$

   As in Sect. 12.1, only *one* equation is needed for the construction of the hedging portfolio with forward contracts, and there is only *one* differential equation for *all* derivatives on the forward price. Different derivatives can be distinguished from one another only through their respective initial and/or boundary conditions.

   Note the absence of a first derivative term with respect to the forward price. Equation 12.5 (which is the corresponding Black-Scholes PDE for derivatives on the *spot* price of an underlying) would look quite similar to Eq. 12.11 for an underlying which earns a dividend yield q being exactly equal to the risk free rate r. In fact, both equations would look identical if this underlying would have $S(t, T')$ as its *spot* price. Herein lies the deeper explanation for the "recipe" for pricing options on futures as given in Eq. 8.8: a derivative on the forward price can be priced as a derivative on the spot price of an (artificial) underlying whose spot price is equal to $S(t, T')$ and whose dividend yield equals the risk-free rate.

12.2.2 Hedging with Futures

If futures are used to construct a synthetic derivative, we have to deal with the additional complexity that the margin account balance depends on when and how many futures contracts have been bought or sold. Since we assume that the margin account is compounded with the same interest rate as the risk-free cash amount borrowed or lend $g(t)$, we consider the margin account as part of the total cash balance or bank account. Then, we need to consider the futures contract's price (which equals the forward price) changes, only. As hedging instrument, we use a futures contract with fixed maturity T.[3] Since the futures price equals the forward price $S(t, T')$, we get

$$\Delta_F(t) S(t, T') + g(t) \text{ with } T' \geq T > t .$$

Changes in the futures contract price and the bank account in the time interval dt cause the value of the portfolio to change as follows

$$d(\Delta_F S(t, T') + g) = \underbrace{\Delta_F \left[S(t, T') + dS(t, T') \right] + g (1 + r dt)}_{\text{new value}} - \underbrace{\Delta_F \left[S(t, T') \right] + g}_{\text{old value}}$$

$$= \Delta_F \, dS(t, T') + gr \, dt + \cdots$$

$$= \Delta_F \frac{\partial S(t, T')}{\partial S(t)} dS(t) + gr \, dt + \cdots , \qquad (12.12)$$

where again we consider only the parameter changes of linear order in dt or dS, neglecting all higher order terms. In the last equation, the change in the futures price has been replaced by the corresponding expression with respect to a change in the spot price, facilitating a direct comparison with the coefficients in Eq. 12.1 for the change in the price of the derivative. Setting the coefficients of dS in both expressions equal yields

$$\Delta_F(t) = \frac{\partial V / \partial S(t)}{\partial S(t, T') / \partial S(t)} = \frac{\partial V}{\partial S(t, T')} .$$

---

[3]Here again the condition $T' \geq T$ is not strictly necessary but convenient since it ensures that the hedge does not "vanish" before maturity T of the derivative. Again, $T' < T$ would require to roll over the hedge into another futures contract (with a later maturity) at time T' or earlier.

Equation 6.1 for the forward price of an underlying earning a dividend yield q gives the partial derivative of V_S with respect to the spot price.

$$\Delta_F(t) = \frac{B(t, T')}{B_q(t, T')} \frac{\partial V}{\partial S(t)} = \frac{\partial V}{\partial S(t, T')} \ . \tag{12.13}$$

In other words, the number of futures required to hedge a derivative is equal to the sensitivity of the derivative with respect to the *spot* price compounded at the rate obtained by using the risk-free rate reduced by the dividend yield. Or equivalently, the number of futures needed equals the sensitivity with respect to the *forward* price.

By definition, the value of the hedging portfolio must be equal to the value of the derivative. It follows that the amount of money needed for the hedge is given by

$$g(t) = V - F_S(t, T', K)\Delta_F(t) \ . \tag{12.14}$$

We have thus completely determined the hedging portfolio using solely the information derived from the value of the derivative instrument and its sensitivities.

12.3 Hedge-Ratios for Arbitrary Combinations of Financial Instruments

An analogous approach to those described above can be taken to derive a formula for the number of an *arbitrary* hedge instrument required to replicate an *arbitrary* derivative. The number of hedging instruments is always equal to the sensitivity of the derivative with respect to this hedging instrument. Intuitively, if the value h of a hedge instrument changes by the amount dh, then the value V of the derivative changes by the amount dV. The quotient dV/dh is precisely the number of hedging instruments needed to compensate for the change dV in the derivative. For very small changes, this quotient approaches the differential quotient. Any instrument can assume the role of either the derivative or the hedging instrument. Since the value obtained by differentiating a derivative with respect to an arbitrary hedging instrument cannot, in general, be calculated directly, a common reference variable, usually the spot price $S(t)$, is introduced with respect to which both the partial derivative of the value as well as the partial derivative of the hedging instrument

Table 12.1 Sensitivities for options and forward contracts with respect to the spot price $S(t)$. x and x' are defined as in Eqs. 8.7 and 8.8

Underlying (Spot)	1	
Futures	$B_q(t,T)B(t,T)^{-1}$	
Forward	$B_q(t,T)$	
	Call	Put
Option on spot price	$B_q(t,T)\,\mathsf{N}(x)$	$-B_q(t,T)\,\mathsf{N}(-x)$
Option on futures	$B_q(t,T')\frac{B(t,T)}{B(t,T')}\mathsf{N}(x')$	$-B_q(t,T')\frac{B(t,T)}{B(t,T')}\mathsf{N}(-x')$
Option on forward	$B_q(t,T')\mathsf{N}(x')$	$-B_q(t,T')\mathsf{N}(-x')$

is known. The equation for this general *hedge ratio* is therefore

$$\Delta_{\substack{\text{Derivate } V \text{ with}\\ \text{Hedging Instrument } h}} = \frac{\partial V}{\partial h} = \frac{\partial V}{\partial S}\left[\frac{\partial h}{\partial S}\right]^{-1}. \tag{12.15}$$

Consequently, the sensitivities of all financial instruments with respect to the spot price must be available if the hedge ratio is to be calculated for an *arbitrary* combination of these instruments. In general, though, it is not recommended to use hedging instruments which introduce additional risk factors the original trade is not sensitive to (e.g. additional foreign exchange risk). The sensitivities already dealt with in the last section, i.e., the sensitivities of the option, forward and spot transactions with respect to the spot price are listed in Table 12.1 (for the Black-Scholes world). For forward contracts, they follow from Eqs. 6.1, 6.5 and 6.6, which were derived using arbitrage arguments. For options, the sensitivities follow directly from the Black-Scholes equations 8.7, 8.8 and 8.9.

Readers wishing to reproduce these results themselves are advised to use the following property[4] of $\mathsf{N}'(x-\sigma\sqrt{T-t})$:

$$\mathsf{N}'(x-\sigma\sqrt{T-t}) = \frac{B_q(t,T)\,S(t)}{B(t,T)\,K}\mathsf{N}'(x). \tag{12.16}$$

---

[4]To prove Eq. 12.16 we first write

$$\mathsf{N}'(x-\sigma\sqrt{T-t}) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\sigma\sqrt{T-t})^2} = \underbrace{\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}}_{\mathsf{N}'(x)}e^{x\sigma\sqrt{T-t}-\frac{1}{2}\sigma^2(T-t)}$$

The relation is now obtained by substituting the definition of x, Eq. 8.5, into the second exponential term.

Here, N′ denotes the derivative of the cumulative standard normal distribution with respect to its argument:

$$N'(y) \equiv \frac{d}{dy} N(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}$$

The entries in Table 12.1 together with Eq. 12.1 provide all the information needed to determine the hedge ratio for every conceivable combination of these financial instruments in the Black-Scholes world. We now present three examples to illustrate this.

a) Hedging a put on the spot price with futures:

$$\Delta_{\substack{\text{Put on spot} \\ \text{with futures}}} = \frac{\partial V}{\partial S} \left[\frac{\partial h}{\partial S} \right]^{-1} = -B_q(t, T) \, N(-x) \left[\frac{B_q(t, T')}{B(t, T')} \right]^{-1}$$

$$= -B(t, T') \frac{B_q(t, T)}{B_q(t, T')} N(-x) \ .$$

A long put on the spot can thus be *replicated* by a *short* position consisting of

$$B(t, T')N(-x)/B_q(T, T'|t)$$

futures. Or interpreted as a hedge, a long put can be *hedged* by a *long* position in this number of futures.

b) Hedging a put on a future with forward contracts:

$$\Delta_{\substack{\text{Put on futures} \\ \text{with forward}}} = \frac{\partial V}{\partial S} \left[\frac{\partial h}{\partial S} \right]^{-1} = -B_q(t, T') \frac{B(t, T)}{B(t, T')} N(-x') \left[B_q(t, T') \right]^{-1}$$

$$= -\frac{B(t, T)}{B(t, T')} N(-x') \ .$$

A long put on the future can be *replicated* with a *short* position of

$$N(-x')/B(T, T'|t)$$

forwards. Or interpreted as a hedge, a long put can be *hedged* by a *long* position in this number of forwards.

c) Hedging a call on the spot with puts on a future:

$$\Delta_{\substack{\text{Call on spot with} \\ \text{put on futures}}} = \frac{\partial V}{\partial S}\left[\frac{\partial h}{\partial S}\right]^{-1} = B_q(t, T)\, N(x)\left[-B_q(t, T')\frac{B(t, T)}{B(t, T')}N(-x')\right]^{-1}$$

$$= -\frac{B(t, T')}{B(t, T)}\frac{B_q(t, T)}{B_q(T, T')}\frac{N(x)}{N(-x')}\ .$$

A long call on the spot can thus be *replicated* by a *short* position in

$$\frac{B(T, T'\,|\,t)N(x)}{B_q(T, T'\,|\,t)N(-x')}$$

puts on the future. Or interpreted as a hedge, the long call can be *hedged* by a *long* position in this number of puts on the future.

12.4 "Greek" Risk Management with Sensitivities

12.4.1 Sensitivities and a Portfolio's Change in Value

The *sensitivity* of a financial instrument with respect to a parameter is defined as a variable that, when multiplied by a (small) parameter change, yields the change in the value of the instrument resulting from this parameter change.

> Price change = sensitivity × parameter change

The sensitivity is thus the derivative of the price with respect to the parameter. Since options depend on several such parameters such as the price of the underlying, the time to maturity, the volatility, the dividend yield of the underlying and the risk-free market rate, the sensitivities listed in Table 12.2 are of particular interest.

The first of these sensitivities, namely Δ, has already been used in Sect. 12.1 to construct a replicating portfolio. In Eq. 12.15 the hedge ratio Δ was defined in greater generality as the sensitivity of the derivative with respect to the price change of the hedging instrument. In the special case where the hedging instrument is the underlying itself, this general hedging ratio and the delta listed in Table 12.2 agree.

Table 12.2 Definitions of the "Greeks"

Symbol	Name	Definition	Interpretation
Δ	Delta	$\dfrac{\partial V}{\partial S}$	Price change resulting from change of underlying spot price
Γ	Gamma	$\dfrac{\partial^2 V}{\partial S^2}$	Change of delta resulting from change of underlying spot price
Ω	Omega	$\dfrac{S}{V}\dfrac{\partial V}{\partial S}$	Relative price change (in %) resulting from relative change (in %) of underlying spot price
Ψ	Vega	$\dfrac{\partial V}{\partial \sigma}$	Price change resulting from change of volatility of underlying
Θ	Theta	$\dfrac{\partial V}{\partial t}$	Price change resulting from change of time to maturity
ρ	Rho	$\dfrac{\partial V}{\partial r}$	Price change resulting from change of the risk-free rate
ρ_q	Rho$_q$	$\dfrac{\partial V}{\partial q}$	Price change resulting from change of dividend yield of underlying

Any method used to calculate the price of an option can be used to calculate the sensitivities, even if the explicit expressions for the sensitivities cannot be obtained directly. The price model must simply be applied twice: the first time to calculate the price with the current parameters and a second time for a valuation with one of the parameters slightly changed. The difference between these two option values divided by the parameter change yields an approximation for the sensitivity of the option with respect to that parameter. Approximating the second derivative (as for gamma, for example) requires the price to be calculated with respect to three different parameter values. In summary: as soon as a pricing method is available we can calculate sensitivities by approximating the differential quotients in Table 12.2 with difference quotients as follows:

$$\text{Delta} \approx \frac{V(S+\delta S) - V(S)}{\delta S}, \ \text{Gamma} \approx \frac{V(S+\delta S) - 2V(S) + V(S-\delta S)}{\delta S^2}$$

$$\text{Vega} \approx \frac{V(\sigma+\delta\sigma) - V(\sigma)}{\delta\sigma}, \quad \text{Theta} \approx \frac{V(t+\delta t) - V(t)}{\delta t}$$

$$\text{Rho}_r \approx \frac{V(r+\delta r) - V(r)}{\delta r}, \quad \text{Rho}_q \approx \frac{V(q+\delta q) - V(q)}{\delta q} \qquad (12.17)$$

The risk of an instrument (or a portfolio) can be measured by its reaction to changes in the parameters influencing its price. In other words, the more sensitive it is to parameter changes, the "riskier" the instrument is. For this reason, the sensitivities are also known as *risk ratios*. A commonly used method of risk management is to control these sensitivities, i.e., to set limits on or targets for their values, for instance. The possibilities and limitations of this kind of risk management become clear, when we consider the Taylor series representation for small parameter changes: The value V depends on all of the above listed parameters: $V = V(t, S, \sigma, q, r)$. A change in value is thus

$$dV = \frac{\partial V}{\partial t}dt + \frac{\partial V}{\partial S}dS + \frac{\partial V}{\partial \sigma}d\sigma + \frac{\partial V}{\partial r}dr + \frac{\partial V}{\partial q}dq + \frac{1}{2}\frac{\partial^2 V}{\partial S^2}dS^2 + \cdots$$

$$= \Theta\, dt + \Delta\, dS + \Psi\, d\sigma + \rho\, dr + \rho_q\, dq + \frac{1}{2}\Gamma\, dS^2 + \cdots$$

The coefficients of all linear parameter changes are given by sensitivities. A linear approximation holds only for small parameter changes. This is especially true for highly non-linear instruments. Because of the particularly strong influence of the underlying price, the effect of its quadratic change has been included in the above expression resulting in the appearance of the sensitivity Γ.

All risk variables in Table 12.2 (with the exception of the variable Ω) have the important property that the sensitivity of a portfolio with respect to a risk factor is equal to the sum of the sensitivities of its component instruments to this risk factor, i.e., the sensitivities are *linear*. For a portfolio consisting of n instruments of type A and m of type B we can therefore write

$$\Delta(n\, A + m\, B) = n\, \Delta(A) + m\, \Delta(B) \quad \text{likewise for } \Gamma, \Theta, \Psi, \rho, \rho_q \tag{12.18}$$

This property makes it possible to make a portfolio *delta-neutral*, for example: given n, $\Delta(A)$ and $\Delta(B)$, it is possible to make the portfolio's delta, $\Delta(nA + mB)$, equal to zero, by choosing $m = -n\Delta(A)/\Delta(B)$. This is called a *delta hedge*. The portfolio is then (in linear approximation) insensitive to small changes in the price of the underlying S. In the special case that the instrument B is the underlying itself we have $\Delta(B) = 1$. This is reflected in the result given in Sect. 12.1 that the number of underlyings needed to hedge a particular instrument is obtained by taking the derivative of the value of the instrument with respect to the underlying. We can now observe that the

hedge methods introduced in Sect. 12.1 were delta hedges and represent only in linear approximation a safeguard against the risk resulting from a change in the underlying. This is no limitation under the (idealized) assumption of continuous trading with zero transaction cost, though.

Likewise, a portfolio can be constructed which is gamma, theta, vega or rho neutral. In general, for Plain Vanilla options, Delta and Vega are hedged at least.[5] Even *several* sensitivities of the portfolio can be neutralized *simultaneously* but more than one hedging instrument is then required to do so. In general, we need at least as many *different* hedging instruments as the number of sensitivities to be neutralized. For example, if a portfolio consisting of n instruments of type A is to be delta and gamma hedged, we would require two hedging instruments B_1 and B_2. The condition that the delta and the gamma of the portfolio be equal to zero means that

$$\Delta(n\,A + m_1\,B_1 + m_2\,B_2) = n\,\Delta(A) + m_1\,\Delta(B_1) + m_2\,\Delta(B_2) = 0$$

$$\Gamma(n\,A + m_1\,B_1 + m_2\,B_2) = n\,\Gamma(A) + m_1\,\Gamma(B_1) + m_2\,\Gamma(B_2) = 0$$

Solving this system of equations yields the number of hedging instruments m_1 and m_2 for the delta and gamma hedged portfolio:

$$m_1 = -n\frac{\Gamma(B_2)\Delta(A) - \Delta(B_2)\Gamma(A)}{\Gamma(B_2)\Delta(B_1) - \Delta(B_2)\Gamma(B_1)} \ , \ \ m_2 = -n\frac{\Gamma(B_1)\Delta(A) - \Delta(B_1)\Gamma(A)}{\Gamma(B_1)\Delta(B_2) - \Delta(B_1)\Gamma(B_2)}$$

From the Black-Scholes differential equation 12.5 we can obtain a relation, which the different sensitivities of *all* financial instruments *must* satisfy. Simply replace the partial derivatives in Eq. 12.5 by the corresponding sensitivities to obtain

$$\Theta + (r - q)S\Delta + \frac{1}{2}b^2\Gamma = r\,V \ .$$

One consequence of this equation is, for example, the following statement: The change value of a delta and gamma neutral portfolio per time (i.e., Θ) is equal to its current value multiplied by the risk-free interest rate.

---

[5]This is common practice even in cases in which the option is priced by means of Black-Scholes or some other model that assumes deterministic volatility. For Hedging, it is assumed in contrast to the model assumptions that the underlying volatility might indeed change. The practice proofs that this improves hedge efficiency, leading to the conclusion that the model with deterministic volatility is insufficient. Such a hedge is also called *out of model hedge*.

Table 12.3 Examples of Omega and Delta

	strike = 70 EUR	strike = 90 EUR
Price	10.18 EUR	0.04 EUR
Delta	0.992	0.024
Omega	7.79	47.41

12.4.2 Omega and Beta

In contrast to the delta which gives the *absolute* price change of the derivative (in monetary units) resulting from an *absolute* change of one monetary unit in the underlying, *omega* is the *relative* price change of the derivative (in %) resulting from a *relative* change in the underlying[6]:

$$\Omega \equiv \frac{[V(S + dS) - V(S)] / |V(S)|}{dS/S} = \frac{S}{|V(S)|}\Delta$$

In situations where a trader is instructed to invest a specific *amount* of money rather than in a certain *number* of instruments, this allows a significantly better assessment of the risk than the (much more prominent) delta. A simple example will serve to clarify this point.

Assume that 10,000 euros are to be invested either in a call with a strike price of 90 euros on an underlying paying no dividend with a spot price of 80 euros or in a call on the same underlying with strike price of 70 euros. Suppose the annualized volatility of this underlying is 20%. Both options mature in 30 days. At a risk-free rate of 3% per annum, Eq. 8.7 (implemented in the Excel workbook BLACKSCHOLESMODEL.xls on the accompanying website [50]) gives the prices and risk ratios shown in Table 12.3.

The sensitivity delta is almost 50 times greater for the call with a strike price of 70 euros than for the call with the strike price of 90 euros. Is the call with a strike of 70 euros riskier? If we have the choice of buying a certain *number* of the strike-70 calls or the same number of strike-90 calls then yes, since we could lose a lot more with a certain number of the strike-70 calls than with the same number of strike-90 calls. But this is solely because the *price* of the strike-70 call is so much higher than the price of the strike-90 call! The higher risk is entirely due to the significantly larger investment: if we invest more euros we can lose more euros.

---

[6]The absolute value of $V(S)$ in this definition has the effect that omega always has the same sign as delta, even for a position with a negative present value.

But if the *same amount* of money is to be invested, a portfolio with strike-90 calls (250,250 of these calls can be purchased for 10,000 euros) is significantly riskier than a portfolio with strike-70 calls (only 982 of these calls can be purchased for 10,000 euros). This risk is impressively quantified by omega, also called the *elasticity*. For the strike-70 call portfolio, a 1.60 euros (=2%) decline in the underlying's price results in a loss of value of 7.79 * 2 %, or 1,558 euros. However, the strike-90 call portfolio loses 47.41*2% of its value, a sum of 9,482 euros. Investing in the strike-90 calls, a relatively small change in the price of the underlying (2%) almost wipes out the entire investment. On the other hand, a slight increase in the price of the underlying can just as easily double the value of the strike-90 call portfolio. Thus, the volatility of the option equals omega times the volatility of the underlying

$$\sigma_{\text{Option}} = \Omega\, \sigma_{\text{Underlying}} \tag{12.19}$$

A similar relation can also be shown to hold for the expected return of an option (in the real world). The market requires, as a compensation for the risk of an investment, a higher expected return of the security under consideration compared to the risk-free rate. Since the risk (as represented by the volatility) of an option on that security is omega times as high as the risk of the security itself, the compensation for taking the risk of investing in the option must be omega times as high as well:

$$\mu_{\text{Option}} - r = \Omega\, (\mu_{\text{Underlying}} - r) \tag{12.20}$$

We can now establish a connection with the *CAPM* (*capital asset pricing model*). The variable β (*beta*) in the CAPM relates the risk premium of a security to the risk premium of a portfolio representing the entire market (with the corresponding diversification). *Indices* such as the *DAX* or the *Dow Jones* are constructed to represent such entire markets. To be more specific: the beta of a security with a price S with respect to an index with a price I is defined as the covariance of the price with the index divided by the variance of the index (see Eq. A.20 in Appendix A):

$$\beta_{SI} = \frac{\text{cov}\,[S,\,I]}{\text{var}\,[I]} = \frac{\sigma_S}{\sigma_I}\rho_{SI} \tag{12.21}$$

This variable can be interpreted as follows: should the value I of the index increase by 1%, then the price S of the security will (on average) increase by beta %. The relative change of S is thus beta times the relative change in I.

$$\beta_{SI} = \frac{\text{relative change in } S \text{ resulting from a change in } I}{\text{relative change in } I} \qquad (12.22)$$

$$= \frac{(dS)/S}{(dI)/I} = \frac{I}{S}\frac{dS}{dI}$$

This corresponds to the definition of omega if S is interpreted as a financial instrument having the index I as an underlying. Within the capital asset pricing model the risk premium for investing in S is shown to be

$$\mu_S - r = \beta_{S,I}(\mu_I - r) \qquad (12.23)$$

Substituting this for the risk premium of the underlying S in Eq. 12.20 gives

$$\mu_{\text{Option}} - r = \Omega\beta_{S,I}(\mu_I - r)$$

Thus, the risk premium for investing in the option is omega times beta multiplied by the risk premium for investing in the entire market. We could therefore also use Eq. 12.23 to define a beta for the option as well.

$$\beta_{\text{Option}} = \Omega\,\beta_{\text{Underlying}} \qquad (12.24)$$

Volatility, risk premium and the beta of an option are thus obtained by simply multiplying the corresponding variables for the underlying by the elasticity omega.

12.4.3 Summation of Sensitivities of Different Underlyings

The summation of sensitivities presented in Sect. 12.4.1 is only possible if they are sensitivities of instruments on the same underlying. It is often the case, however, that a portfolio consists of instruments on several different underlyings. The net delta of such a portfolio is *not* simply the sum of the deltas of the individual instruments. For example, in a chemical portfolio with options on Bayer and BASF stocks, the delta of the Bayer option with respect to the Bayer stock price cannot be added to the delta of the BASF option with respect to the BASF stock price to obtain the portfolio's delta. We must first answer the question: Which stock should we chose as the reference underlying

in calculating the sensitivity of our portfolio in the above example? Bayer or BASF?

Having chosen a reference underlying S_i from among all the instruments S_k in the portfolio with respect to which the sensitivities of the entire portfolio are to be calculated, we must then proceed by expressing the sensitivities of the remaining underlyings in terms of the sensitivity of this chosen reference. The Betas of the S_k with respect to the S_i can be used in the calculation of such an expression. From the volatilities of the prices and the pair wise correlation between the underlyings, we can use Eqs. 12.21 or A.20 to calculate the necessary Betas, even when neither of the prices is an index. From the sensitivity Δ_k of an instrument with value V_k on the underlying S_k we can then deduce the sensitivity Δ_i of this instrument with respect to *another* underlying S_i by using Eq. 12.22:

$$\Delta_i = \frac{dV_k}{dS_i} = \frac{dS_k}{dS_i}\frac{\partial V_k}{\partial S_k} = \frac{dS_k}{dS_i}\Delta_k = \frac{S_k}{S_i}\beta_{ki}\Delta_k = \frac{S_k\sigma_k}{S_i\sigma_i}\rho_{ki}\Delta_k . \qquad (12.25)$$

In the second to last step, Eq. 12.22 in the form "$dS/dI = \beta_{SI}S/I$" is applied while in the last step we make use of Eq. 12.21. Having in this manner transformed the sensitivities with respect to the underlyings S_k into sensitivities with respect to the selected reference underlying S_i, we can now simply add them up to calculate the total sensitivity of the portfolio since they now refer to the same underlying. The sensitivity of the entire portfolio with respect to the reference underlying is thus given by

$$\Delta_i^{\text{total}} = \frac{1}{S_i}\sum_k S_k\,\beta_{ik}\,\Delta_k = \frac{1}{S_i\sigma_i}\sum_k S_k\,\sigma_k\,\rho_{ik}\,\Delta_k . \qquad (12.26)$$

This reference underlying can (but need not) be an index on a relevant business sector, for example. Such a combination of securities on different underlyings makes sense when the underlyings are highly correlated. The correlation of the underlyings, or at least the correlation with the selected reference underlying, must be known for the following two reasons:

- They are necessary for the transformation of the deltas in accordance with Eq. 12.25.
- Sound decisions as to a reasonable choice of the above mentioned combination of securities can only be made on the basis of the correlations.

12.5 Computation of the Greek Risk Variables

Having introduced the *Greeks* and their applications in the last section, we will now give a short discussion on their determination using the most important pricing methods.

12.5.1 Sensitivities in the Binomial Model

In order to obtain the price changes in the binomial model resulting from a change in the underlying's spot price, under the assumption that all other parameters (including the time) remain the same, we require several different spot prices for the same time. This is the case for all times except at the beginning of the tree. To obtain more than one spot price at time t, we start the tree at least one time step earlier for each order of differentiation desired. The highest order derivative with respect to S which will be needed is that for gamma, i.e., the second derivative. Thus we let the tree start two time steps in the past, i.e., at time $t - 2dt$. In doing so, we generate (in a recombinant tree) three different price evolutions uuS, udS and ddS. The tree should be constructed in such a way that the middle price evolution udS represents the actual existing spot price $S(t)$. This implies that the tree must start at $S(t - 2dt) = e^{-2\mu dt} S(t)$ if we wish to apply the procedure given by Eq. 9.30. If Eq. 9.31 is to be utilized instead, the tree should begin with $S(t - 2dt) = S(t)$. Let $V(t)$ denote the value of any given financial derivative (or portfolio) on the spot price S. For a binomial tree beginning two time steps before the present time t, the following sensitivities can be calculated *directly* within the same tree:

$$\Delta \approx \frac{V^{uu}(t) - V^{dd}(t)}{(u^2 - d^2)S(t - 2\delta t)} \approx \frac{V^{uu}(t) - V^{ud}(t)}{u(u - d)S(t - 2\delta t)} \approx \frac{V^{ud}(t) - V^{dd}(t)}{d(u - d)S(t - 2\delta t)}$$

$$\Omega = \frac{u\, d\, S(t - 2\delta t)}{V^{ud}(t)}\Delta \tag{12.27}$$

$$\Gamma = 2\frac{d\, V^{uu}(t) - (u + d)V^{ud}(t) + uV^{dd}(t)}{ud(u - d)(u^2 - d^2)S^2(t - 2\delta t)}$$

$$\Theta = \frac{V^{uudd}(t + 2\delta t) - V(t - 2\delta t)}{4\delta t} \approx \frac{V^{uudd}(t + 2\delta t) - V^{ud}(t)}{2\delta t}$$

$$\approx \frac{V^{ud}(t) - V(t - 2\delta t)}{2\delta t} \ .$$

Other sensitivities such as vega and rho can only be determined via Eq. 12.17: we have to calculate the price for a slightly modified volatility or risk-free rate, respectively, with a completely new tree, subsequently dividing the difference between the old and the new option price by the respective change in the parameter to obtain the desired sensitivity.

For the first-order partial derivatives of the price, namely delta, omega and theta, the binomial model provides three different choices: a symmetric derivative, an asymmetric derivative for which the price change with respect to an increasing parameter value is used, and an asymmetric derivative for which the price change with respect to a declining parameter value is used. All these alternatives are displayed in Eq. 12.27.

We have previously introduced a somewhat different equation for delta (see Eq. 9.8). The difference is that in the earlier expressions for delta, the time was not assumed to be strictly constant but changed by dt. The price change induced by this time change is compensated for by the other form of the delta so that both methods are correct up to the order of precision obtained by using the binomial model.

The equation for Theta is, strictly speaking, only correct for the procedure given by Eq. 9.31 since in this procedure, the derivative prices used all contain the *same* underlying price. In the procedure given by Eq. 9.30, however, not only does the time change but, because of the drift appearing in u and d, the underlying S changes as well. Consequently, a change in the option price resulting *solely* because of a time change is not obtained. This is one (more) reason why Eq. 9.31 is used almost exclusively in this text.

The expression for gamma can be obtained from the following argument: gamma is by definition the change of delta per change of the underlying. At time t we have two deltas at our disposal, these being the two unsymmetrical derivatives in Eq. 12.27

$$\Delta_u \approx \frac{V^{uu}(t) - V^{ud}(t)}{u(u-d)S(t-2\delta t)} \quad , \quad \Delta_d \approx \frac{V^{ud}(t) - V^{dd}(t)}{d(u-d)S(t-2\delta t)} \ .$$

Δ_u is the delta between the upper and middle node at time t. Similarly Δ_d is the delta between the middle and lower node. The difference between these two deltas will serve as the delta change needed to calculate gamma. In addition we need the values of the underlying belonging to those two deltas since gamma is the delta change per corresponding underlying change. As the underlying value belonging to Δ_u we use the average (denoted by $\overline{S}_u$) of the underlying values at the upper and middle nodes. Similarly we use the average (denoted by $\overline{S}_d$) of the underlying values at the lower and middle nodes as the

underlying value belonging to Δ_d.

$$\overline{S}_u = \frac{1}{2}\left[uu + ud\right]S(t - 2\delta t) = \frac{1}{2}u(u + d)S(t - 2\delta t)$$

$$\overline{S}_d = \frac{1}{2}\left[dd + ud\right]S(t - 2\delta t) = \frac{1}{2}d(u + d)S(t - 2\delta t)$$

With respect to *these* underlying values Δ_u and Δ_d are even *symmetric* derivatives. Gamma is now simply the delta difference divided by the difference in underlying values:

$$\begin{aligned}
\Gamma &= \frac{\Delta_u - \Delta_d}{\overline{S}_u - \overline{S}_d} \\
&= \frac{1}{\overline{S}_u - \overline{S}_d}\frac{\left[V^{uu}(t) - V^{ud}(t)\right]/u - \left[V^{ud}(t) - V^{dd}(t)\right]/d}{(u - d)S(t - 2\delta t)} \\
&= \frac{1}{\overline{S}_u - \overline{S}_d}\frac{d\left[V^{uu}(t) - V^{ud}(t)\right] - u\left[V^{ud}(t) - V^{dd}(t)\right]}{ud(u - d)S(t - 2\delta t)} \\
&= \frac{2}{(u - d)(u + d)S(t - 2\delta t)}\frac{dV^{uu}(t) - (u + d)V^{ud}(t) + uV^{dd}(t)}{ud(u - d)S(t - 2\delta t)} \; .
\end{aligned}$$

This corresponds exactly to the gamma in Eq. 12.27.

A detailed demonstration on how to determine the sensitivities of an option portfolio using the binomial model is provided in the Excel workbook BINOMIALTREE.xls.

12.5.2 Sensitivities in the Black-Scholes Model

Since the Black-Scholes model provides analytic solutions for option prices, the sensitivities can be determined immediately by calculating the required partial derivatives directly. The sensitivities of a plain vanilla call and put on an underlying with spot price S earning a yield q will serve as an example and are given below. The expressions appearing here arise from differentiating[7] the

---

[7]Only the *explicit* dependence and not the *implicit* dependence of the option on the parameters under consideration play a role when calculating these partial derivatives. In determining theta, for example, the time dependence of S should not be taken into consideration, since this dependence is *implicit* for the option price. This means that in taking the partial derivatives, S, t, r, q and σ are to be viewed as independent parameters.

function in Eq. 8.7 and exploiting the relation in Eq. 12.16. The sensitivities of options on futures and forwards can be calculated similarly by differentiating the functions given in Eqs. 8.8 and 8.9.

$$\Delta_{\text{Call}} = B_q(t, T)\, \text{N}(x) \ , \quad \Delta_{\text{Put}} = B_q(t, T)\, [\text{N}(x) - 1]$$

$$\Gamma_{\text{Call}} = \frac{B_q(t, T)\, \text{N}'(x)}{S(t)\sigma\sqrt{T - t}} = \Gamma_{\text{Put}}$$

$$\Theta_{\text{Call}} = -\frac{B_q(t, T)S(t)\, \text{N}'(x)\sigma}{2\sqrt{T - t}} + S(t)\text{N}(x)\frac{\partial B_q(t, T)}{\partial t} - K\text{N}(x - \sigma\sqrt{T - t})\frac{\partial B(t, T)}{\partial t}$$

$$\Theta_{\text{Put}} = -\frac{B_q(t, T)S(t)\, \text{N}'(x)\sigma}{2\sqrt{T - t}} - S(t)\text{N}(-x)\frac{\partial B_q(t, T)}{\partial t}$$

$$\qquad + K\text{N}(-x + \sigma\sqrt{T - t})\frac{\partial B(t, T)}{\partial t}$$

$$\Psi_{\text{Call}} = B_q(t, T)S(t)\, \text{N}'(x)\sqrt{T - t} = \Psi_{\text{Put}}$$

$$\rho_{\text{Call}} = -K\text{N}(x - \sigma\sqrt{T - t})\frac{\partial B(t, T)}{\partial r} \ , \quad \rho_{\text{Put}} = K\text{N}(-x + \sigma\sqrt{T - t})\frac{\partial B(t, T)}{\partial t}$$

$$\rho_{q_{\text{Call}}} = S(t)\text{N}(x)\frac{\partial B_q(t, T)}{\partial q} \ , \quad \rho_{q_{\text{Put}}} = -S(t)\text{N}(-x)\frac{\partial B_q(t, T)}{\partial q} \ . \tag{12.28}$$

Here, x is as defined in Eq. 8.5 and N' again denotes the derivative of the cumulative standard normal distribution with respect to its argument:

$$\text{N}'(x) = \frac{d\text{N}(x)}{dx} = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

The Excel workbooks BLACKSCHOLES.XLS and STRADDLE.XLS demonstrate these sensitivities. In BLACKSCHOLES.XLS, the hedging of a portfolio consisting of puts, calls and underlyings is demonstrated and in STRADDLE.XLS, the behavior of the sensitivities of a straddle as a function of the underlying is described.

12.5.3 Sensitivities by Means of Finite Difference Methods

In principle, every numerical algorithm used to compute the price of an option can also be applied for the determination of the sensitivities or Greeks. All numerical procedures for computing the sensitivities are in essence based on the same fundamental idea displayed in Eq. 12.17. Two option prices

are computed. These prices differ since one of the risk factors used in their determination was slightly different. The sensitivity of the option price with respect to that particular risk factor is then the difference between the two option prices divided by the difference in the risk factor.

With finite difference methods—as with binomial trees—some sensitivities can be calculated (as finite differences) *directly on the grid* used for pricing without having to go through the lengthy procedure based on Eq. 12.17. In fact, a finite difference scheme does nothing other than calculate (approximations for) the derivatives with respect to S and t at each time step. The relevant formulas have already been presented in Chap. 10. Explicitly, Eq. 10.8 directly yields *delta*, Eq. 10.9 *gamma*, and Eq. 10.11 *Theta.* Calculating Greeks directly on the grid has proved to yield results which are significantly more exact[8] than procedures based on Eq. 12.17. Hence, the Greeks should be calculated on the grid whenever possible.

Vega and *rho,* however, cannot be calculated directly from an (S, t)-grid and so for these sensitivities we have to rely on Eq. 12.17, i.e., the calculation must be run through twice, in each case with a slight change in the respective parameter.

The determination of the sensitivities of an options portfolio using finite difference methods is shown in the Excel workbook FINITEDIFFER-ENCEMETHOD.xlsm from the download section [50].

12.5.4 Sensitivities by Means of Monte Carlo Simulations

When using Monte Carlo simulations in determining the sensitivities, the computations are always performed in accordance with Eq. 12.17. This means the two simulations are performed with slightly different values for the parameter under consideration but with *exactly the same random numbers*.

The difference between the two option prices divided by the difference in the parameter value gives an approximation of the desired sensitivity. Differentiating twice, as for gamma, for example, is accomplished in accordance with Eq. 12.17 with three different simulations for three different parameter values. Each simulation must be performed *with the same random numbers*. The computation of the sensitivities by means of Monte Carlo simulations is demonstrated in full detail for an option portfolio in the Excel workbook MONTECARLOSIMULATION.xlsm from the download section [50].

---

[8]For instance, we might otherwise observe that the sensitivities oscillate as a function of the spot price of the underlying.

13

Martingales and Numeraires

13.1 The Martingale Property

The most important and profound concept that the reader may have gained from the material presented in this book so far is that of risk neutrality, which can be summarized as follows:

> Today's price of a (tradable) financial instrument is equal to the discounted expectation of its future price if this expectation is calculated with respect to the risk-neutral probability measure.

At this point, we recommend that the reader reviews Sects. 7.1.3, 7.2.1 and 9.2.3. We will now elaborate on the concept of risk neutrality.

The *risk-neutral probability* is an example of a *martingale measure*. Martingale measures are a specific class of probability measures satisfying the property—as we are about to see—described in Eq. 13.1. For an intuitive explanation of the term *measure*: probability distributions can be interpreted as *measures* since the expectation of a function $f(X)$ of a random variable X, having a distribution with density function p, can be interpreted as an integral with respect to a certain *integral measure*:

$$\mathrm{E}[f(X)] = \int f(X) \, \underbrace{p(x)dx}_{\text{Integral Measure}}$$

By replacing the *Riemann integral measure dx* with the *probability measure* $p(x)dx$, we obtain the expectation of the random variable $f(X)$ by integrating the function f with respect to this probability measure.

Martingale theory is an intensively researched field of mathematics, and recognizing that the price process of a derivative can be interpreted as a martingale allows the application of martingale theory in the valuation of financial instruments. Because of its generality, the utilization of martingale theory requires very few of the assumptions listed in Chap. 4, namely Assumptions 1, 2, 3, 5. For sake of simplification, we assume that Assumption 4 from Chap. 4 is fulfilled, too. The methods presented here will not be discussed in complete mathematical detail, but will be motivated by our "experience" of the subject gained in previous sections.

The results in Sect. 9.2.3 are based on the somewhat surprising observation that the probability p' of an upward move u in the underlying price in the real world does not appear in Eqs. 9.8 or 9.13. Only the risk-neutral probability plays a role. This observation led to Eq. 9.20 and the interpretation of derivative prices as expectations taken with respect to a certain probability measure.

At this stage, we wish to generalize this concept. Consider again the derivation of Eq. 9.8. The specific type of the financial instrument V does not come into play. The discussion took place in the context of options merely for simplicity's sake, making use of none of the properties particular to options. Hence, Eq. 9.20 holds for all financial instruments whose value is governed by the price of the underlying S. Furthermore, at no point in the derivation did the process S have to have a particular form. This means that Eq. 9.20 holds for *arbitrary* instruments on *arbitrary* underlyings, in other words, for all general processes of the form 2.19. Such processes are much more general than the simple random walk given by Eq. 2.23, for example. This is of decisive importance in the analysis of term structure models. In fact, the generality of Eq. 9.20 goes still further. Using $B(T, T) = 1$, Eq. 9.20 can be written as

$$\frac{V(t)}{B(t, T)} = \mathrm{E}\left[\frac{V(T)}{B(T, T)}\right] ,$$

where we recall that the expectation at time t has been taken with respect to the risk-neutral probability p.

Defining the *normalized* price as

$$Z(t) := \frac{V(t)}{B(t, T)} ,$$

we arrive at the elegant form

$$Z(t) = \mathrm{E}\left[Z(T)\right] . \qquad (13.1)$$

In other words:

> The normalized price at time t is given by the expectation (with respect to p) of the future normalized price.

This expresses the martingale property in its "purest" form. We can also say:

> The normalized price is a martingale (with respect to p).

Normalizing the price means nothing other than expressing the price of the instrument in units of zero bonds maturing at T rather than in monetary units such as euros. Thus, the numerical price does not tell us how much the instrument costs in euros but how much the instrument costs in terms of zero bonds.

Having shown that the martingale property holds (within limits[1]) for *arbitrary financial instruments* on *arbitrary underlyings*, we now show that it also holds for *arbitrary normalizing factors*. This indicates the truly general character of the martingale property.

13.2 The Numeraire

In financial literature, the normalizing instrument is commonly referred to as *numeraire*. In our discussions here, we will frequently use the more intuitive expression *normalization*. We will now show that not only zero bonds but *arbitrary* (tradable) financial instruments may serve as normalizing factors, i.e., as numeraires. The numeraire used does not even have to refer to the underlying S (the zero bond does not do so either). Let S be the price of an underlying, V the price of an arbitrary financial instrument on this underlying,

---

[1]The limits are that the underlying has to follow a general Ito process of the form 2.19 and that the financial instrument as well as the normalizing factor have to be traded instruments, see Sect. 14.4.

and Y another arbitrary financial instrument (in our previous discussions, the zero bond was chosen to play the role of Y). As in the derivation of Eq. 9.8, let the time evolution of the price of the underlying be described by a *binomial tree* such as given in Eq. 9.2, for example.

We intend to construct a portfolio composed of α underlyings whose prices are given by the process S and β of the financial instruments with a price process given by Y:

$$\Pi(t) = \alpha_t S(t) + \beta_t Y(t) \ . \tag{13.2}$$

We require this portfolio to have a value equal to that of the derivative V in all states of the world one time step later. Note that both the normalizing factor Y and the underlying S must be *tradable* since otherwise the "α underlyings at price S" or the "β instruments at price Y" could be neither purchased nor sold on the market and the construction of such a portfolio would be impossible. Since the normalizing instrument Y can, as will be shown, be chosen arbitrarily, there is no shortage of candidates for Y; we simply select any *tradable* instrument as the numeraire.

However, it is often the case that while a financial instrument on an underlying is tradable, the underlying *itself* is *not* tradable. This situation arises quite frequently. For example, the forward price $S(t, T)$ (see Eq. 6.1) is generally not tradable even if the associated spot price $S(t)$ is the price of a tradable instrument (such as a stock). Despite this fact, the forward price is often used as an underlying; the reader is referred to Eqs. 8.8 or 8.9 for examples. Also, as we will show in detail in Sect. 14.4, interest rates are not tradable either, in contrast to bonds which are financial instruments having the interest rate as an "underlying". In such cases, a second tradable instrument U_S having S as its underlying is chosen in addition to the numeraire instrument Y. The only restriction in the choice of this second instrument is that it must not be possible to construct U_S by a portfolio consisting solely of the numeraire instrument (we need two truly "linearly independent" instruments).

A portfolio can now be constructed similar to Eq. 13.2 by replacing the non-tradable underlying S with the tradable instrument U_S. A non-tradable underlying does not, in principle, complicate the situation as long as a tradable instrument on the underlying U_S (which cannot be represented by the numeraire instrument) can be found.[2]

---

[2]If S itself is tradable, it can, of course, be chosen as the (tradable) instrument U_S. If this is the case, we merely need to replace U_S with S in all pertinent equations derived in the following material.

$$\Pi(t) = \alpha_t U_S(t) + \beta_t Y(t) \; . \tag{13.3}$$

The portfolio thus constructed is required to have the same value as the derivative V after proceeding by one time step. Hence

$$\alpha_t U_{S_u} + \beta_t Y_u = \Pi_u \overset{!}{=} V_u$$

$$\alpha_t U_{S_d} + \beta_t Y_d = \Pi_d \overset{!}{=} V_d \; . \tag{13.4}$$

We end up with two equations for both cases $S(t + \delta t) = S_u$ and $S(t + \delta t) = S_d$.

The weights α and β satisfying both of these equations can now be uniquely determined.[3]

$$\alpha_t = \frac{V_u Y_d - V_d Y_u}{U_{S_u} Y_d - U_{S_d} Y_u} \;\; , \quad \beta_t = \frac{V_d U_{S_u} - V_u U_{S_d}}{U_{S_u} Y_d - U_{S_d} Y_u} \; . \tag{13.5}$$

If in all events the derivative and the portfolio have the same value at time $t + \delta t$, their values must also be equal at time t. Otherwise an arbitrage opportunity would exist.

$$\begin{aligned} V(t) = \Pi(t) &= \alpha_t U_S(t) + \beta_t Y(t) \\ &= \frac{V_u Y_d - V_d Y_u}{U_{S_u} Y_d - U_{S_d} Y_u} U_S(t) + \frac{V_d U_{S_u} - V_u U_{S_d}}{U_{S_u} Y_d - U_{S_d} Y_u} Y(t) \; . \end{aligned} \tag{13.6}$$

It is exactly at this point that the assumption of an *arbitrage free market* enters into our discussion. As we continue with the derivation, we will clearly recognize how this assumption, together with the normalizing factor Y, uniquely determines the martingale probability p. Collecting terms with respect to the coefficients of V_u and V_d, we obtain

$$V(t) = V_u \frac{Y_d U_S(t) - U_{S_d} Y(t)}{U_{S_u} Y_d - U_{S_d} Y_u} + V_d \frac{U_{S_u} Y(t) - Y_u U_S(t)}{U_{S_u} Y_d - U_{S_d} Y_u} \; .$$

In view of our goal of finding a representation of the value of the derivative normalized with respect to the numeraire instrument Y, we rewrite the above

---

[3]E.g., α may be determined by multiplying the first equation with Y_d and the second equation with Y_u and subtraction of the resulting equations. Similar, β could be determined by multiplying the first equation with U_{S_d} and the second equation with U_{S_u} and, again, subtraction of the resulting equations.

equation in the form

$$\frac{V(t)}{Y(t)} = \frac{V_u}{Y_u} \frac{Y_d U_S(t)/Y(t) - U_{S_d}}{U_{S_u} Y_d/Y_u - U_{S_d}} + \frac{V_d}{Y_d} \frac{U_{S_u} - Y_u U_S(t)/Y(t)}{U_{S_u} - U_{S_d} Y_u/Y_d} \ .$$

This equation can now be written as

$$Z(t) = Z_u \, p_u + Z_d \, p_d \tag{13.7}$$

by defining the normalized prices as

$$Z_S(t) := \frac{V(t)}{Y(t)} \ , \quad Z_S^{u,d}(t + \delta t) := \frac{V_{S_{u,d}}(t + \delta t)}{Y_{u,d}(t + \delta t)}$$

and the "probabilities" as

$$p_u := \frac{Y_d \frac{U_S}{Y} - U_{S_d}}{U_{S_u} \frac{Y_d}{Y_u} - U_{S_d}} = \frac{Y_u Y_d U_S - Y_u Y U_{S_d}}{Y_d Y U_{S_u} - Y_u Y U_{S_d}} = \frac{\frac{U_S}{Y} - \frac{U_{S_d}}{Y_d}}{\frac{U_{S_u}}{Y_u} - \frac{U_{S_d}}{Y_d}}$$

$$p_d := \frac{U_{S_u} - Y_u \frac{U_S}{Y}}{U_{S_u} - U_{S_d} \frac{Y_u}{Y_d}} = \frac{Y_d Y U_{S_u} - Y_d Y_u U_S}{Y_d Y U_{S_u} - Y_u Y U_{S_d}} = \frac{\frac{U_{S_u}}{Y_u} - \frac{U_S}{Y}}{\frac{U_{S_u}}{Y_u} - \frac{U_{S_d}}{Y_d}} \ . \tag{13.8}$$

where the last expressions are obtained by dividing both the numerator and denominator by $Y_u Y_d Y$. Note that these p are independent of the derivative V. They depend explicitly only on the normalizing factor Y and the instrument U_S. The form of these functions arose from the assumption of an arbitrage-free market as expressed in Eq. 13.6. If p_u and p_d could actually be interpreted as probabilities, Eq. 13.7 would have the form indicated in Eq. 13.1. Before showing that this is the case, we make the following remark on the computation of p in practice: Instead of expressing p in terms of U_S and Y, as just presented, Eq. 13.7 can be used to write p as a function of Z (and thus of V and Y). Making use of the equality $p_d = 1 - p_u$ yields

$$p_u = \frac{Z(t) - Z_d}{Z_u - Z_d} \ . \tag{13.9}$$

This is a method frequently employed in explicitly computing martingale probabilities in practice. This expression has the disadvantage that the independence of p on the derivative V is not immediately recognizable.

Before we can interpret p as a probability, it remains to show that p does in fact satisfy all requisite properties. The following conditions, holding for all probabilities, must be checked[4]:

$$p_u + p_d = 1 \quad , \quad p_u \geq 0 \quad , \quad p_d \geq 0 \tag{13.10}$$

Using simple algebra, it follows immediately from the explicit representation in Eq. 13.8 that $p_u + p_d = 1$ holds. In order to recognize the implications of the other two conditions, note that p_u and p_d have a *common* denominator $\frac{U_{S_u}}{Y_u} - \frac{U_{S_d}}{Y_d}$. This factor is greater than zero if and only if the normalized price of U_S in the "down state" is smaller than in the "up state". This is not necessarily always the case,[5] since "up" and "down" are defined by the *un*normalized *underlying* price S ($S_u > S_d$ by definition) and not by the normalized price of the instrument U_S. If such should be the case, i.e., if the *denominator* should be less than zero, *both numerators* must be less than zero as well. Just as both numerators must be greater then zero if the denominator is greater than zero. In summary,

$$\frac{U_{S_u}}{Y_u} > \frac{U_S}{Y} > \frac{U_{S_d}}{Y_d} \quad \text{for} \quad S_u > S > S_d$$

$$\text{or} \tag{13.11}$$

$$\frac{U_{S_u}}{Y_u} < \frac{U_S}{Y} < \frac{U_{S_d}}{Y_d} \quad \text{for} \quad S_u > S > S_d$$

must hold. The normalized price of the instrument U_S must therefore be a strictly monotone function of the underlying price. If this is not the case, we are immediately presented with an arbitrage opportunity. Let us assume for instance, that $\frac{U_{S_u}}{Y_u} < \frac{U_S}{Y} > \frac{U_{S_d}}{Y_d}$. This market inefficiency could be exploited by selling (short selling) the instrument U_S at time t and using the proceeds to purchase $a = U_S/Y$ of the instrument Y. This is always possible as both Y and U_S are tradable instruments. This portfolio has a value at time t of

$$-U_S + aY = -U_S + (\frac{U_S}{Y})Y = 0$$

---

[4]If all three of these conditions hold, it follows immediately that $p_u \leq 1$ and $p_d \leq 1$ as well.

[5]Even when S is tradable, allowing U_S to be replaced by the underlying, there are several common instruments that violate this condition when used as a normalizing instrument. For example, the value C of a plain vanilla call on S increases *faster* (in percentage terms) than S itself, implying for the quotient $\frac{S}{C(S)}$ that $\frac{S_2}{C(S_2)} < \frac{S_1}{C(S_1)}$ for $S_2 > S_1$.

One time step later, the portfolio's value is, in all events u and d, positive since

$$-U_{S_u} + aY_u = -U_{S_u} + (\frac{U_S}{Y})Y_u = Y_u \underbrace{\left[-\frac{U_{S_u}}{Y_u} + \frac{U_S}{Y} \right]}_{>0} > 0$$

$$-U_{S_d} + aY_d = -U_{S_d} + (\frac{U_S}{Y})Y_d = Y_d \underbrace{\left[-\frac{U_{S_d}}{Y_d} + \frac{U_S}{Y} \right]}_{>0} > 0$$

This strategy leads to a certain profit without placing investment capital at risk. The fact that the value of the portfolio is positive in both possible states u and d can be directly attributed to the assumption that $\frac{U_S}{Y}$ is greater than both $\frac{U_{S_d}}{Y_d}$ and $\frac{U_{S_u}}{Y_u}$, thereby violating the condition in Eq. 13.11. Conversely, if $\frac{U_{S_u}}{Y_u} > \frac{U_S}{Y} < \frac{U_{S_d}}{Y_d}$, an analogous arbitrage opportunity arises by following the strategy of going long in U_S and short in $a = U_S/Y$ of the instrument Y.

It follows immediately from these arbitrage considerations that, if the market is arbitrage free, the condition in Eq. 13.11 is *automatically* satisfied by every tradable financial instrument playing the role of the numeraire and every tradable instrument U_S on the underlying and, in consequence, need not be verified in practice. p_u and p_d in Eq. 13.8 are therefore actually probabilities (they satisfy all the properties in 13.10), implying that the normalized price $Z = V/Y$ (and thus the normalized prices of all tradable instruments on S) is a martingale. Or conversely, if it were possible to find a tradable instrument U_S on S whose price, normalized with respect to numeraire instrument Y, is not a martingale, a portfolio consisting of instruments U_S and Y could be constructed, making arbitrage possible.

13.3 Self-financing Portfolio Strategies

The discussion given above was restricted to one single discrete time step. The extension to arbitrarily many discrete time steps is completely analogous. As described in Chap. 9, we obtain at every node of the tree a replicating portfolio

consisting of the numeraire instrument and a financial instrument on the underlying (or the underlying itself if it is tradable[6]) in order to replicate both possible derivative prices in the next step. Such a replication of the derivative V through a portfolio composed of U_S and Y (or S and Y, if S is tradable) corresponds to a *hedge* of V (for more on this subject, see Chap. 12). As we move from one time step to the next, this hedge is constantly adjusted by adjusting the value of the weights α_t and β_t of the (financial instrument on the) underlying and the numeraire.

Two conditions must be fulfilled to make this adjustment. First, it must be possible to specify the weights at the beginning of a time step and they must not depend on later realizations of the spot price. That is, the weights α_t and β_t required to replicate exactly the derivative one time step later at $t + \delta t$ in all cases need to be known at time t. In other words, it is predictable (previsble) at time t which weights α_t and β_t will replicate the derivative at time $t + \delta t$. Such a process, which's value at one time step later is known at any time step is called *previsable process*. The second condition demands that the portfolio is self financing. This implies that the required adjustments of the weights by selling or buying U_S must be reflected by similar buying or selling an equal amount of the numeraire Y. In other words, at no time money must be injected to or withdrawn from the hedge portfolio of underlying and numeraire. Such a strategy is called *self-financing*.

To ensure that the first condition is met, we consider again Eq. 13.5:

$$\alpha_t = \frac{V_u Y_d - V_d Y_u}{U_{S_u} Y_d - U_{S_d} Y_u}, \quad \beta_t = \frac{V_d U_{S_u} - V_u U_{S_d}}{U_{S_u} Y_d - U_{S_d} Y_u} \,.$$

The terms $V_{u,d}$, $Y_{u,d}$ and U_{S_u, S_d} depend on $S(t)$ (since then S_u resp. S_d are determined), but don't depend on $S(t + \delta t)$, i.e., the actual realization of S at time $t + \delta t$. Therefore, it is a previsible process and the first condition is met.

Next, we consider the time development of the hedge-portfolio $\Pi(t)$ and the derivative $V(t)$ to verify the second condition. At the start time t we have by construction

$$V(t) = \Pi_t(t) = \alpha_t U_S(t) + \beta_t Y(t) \,.$$

---

[6]In Chap. 9, the normalizing instrument Y was taken to be the zero bond B with a face value of $B(T, T) = 1$ and weight $\beta_t = g(t)$. The value of this zero bond changed from t to $t + \delta t$ from $Y = B(t, T)$ to $Y_u = Y_d = B(t, T)B(t)^{-1}$ with $B(t) = B(t, t + dt)$. The weight α_t of the underlying was denoted by Δ.

The index t of $\Pi_t(t)$ indicates here that we consider the hedge portfolio with parameters α_t and β_t, as determined at time t. One time step late, we have

$$\Pi_t(t + \delta t) = \alpha_t U_S(t + \delta t) + \beta_t Y(t + \delta t) \ .$$

Now we determine new values for α and β belonging to time $t + \delta t$ and the hedge portfolio will be re-adjusted. Thus, we get for the new portfolio

$$V(t + \delta t) = \Pi_{t+\delta t}(t + \delta t) = \alpha_{t+\delta t} U_S(t + \delta t) + \beta_{t+\delta t} Y(t + \delta t) \ .$$

Our hedging strategy is self-financing if and only if the portfolio value before and after the re-adjustment is the same:

$$\Pi_{t+\delta t}(t + \delta t) = \Pi_t(t + \delta t) = \alpha_t U_S(t + \delta t) + \beta_t Y(t + \delta t) \ . \qquad (13.12)$$

Since α_t and β_t are determined according to Eq. 13.4, it follows immediately that Eq. 13.12 is also true in all possible scenarios of our binomial world ($S(t + \delta t) = S_u$ or $S(t + \delta t) = S_d$). Therefore, the strategy is self-financing by construction.

It is worthwhile to consider this fact from another point of view which allows us to quickly verify whether or not a portfolio is self financing. The total difference $\delta\Pi(t)$ in the value of a portfolio for a certain trading strategy over a time span δt equals the difference in the value at time $t + \delta t$ of the portfolio set up at time $t + \delta t$ and the value at time t of the portfolio set up at time t, thus

$$\delta\Pi(t) = \Pi_{t+\delta t}(t + \delta t) - \Pi_t(t) = \underbrace{\Pi_t(t + \delta t) - \Pi_t(t)}_{\delta\Pi^{\mathrm{Market}}} + \underbrace{\Pi_{t+\delta t}(t + \delta t) - \Pi_t(t + \delta t)}_{\delta\Pi^{\mathrm{Trading}}} \ .$$

We arrive at the second equation by simply inserting a zero in the form $0 = \Pi_t(t + \delta t) - \Pi_t(t + \delta t)$. It is now easy to recognize both components contributing to the total change in Π: $\delta\Pi^{\mathrm{Market}}$ is the change in value of the portfolio resulting from changes in the market without having adjusted the positions in the portfolio. $\delta\Pi^{\mathrm{Trading}}$ is the difference at time $t + \delta t$ between the value of the new portfolio and that of the old portfolio, i.e., the value change resulting solely from trading. For the strategy to be self financing, $\delta\Pi^{\mathrm{Trading}}$ must equal zero since the value of the old portfolio must provide exactly the funds necessary to finance the new portfolio. Consider for example a portfolio

composed of two instruments having the form $\Pi(t) = \alpha_t U_S(t) + \beta_t Y(t)$ as above:

$$\delta\Pi^{\text{Market}} = \Pi_t(t + \delta t) - \Pi_t(t) = \alpha_t \underbrace{[U_S(t + \delta t) - U_S(t)]}_{\delta U_S(t)} + \beta_t \underbrace{[Y(t + \delta t) - Y(t)]}_{\delta Y(t)}$$

$$\delta\Pi^{\text{Trading}} = \Pi_{t+\delta t}(t + \delta t) - \Pi_t(t)$$

$$= U_S(t + \delta t)\underbrace{[\alpha(t + \delta t) - \alpha_t]}_{\delta\alpha_t} + Y(t + \delta t)\underbrace{[\beta(t + \delta t) - \beta_t]}_{\delta\beta(t)},$$

that is,

$$\delta\Pi(t) = \alpha_t\delta U_S(t) + \beta_t\delta Y(t) + U_S(t + \delta t)\delta\alpha_t + Y(t + \delta t)\delta\beta_t$$

$$= \alpha_t\delta U_S(t) + \beta_t\delta Y(t) \quad \text{for a self financing strategy.} \quad (13.13)$$

This implies that the strategy is self financing if and only if the *total* change in the portfolio's value from one adjustment period to the next can be explained exclusively by *market* changes. This holds for infinitesimal time steps as well:

$$\Pi(t) = \alpha_t U_S(t) + \beta_t Y(t) \quad \text{self financing strategy}$$

$$\Longleftrightarrow \quad d\Pi(t) = \alpha_t dU_S(t) + \beta_t dY(t). \quad (13.14)$$

From this point of view, we again consider the structure of Eq. 13.13. The change $\delta\Pi$ over the next time step is composed of the change δU and the change δY. At time t, neither the value of δU nor of δY are known (we cannot even say if these are "upward" or "downward"), but the *coefficients* α_t and β_t controlling the influence of these changes on $\delta\Pi$ are known at time t.

Compare this with the general Ito process given in Eq. 2.19. The coefficients $a(S, t)$ and $b(S, t)$ which control the next step in the process are also already known at time t. These coefficients a and b are also *previsible processes*.

13.4 Generalization to Continuous Time

The profound and important concepts presented above can be summarized as follows:

- For any arbitrary underlying S which follows a stochastic process of the general form indicated in Eq. 2.19,

- and any tradable instrument U with S as its underlying (or S itself if it is tradable)
- and any other arbitrary, tradable, financial instrument Y (*numeraire*),
- the assumption of an arbitrage-free market implies the existence of a (numeraire-dependent) unique probability *measure p*,
- such that the current, normalized price $Z = V/Y$ (also called the *relative price*) of an arbitrary, tradable, financial instrument V on S
- equals the expectation of the future normalized price and thus, Z is a martingale with respect to p.
- This statement holds because a *self financing* portfolio strategy with *previsible* weights can be followed which replicates (*hedges*) the price of the financial instrument V at all times.

The expectation is taken at time t with respect to the Y-dependent probability measure p. This dependence is given explicitly by Eq. 13.8 (in the context of a binomial tree over one time step). To emphasize this dependence, the expectation is often equipped with the subscript t and the superscript Y:

$$Z(t) = \mathrm{E}_t^Y [Z(u)] \quad \forall u \geq t . \tag{13.15}$$

The intuitive interpretation of Eq. 13.15 is that the *expected* change in the normalized price of the tradable instrument is zero with respect to this probability measure, in other words, the normalized price has *no drift*. This implies that in this measure such a normalized price process has the form given in Eq. 2.19 with no drift term:

$$dZ(t) = b_Z(S, t) \, d\widetilde{W} \quad \text{mit} \quad d\widetilde{W} \sim X\sqrt{dt}, \ X \sim \mathrm{N}(0, 1) . \tag{13.16}$$

Here, $d\widetilde{W}$ is a Brownian motion and b_Z a (*previsible*) process which is different for each different instrument.

The material presented thus far has been restricted to *discrete* time steps. But the above statements hold in continuous time as well. To see this, several fundamental theorems from stochastic analysis are required. In the following, the insights gained in the study of discrete processes will be carried over to the continuous case and the necessary theorems from stochastic analysis will be used (without proof) when needed. In addition to extending the results already obtained to the time continuous case, the object of the following discussion is to provide a deeper understanding of the general approach to pricing derivatives, in particular, an understanding of role of the *drift* of an underlying.

We therefore consider a very general (not necessarily tradable) underlying S, which, *in the real world*, is governed by an Ito process satisfying Eq. 2.19, i.e.

$$dS(t) = a(S, t)\, dt + b(S, t)\, dW \quad \text{with} \quad dW \sim X\sqrt{dt},\ X \sim \mathrm{N}(0, 1)\ . \tag{13.17}$$

Let $U(S, t)$ denote the price of a tradable, financial instrument with an underlying S. The process for U *in the real world,* according to Ito's lemma, is given by Eq. 2.21 as

$$dU(S, t) = a_U(S, t)dt + \frac{\partial U}{\partial S}\, b(S, t)\, dW \quad \text{with}$$

$$a_U(S, t) := \frac{\partial U}{\partial S} a\, (S, t) + \frac{\partial U}{\partial t} + \frac{1}{2}\frac{\partial^2 U}{\partial S^2} b(S, t)^2\ , \tag{13.18}$$

Here $a_U(S, t)$ denotes the drift of U.

Furthermore, we select an arbitrary, tradable instrument Y as the numeraire instrument. Note, however, that the choice of the numeraire Y is not completely arbitrary. It has always been tacitly assumed that our market is driven by just one *single* random factor (*one-factor model*), namely the *Brownian motion* dW in Eq. 2.19. The numeraire instrument may indeed be any arbitrary, deterministic or stochastic instrument, but if it has a stochastic component, it must be driven by *the same* random walk as the underlying S. If not, the resulting model would be a *multi-factor model,* and in consequence could not be completely "spanned" by the two instruments U and Y. Analogous to Eq. 2.19, the most general process describing the numeraire instrument Y satisfies

$$dY(t) = m(Y, t)\, dt + n(Y, t)\, dW \quad \text{with} \quad dW \sim X\sqrt{dt},\ X \sim \mathrm{N}(0, 1) \tag{13.19}$$

with (*previsible*) processes m and n and the same random walk dW, which drives the random component of the underlying S in Eq. 13.17.

Motivated by our experience with Sect. 13.1, we seek a probability measure with respect to which the prices of all tradable instruments (which depend on no stochastic factors other than the Brownian motion dW in Eq. 13.17) normalized with the numeraire instrument Y are martingales. We are as yet quite far from attaining this goal. We start by aiming at the short-term goal of

finding a measure for which the normalized price

$$Z(S,t) := \frac{U(S,t)}{Y(t)} \tag{13.20}$$

of a *single* selected instrument U is a martingale. The product rule establishes the following equation for the process Z[7]:

$$dZ = d\left[Y^{-1}U\right] = U d\left[Y^{-1}\right] + Y^{-1}dU + dU d\left[Y^{-1}\right] \tag{13.21}$$

The differential dU was already specified above. The differential of $f(Y) := Y^{-1}$ is obtained through an application of Ito's lemma, Eq. 2.21, under consideration of $\partial f/\partial Y = -1/Y^2$, $\partial^2 f/\partial Y^2 = 2/Y^3$ and $\partial f/\partial t = 0$. Simple substitution gives

$$d\left[Y^{-1}\right] = \left[-\frac{1}{Y^2}m + \frac{1}{Y^3}n^2\right]dt - \frac{1}{Y^2}n\,dW$$

The last term in Eq. 13.21 appears since the product of the two differentials contains not only higher order terms but also a term $\sim dW^2$ which is *linear* in dt (see Eq. 2.20), explicitly:

$$dU d\left[Y^{-1}\right] = \left(a_U dt + \frac{\partial U}{\partial S} b\,dW\right)\frac{1}{Y^2}\left(\left[-m + n^2/Y\right]dt - n\,dW\right)$$

$$= -\frac{\partial U}{\partial S} b \frac{n}{Y^2}\underbrace{(dW)^2}_{\sim dt} + \mathcal{O}\left(dt\,dW\right)\ .$$

This effect clearly stems from the fact that both dU and dY are no ordinary but *stochastic* differentials.[8] Altogether dZ becomes

$$dZ = \frac{U}{Y}\left(\left[\frac{n^2}{Y^2} - \frac{m}{Y}\right]dt - \frac{n}{Y}dW\right) + \frac{a_U}{Y}dt + \frac{\partial U}{\partial S}\frac{b}{Y}dW - \frac{\partial U}{\partial S}\frac{bn}{Y^2}dt$$

---

[7]We suppress the arguments of U, Y, a, b, m and n in order to keep the notation simple. The arguments of these variables are always those as given in Eqs. 13.17 and 13.19.

[8]Equation 13.21 can be proven formally by applying Ito's lemma (in the version for two stochastic variables) to the function $f(U, Y) = U Y^{-1}$.

$$= \left(\frac{b}{Y} \frac{\partial U}{\partial S} - \frac{n}{Y} \frac{U}{Y} \right) dW + \left(\frac{a_U}{Y} + \left[\frac{n^2}{Y^2} - \frac{m}{Y} \right] \frac{U}{Y} - \frac{bn}{Y^2} \frac{\partial U}{\partial S} \right) dt$$

$$= \left(\frac{b}{Y} \frac{\partial U}{\partial S} - \frac{n}{Y} \frac{U}{Y} \right) \left\{ dW + \frac{a_U + \left[\frac{n^2}{Y^2} - \frac{m}{Y} \right] U - b \frac{n}{Y} \frac{\partial U}{\partial S}}{b \frac{\partial U}{\partial S} - n \frac{U}{Y}} dt \right\},$$

$$(13.22)$$

where the coefficient of dW is factored out "by force" in the final step. We seek a probability measure with respect to which Z is a martingale, or in other words, a process of the form specified in Eq. 13.16. The process dZ would have the desired form if a measure existed with respect to which

$$d\widetilde{W} := dW + \frac{a_U + \left[\frac{n^2}{Y^2} - \frac{m}{Y} \right] U - b \frac{n}{Y} \frac{\partial U}{\partial S}}{b \frac{\partial U}{\partial S} - n \frac{U}{Y}} dt \qquad (13.23)$$

were a standard Brownian motion, i.e.,

$$d\widetilde{W} \sim X\sqrt{dt} \ \ \mathrm{mit} X \sim N(0, 1) \,.$$

Stochastic analysis delivers just such a theorem, namely the famous Girsanov Theorem:

Theorem 1 (Girsanov) *Let $W(t)$ be a Brownian motion with respect to a probability measure $\mathcal{P}$, and $\gamma(t)$ a previsible process which (for some future time T) satisfies the boundedness condition*

$$\mathrm{E}_{\mathcal{P}} \left[\exp \left(\frac{1}{2} \int_0^T \gamma(t) dt \right) \right] < \infty$$

Then there exists a measure $\mathcal{Q}$, equivalent[9] to $\mathcal{P}$, with respect to which

$$\widetilde{W}(t) = W(t) + \int_0^t \gamma(s) ds$$

---

[9]Two probability measures are called equivalent if they agree exactly on what is possible and what is impossible. I.e. an event is impossible (probability zero) in one probability measure if and only if it is impossible in all equivalent probability measures.

is a Brownian motion. This implies that

$$dW(t) + \gamma(t)dt = d\widetilde{W}(t) \sim X\sqrt{dt} \ \ with \ X \sim N(0, 1)$$

Conversely, in the measure $\mathcal{Q}$ the original process $W(t)$ is a Brownian motion with an additional drift component, $-\gamma(t)$: $dW(t) = d\widetilde{W}(t) - \gamma(t)dt$

In order to apply the theorem, the coefficient of dt in Eq. 13.23 must be identified with the process $\gamma(t)$ in the Girsanov Theorem. We begin by observing that this coefficient depends only on variables which can be evaluated at time t and as such can itself be determined at time t. This implies that it is previsible. Proceeding under the assumption that the technical boundedness condition in the theorem is satisfied (this will always be the case in our models), the theorem provides a measure with respect to which $d\widetilde{W}$ is in fact a simple Brownian motion. dZ in Eq. 13.22 then has a form as in Eq. 13.16 with

$$b_Z(S, t) = \frac{b}{Y}\frac{\partial U}{\partial S} - \frac{n}{Y}\frac{U}{Y} \ . \tag{13.24}$$

Can we now conclude that Z is a martingale with respect to this measure? Does the absence of a drift in Eq. 13.16 directly imply[10] the martingale property in Eq. 13.1? Again, stochastic analysis provides theorems which ensures (if certain technical conditions are satisfied, see [11, page 79], for example) that this is the case. The measure for which dZ has the form 13.16 is thus a martingale measure.

We have thus attained our first goal by finding a measure with respect to which the normalized price of one selected instrument U (or for S if S should be tradable) is a martingale. The only requirement made of the instrument U is that it be tradable. Thus, for every arbitrary, tradable financial instrument (with S as an underlying) there exists a martingale measure. This measure could, at this point in the discussion, be different for each instrument U, in other words, it may be dependent on our choice of U (just as it depends on the choice of numeraire instrument Y). It remains to show that the normalized price of *every* tradable instrument (with S as its underlying) is a martingale with respect to *the same* probability measure. In the discrete case, the essential point was that a self financing strategy for a portfolio consisting of U and Y

---

[10] We have already shown the reverse implication above.

exists which replicates V exactly. We will need another important theorem from stochastic analysis to establish this for the time continuous case:

Theorem 2 (Martingale Representation) *If Z is a martingale with respect to the probability measure $\mathcal{P}$ with a volatility which is non-zero almost everywhere with probability $P = 1$, i.e., if Z follows a stochastic process satisfying*

$$dZ = b_Z(t)dW \quad with \quad P[b_Z(t) \neq 0] = 1 \quad \forall t$$

with a previsible process $b_Z(t)$, and if there exists in this measure another martingale X, then there exists a previsible process $\alpha(t)$ such that

$$dX = \alpha(t)dZ$$

Or equivalently in integral form

$$X(t) = X(0) + \int_0^t \alpha(s)dZ(s)$$

The process $\alpha(t)$ is unique. Furthermore, α and b_Z together satisfy the boundedness condition

$$\mathrm{E}\left[\exp\left(\frac{1}{2}\int_0^T \alpha^2(t)b_Z^2(t)dt\right)\right] < \infty .$$

This theorem states intuitively that (if the volatility is non-zero), two martingales differ at most by a previsible process. This implies that any martingale can be represented by any other martingale and a previsible process.

As yet, we only have one martingale in our measure, namely Z, the normalized price of U. In order to apply the theorem, we need a second martingale. And since we wish to gather information about the price of another arbitrary financial instrument V, we must construct a second martingale from this instrument V. We do this with the help of yet another quite simple theorem:

Theorem 3 (Tower Law) *For any arbitrary function V, depending on events occurring up to some specified future time $T > t$, the expectation at time t of $V(T)$ with respect to any arbitrary probability measure $\mathcal{P}$,*

$$E(t) := E_t^{\mathcal{P}}[V(T)]$$

is a martingale with respect to $\mathcal{P}$, explicitly

$$E(t) = E_t^{\mathcal{P}}\left[E(u)\right] \quad \forall u > t$$

It is easy to see that this theorem is true: substituting the definition of E into the claim that $E(t)$ is a martingale reads

$$E_t^{\mathcal{P}}\left[V(T)\right] = E_t^{\mathcal{P}}\left[E_u^{\mathcal{P}}\left[V(T)\right]\right] \quad \forall u > t \tag{13.25}$$

This implies that in taking the expectation at time u and subsequently taking the expectation *of this expectation* at an earlier time t we arrive at the same result as if we had directly taken the expectation with respect to the earlier time t in the first place. The reader should become familiar with this idea by verifying it using the binomial or trinomial trees presented in Figs. 9.1 or 9.2.

The payoff profile $V(T)$ of a financial instrument with maturity T is a function depending only on events (values of the underlying process S) occurring up to time T. The Tower Law states that the *expectation* of this payoff profile is a martingale with respect to *every* probability measure, in particular with respect to the martingale measure of Z from Eq. 13.20. Thus, we have found two processes which are martingales with respect to this measure, namely $Z(t)$ and $E_t^Y\left[V(T)\right]$.

But we want more. We want the (appropriately normalized) price $V(t)$ itself to be a martingale, not merely the expectation of the payoff profile $V(T)$. In the discrete case (and in the continuous case for U as well), this was accomplished by considering the *normalized* prices. We therefore consider instead of V the payoff profile normalized with Y, $V(T)/Y(T)$. The expectation (taken at time t with respect to the martingale measure of Z) of this function

$$E(t) := E_t^Y\left[\frac{V(T)}{Y(T)}\right]$$

is, because of the Tower Law, also a martingale, which we denote by $E(t)$ in what follows. Furthermore, the payoff of V at maturity $t = T$ is exactly replicated by the product $Y(t)E(t)$ since

$$Y(T)E(T) = Y(T)E_T^Y\left[\frac{V(T)}{Y(T)}\right] = Y(T)\frac{V(T)}{Y(T)} = V(T) \tag{13.26}$$

It is now clear how the existence of a replicating portfolio can be established through an application of the martingale representation theorem: the martingale in question is $E(t)$, the expectation of the normalized derivative price at maturity, and as our second martingale we take $Z(t)$ from Eq. 13.20, the normalized price of the initially selected tradable instrument U. The martingale representation theorem now states that (if the volatility of Z is always non-zero) the process $E(t)$ differs from the process $Z(t)$ only by a previsible process α_t:

$$dE = \alpha_t dZ \tag{13.27}$$

We use this previsible process now to construct a portfolio consisting of α_t of the instrument U and β_t of the numeraire instrument Y as was done in Eq. 13.3. This is always possible since α_t is previsible by the Martingale Representation Theorem and both U and Y are tradable.

$$\Pi(t) = \alpha_t U(t) + \beta_t Y(t) \ . \tag{13.28}$$

This portfolio should equal $Y(t)E(t)$ for all times $t \leq T$ since, according to Eq. 13.26, it then replicates the payoff of V upon its maturity exactly, i.e., when $t = T$. From this condition, we can derive the number β_t of numeraire instruments required for the replicating portfolio:

$$Y(t)E(t) = \Pi(t) = \alpha_t U(t) + \beta_t Y(t)$$

$$\Longleftrightarrow \qquad \beta_t = E(t) - \alpha_t \frac{U(t)}{Y(t)} = E(t) - \alpha_t Z(t) \ . \tag{13.29}$$

We have thus established the existence of a replicating portfolio. It remains to show that this portfolio is *self financing*, because only if no injection or withdrawal of capital is required throughout the lifetime of the derivative can we deduce the equality of the portfolio's value and the value of the instrument V. To this end, we consider the total change in the value of the portfolio in light of Eq. 13.14:

$$
\begin{aligned}
d\Pi &= d\,(YE) \\
&= E dY + Y dE + dE dY \\
&= E dY + Y\alpha dZ + \alpha dZ dY \\
&= [\beta + \alpha Z]\, dY + \alpha Y dZ + \alpha dZ dY
\end{aligned}
$$

$$= \alpha \left[Z dY + Y dZ + dZ dY \right] + \beta dY$$
$$= \alpha \, d\underbrace{(ZY)}_{U} + \beta \, dY \,,$$

where in the second equality the (stochastic) product rule for stochastic integrals has been applied, in the third equality the Martingale Representation Theorem in form of Eq. 13.27, and in the fourth, Eq. 13.29 in the form $E(t) = \beta_t + \alpha_t Z(t)$ has been used. In the last equation the (stochastic) product rule has been applied again. The equation now states that the total change in the portfolio defined in Eq. 13.28 results solely from the change in price of the instruments U and Y and *not* from any adjustment of the positions α or β:

$$d\Pi(t) = \alpha_t dU(t) + \beta_t dY(t) \,.$$

Thus, via Eq. 13.14 the portfolio is self financing. The value of the portfolio is by construction $\Pi(t) = Y(t)E(t)$ for all times. According to Eq. 13.26, this replicates the payoff profile $V(T)$ exactly at time T. The value of the portfolio must therefore equal that of the derivative for all previous times as well:

$$V(t) = \Pi(t) = Y(t)E(t) = Y(t)\mathrm{E}_t^Y \left[\frac{V(T)}{Y(T)} \right] \qquad (13.30)$$

and thus

$$\frac{V(t)}{Y(t)} = \mathrm{E}_t^Y \left[\frac{V(T)}{Y(T)} \right] \qquad (13.31)$$

Therefore, the normalized price of the tradable financial instrument V is a martingale in the *same* probability measure with respect to which the normalized price of the instrument U is a martingale. Since V was chosen arbitrarily, this implies that the normalized price of *all* tradable instruments are martingales with respect to the *same* probability measure.

Furthermore, the process α_t can be calculated explicitly. Equation 13.27 states that α_t is the change of E per change in Z or in other words the derivative of E with respect to Z. Using Eqs. 13.20 and 13.30, both Z and E can be expressed in terms of the prices of tradable instruments (known at time t)

$$\alpha_t = \frac{\partial E(t)}{\partial Z(t)} = \frac{\partial \left[V(t)/Y(t) \right]}{\partial \left[U(t)/Y(t) \right]} \,. \qquad (13.32)$$

The process $\alpha(t)$ corresponds to the sensitivity Δ introduced in Chap. 12. It follows that Eq. 13.29 can be applied to calculate $\beta(t)$ at time t explicitly as well.

$$\beta_t = E(t) - \alpha_t Z(t) = \frac{V(S,t)}{Y(t)} - \frac{U(S,t)}{Y(t)} \frac{\partial \left[V(S,t)/Y(t)\right]}{\partial \left[U(S,t)/Y(t)\right]} \, .$$

We have thus accomplished our goal, having shown that the normalized prices V/Y of all tradable instruments are martingales with respect to the martingale measure of $Z = U/Y$. The only question remaining is whether this measure is unique or whether several such measures may exist. To answer this question we apply yet another theorem from stochastic analysis which states that for *complete markets*[11] the martingale measure obtained above is unique. The theorem [89] is stated explicitly here[12]:

Theorem 4 (Harrison-Pliska) *A market consisting of financial instruments and a numeraire instrument is arbitrage free if and only if there exists a measure, equivalent to the real world measure, with respect to which the prices of all financial instruments normalized with the numeraire instrument are martingales. This measure is unique if and only if the market is complete.*

Summary
At this stage, it is helpful to summarize what has been done in this section. The summary corresponds to the summary at the beginning of the Sect. 13.4 (which was done for discrete time steps).

- We select a tradable instrument Y as the numeraire and another tradable instrument U which has S as an underlying (if S itself is tradable, S can be chosen directly).
- We then find the probability measure for which $Z = U/Y$ is a martingale. The *Girsanov-Theorem* guarantees that this is always possible via a suitable drift transformation as long as a technical boundedness condition is satisfied.
- The *Martingale Representation Theorem* and the *Tower Law* enable the construction of a *self financing portfolio* composed of the instruments U and Y which replicates the payoff profile at maturity of any arbitrary, tradable

---

[11]A market is called *complete* if there exists a replicating portfolio for each financial instrument in the market.

[12]As always, the term "if and only if" means that one follows from the other *and vice versa*.

instrument V having S as an underlying. The value of this replicating portfolio is given by Eq. 13.30 where the expectation is taken with respect to the martingale measure of $Z = U/Y$.

- This portfolio must be equal to the value of the derivative $V(t)$ for all times before maturity if the market is arbitrage free. This means that, according to Eq. 13.31, the normalized price V/Y with respect to the martingale measure of Z is likewise a martingale. Thus, having obtained (via Girsanov) a martingale measure for Z, the normalized price V/Y of all other tradable instruments are martingales with respect to this same measure.

- Finally, the *Harrison-Pliska Theorem* states that this measure is unique in complete markets: in complete markets there exists for each numeraire instrument *one single* measure with respect to which all tradable instruments normalized with this numeraire are martingales.

13.5 The Drift

With respect to the martingale measure, the price processes of all instruments normalized by the numeraire instrument Y are drift-free. The *expected* changes in the normalized prices of tradable instruments are thus exactly equal to zero. What can we say about the process of the *underlying* with respect to this measure? The model Eq. 2.19 was set up to describe the underlying in the *real* world. As we have seen however, the *valuation* of financial instruments is accomplished in a world governed by the probability with respect to which normalized prices of tradable instruments are martingales, i.e., processes of the form indicated in Eq. 13.16. Hence, it is important to know how the *underlying* process is transformed when the real probability measure is transformed into the martingale measure.

Not only do we know that a martingale measure exists (and is unique in a complete market) for $Z = U/Y$ with respect to which Z can be represented by a process of the form in Eq. 13.16. We also know from Eqs. 13.23 and 13.24 the explicit relationships between the variables in the real world and the world governed by the martingale measure. We now make use of Eq. 13.23 in particular, to express the underlying S as given by the process 13.17 with respect to the Brownian motion dW in the real world in terms of the Brownian

motion $d\widetilde{W}$ with respect to the martingale measure (using the explicit form for a_u from Eq. 13.18):

$$dS = a\,dt + b\,dW$$

$$= adt + b\left\{ d\widetilde{W} - \frac{a_u + \left[\frac{n^2}{Y^2} - \frac{m}{Y}\right]U - b\frac{n}{Y}\frac{\partial U}{\partial S}}{b\frac{\partial U}{\partial S} - n\frac{U}{Y}}dt \right\}$$

$$= adt - \frac{\frac{\partial U}{\partial S}a + \frac{\partial U}{\partial t} + \frac{1}{2}\frac{\partial^2 U}{\partial S^2}b^2 + \left[\frac{n^2}{Y^2} - \frac{m}{Y}\right]U - b\frac{n}{Y}\frac{\partial U}{\partial S}}{\frac{\partial U}{\partial S} - \frac{n}{b}\frac{U}{Y}}dt + bd\widetilde{W}$$

$$= \frac{b\frac{n}{Y}\frac{\partial U}{\partial S} - a\frac{n}{b}\frac{U}{Y} - \frac{\partial U}{\partial t} - \frac{1}{2}\frac{\partial^2 U}{\partial S^2}b^2 - \left[\frac{n^2}{Y^2} - \frac{m}{Y}\right]U}{\frac{\partial U}{\partial S} - \frac{n}{b}\frac{U}{Y}}dt + bd\widetilde{W}\ .$$

$$(13.33)$$

This equation explicitly specifies the underlying process with respect to the martingale measure.

Note that in the transition from the real world measure to the martingale measure, only the drift of the underlying has changed and *not* the volatility; the coefficient of the Brownian motion remains $b(S,t)$. This is not merely coincidence but a natural consequence of the Girsanov Theorem which intuitively states that a transformation between two equivalent probability measures effects nothing more than a change in the drift.

We should further note that the Harrison-Pliska Theorem states that the martingale measure in complete markets (we will from now on always assume that the market is complete, if nothing else is explicitly stated) is *unique.* This implies that the drift of S, used in the valuation of financial instruments on S is *unique* as well, up to the choice of the numeraire instrument Y.

In the last equality in 13.33 the drift $a(S,t)$ in the real world has been written to share a common denominator with the second term $\sim dt$. In doing so, we observe that the terms $\sim \frac{\partial U}{\partial S}a$ cancel each other and the drift in the real world enters into the equation corresponding to the martingale measure only in the form of $\frac{n}{b}\frac{U}{Y}a$. For numeraire instruments Y satisfying Eq. 13.19 with $n(Y,t) = 0\ \forall Y, t$, i.e., for numeraires with processes of the form

$$dY(t) = m(Y,t)\,dt \qquad\qquad (13.34)$$

the drift a of the *real* world disappears completely for the martingale measure associated with Y! Regardless of the drift $a(S, t)$ chosen in the model, $a(S, t)$ is completely irrelevant for the valuation of financial instruments if the numeraire chosen is a process of the form specified in Eq. 13.34.

Equation 13.34 by no means implies that Y must be deterministic, since $m(Y, t)$ is not assumed to be deterministic but only *previsible*.[13] This means that at time t the evolution of Y is only known for the step *immediately following* t, but not for later steps.

As has already been mentioned on more than one occasion, the choice of instrument to be used as the normalizing factor (the *numeraire*) is arbitrary but this choice significantly affects whether a specific problem can be solved elegantly or awkwardly. This is analogous to the selection of a suitable system of coordinates when solving problems in physics, for example. We see from Eq. 13.33 that an appropriate choice of numeraire can simplify calculations substantially. The numeraire should always be chosen to be of the form 13.34 for some previsible process m. This is always possible in practice and all numeraire instruments in this book satisfy this property.[14] For such numeraires, m/Y in Eq. 13.33 is precisely the *yield of the numeraire instrument*, since Eq. 13.34 obviously implies:

$$m = \frac{dY}{dt} \implies \frac{m}{Y} = \frac{1}{Y}\frac{dY}{dt} = \frac{d\ln Y}{dt} . \tag{13.35}$$

Using this equation and $n = 0$, the drift transformation equation 13.23 becomes

$$d\widetilde{W} = dW + \frac{a_U - \frac{d\ln Y}{dt}U}{b\frac{\partial U}{\partial S}}dt$$

$$= dW + \frac{1}{b}\left[a_U\left(\frac{\partial U}{\partial S}\right)^{-1} - \frac{d\ln Y}{dt}\left(\frac{\partial \ln U}{\partial S}\right)^{-1}\right]dt \tag{13.36}$$

---

[13]A previsible process is a *stochastic* process whose current value can be determined from information available at the previous time step. Intuitively, it is a stochastic process "shifted back" one step in time.

[14]This will be shown below explicitly for all numeraire instruments used.

and the underlying process in Eq. 13.33 reduces to

$$dS(t) = \tilde{a}(S, t)\, dt + b(S, t) d\widetilde{W} \qquad \text{mit}$$

$$\tilde{a}(S, t) = \left(\frac{\partial U(S, t)}{\partial S}\right)^{-1} \left[U(S, t)\frac{d \ln Y(t)}{dt} - \frac{\partial U(S, t)}{\partial t} - \frac{b^2(S, t)}{2}\frac{\partial^2 U(S, t)}{\partial S^2} \right].$$

$$(13.37)$$

Only the tradable instrument U and the numeraire Y (and their respective derivatives) and the "volatility" $b(S, t)$ of the underlying process S appear in this expression. As already noted, the real underlying drift $a(S, t)$ has disappeared completely.

Since the prices of financial instruments must obviously be independent of the method used to compute them, the following theorem holds irrespective of the choice of numeraire:

Theorem 5 *Suppose there exists a numeraire Y in an arbitrage-free market satisfying Eq. 13.34 with a previsible process $m(Y, t)$. Then the drift of the underlying in the real world is irrelevant to the prices of financial instruments. Arbitrage freedom alone determines the prices of financial instruments and not the expectation of the market with respect to the evolution of the underlying.*

In this context, it is interesting to consider the behavior of the *non-normalized* process of the tradable instrument U in the martingale measure. The process 13.18 in the real world is transformed via Eq. 13.36 to

$$dU = a_U dt + \frac{\partial U}{\partial S} b\, dW$$

$$= a_U dt + \frac{\partial U}{\partial S} b \left\{ d\widetilde{W} - \frac{1}{b}\left[a_U \left(\frac{\partial U}{\partial S}\right)^{-1} - \frac{d \ln Y}{dt}\left(\frac{\partial \ln U}{\partial S}\right)^{-1} \right] dt \right\}$$

$$= a_U dt + \frac{\partial U}{\partial S} b\, d\widetilde{W} - \left[a_U - U\frac{d \ln Y}{dt} \right] dt ,$$

and thus

$$dU(S, t) = \frac{d \ln Y(t)}{dt} U(S, t) dt + b(S, t)\frac{\partial U(S, t)}{\partial S} d\widetilde{W} . \qquad (13.38)$$

The drift of a tradable instrument in the martingale measure for a numeraire of the form specified in Eq. 13.34 is thus simply the product of the price of the instrument and the yield of the numeraire!

Theorem 6 *The expected yield (defined as the expected logarithmic price change per time) of a tradable financial instrument in the martingale measure with a numeraire of the form 13.34 is always equal to the yield of the numeraire instrument.*

$$\mathrm{E}_t^Y \left[\frac{dU(S,t)}{U(S,t)} \right] = \frac{d \ln Y(t)}{dt} \, dt \; . \tag{13.39}$$

In Eqs. 13.38 and 13.39, a tradable instrument is denoted by the letter U but these properties naturally hold for all tradable instrument since, as discussed in detail in the previous section, all tradable instruments are martingales with respect to the same probability measure. The instrument U is not essentially different from other tradable instruments in the market.

13.6 The Market Price of Risk

As emphasized several times previously, all tradable instruments U in a complete, arbitrage-free market (normalized with respect to a selected numeraire) have the same martingale measure. This implies that $d\widetilde{W}$ in Eq. 13.23 is always *the same* Brownian motion for this measure. Since the Brownian motion dW of the underlying in the real world is not dependent on the specific financial instrument either, this must hold for the difference $d\widetilde{W} - dW$ as well. The change in drift from dW to $d\widetilde{W}$ in Eq. 13.23 must be *the same* for every tradable instrument U. This implies that for two arbitrary tradable instruments, $U_1(S,t)$ and $U_2(S,t)$ the following must hold:

$$\frac{a_{U_1} + \left[\frac{n^2}{Y^2} - \frac{m}{Y} \right] U_1 - b \frac{n}{Y} \frac{\partial U_1}{\partial S}}{b \frac{\partial U_1}{\partial S} - n \frac{U_1}{Y}} = \frac{a_{U_2} + \left[\frac{n^2}{Y^2} - \frac{m}{Y} \right] U_2 - b \frac{n}{Y} \frac{\partial U_2}{\partial S}}{b \frac{\partial U_2}{\partial S} - n \frac{U_2}{Y}} \; , \tag{13.40}$$

where a_{U_i} denotes the drift of U_i in the *real* world in accordance with Eq. 13.18. The existence of a unique measure with respect to which all

normalized, tradable instruments are martingales satisfying Eq. 13.16 can be formulated equivalently as follows: regardless of the appearance of the drift of the financial instrument in the real world, the combination of this drift with other characteristics of the financial instrument and the numeraire as specified in Eq. 13.40 must be the same for *all* financial instruments. This combination has its own name; it is known as the *market price of risk*.[15] The market price of risk γ_U for an instrument U is defined by

$$
\gamma_U(t) := \frac{a_U + \left[\frac{n^2}{Y^2} - \frac{m}{Y}\right] U - b\frac{n}{Y}\frac{\partial U}{\partial S}}{b\frac{\partial U}{\partial S} - n\frac{U}{Y}} \tag{13.41}
$$

$$
= \frac{1}{b}\left[a_U\left(\frac{\partial U}{\partial S}\right)^{-1} - \frac{d\ln Y}{dt}\left(\frac{\partial \ln U}{\partial S}\right)^{-1}\right]
$$

with a_U as in Eq. 13.18. This expression simplifies further if the numeraire fulfills Eq. 13.34:

$$
\gamma_U(t) = \frac{1}{b}\left[a_U\left(\frac{\partial U}{\partial S}\right)^{-1} - \frac{d\ln Y}{dt}\left(\frac{\partial \ln U}{\partial S}\right)^{-1}\right].
$$

From this definition and Eq. 13.40 it is now immediate that the following theorem holds.

Theorem 7 *The market price of risk is identical for all tradable instruments in a complete, arbitrage-free market.*

Comparing the definition in Eq. 13.41 with Eq. 13.23 and Eq. 13.22 immediately gives

Theorem 8 *The previsible process $\gamma(t)$ in the Girsanov Theorem effecting the drift transformation for the transition from the probability measure in the real world to the martingale measure is the market price of risk.*

Let us consider the process given by Eq. 13.18 representing a tradable instrument U in the real world from this point of view. The drift a_U in the real world can by Definition 13.41 be expressed in terms of the market price

---

[15] The motivation for this name will become clear further below, when we look at certain special cases.

of risk:

$$a_U = \gamma_U \, b \frac{\partial U}{\partial S} + U \frac{d \ln Y}{dt} \, .$$

Substituting this expression into Eq. 13.18 yields the process for a tradable financial instrument in the *real* world, expressed in terms of the market price of risk:

$$dU(S, t) = \left[\gamma_U(t) \, b(s, t) \frac{\partial U(S, t)}{\partial S} + \frac{d \ln Y(t)}{dt} U(S, t) \right] dt + b(S, t) \frac{\partial U(S, t)}{\partial S} dW \, .$$
$$(13.42)$$

The *valuation* of the financial instrument is accomplished not in the real world but in that governed by the martingale measure. In the martingale measure, the instrument U is a process satisfying Eq. 13.38. Comparing this process with Eq. 13.42 directly yields the following "recipe":

Theorem 9 *Setting the market price of risk equal to zero in the expression for the stochastic process (more explicitly in the differential equation which is satisfied by this process) which governs the financial instrument in the real world immediately yields the stochastic process (i.e., the differential equation) which is to be applied in the valuation of this instrument.*

13.7 Tradable Underlyings

The equations in the previous section appear relatively complicated because we have assumed throughout that the underlying S is not necessarily tradable. If the underlying is in fact tradable, it can be used in place of the instrument U directly with the consequence that

$$U = S \Longrightarrow \frac{\partial U}{\partial S} = 1 \, , \quad \frac{\partial^2 U}{\partial S^2} = 0 = \frac{\partial U}{\partial t} = 0 \, . \tag{13.43}$$

The general equation 13.33 for this special case reduces to

$$dS = \frac{b \frac{n}{Y} - a \frac{n}{b} \frac{S}{Y} - \left[\frac{n^2}{Y^2} - \frac{m}{Y} \right] S}{1 - \frac{n}{b} \frac{S}{Y}} dt + bd\widetilde{W} \, .$$

The corresponding Eq. 13.37 for a more appropriate choice of numeraire instrument satisfying Eq. 13.34 further reduces to

$$dS(t) = S(t)\frac{d\ln Y(t)}{dt}\,dt + b(S,t)d\widetilde{W}\,, \qquad (13.44)$$

which, of course, agrees with Eq. 13.38, this equation holding for every tradable instrument. The expectation of the yield (defined as the expected relative price change per time) of a tradable underlying in the martingale measure is thus (as for every tradable instrument) always equal to the yield of the numeraire instrument.

The market price of risk $\gamma_S(t)$ for a tradable underlying is obtained from Eq. 13.41 with 13.43 as

$$\gamma_S(t) = \frac{a + \left[\frac{n^2}{Y^2} - \frac{m}{Y}\right]S - b\,\frac{n}{Y}}{b - n\frac{S}{Y}}\,, \qquad (13.45)$$

where $a = a(S,t)$ is the underlying drift in the real world (see Eq. 13.17). For a well-chosen numeraire of the form 13.34 we get:

$$\gamma_S(t) = \frac{a(S,t)}{b(S,t)} - \frac{S(t)}{b(S,t)}\frac{d\ln Y}{dt}\,.$$

With this, the process Eq. 13.17 in the real world (in consistence with Eq. 13.42) becomes

$$dS(t) = \left[b(S,t)\gamma_S(t) + \frac{d\ln Y(t)}{dt}S(t)\right]dt + b(S,t)\,dW\,. \qquad (13.46)$$

Comparison with Eq. 13.44 again yields the "recipe" (which holds for all tradable instruments and hence for S as well) that the market price of risk must simply be set equal to zero for valuation purposes.

Note that these considerations are valid for a still very general case: In our treatment, it has only been assumed that the underlying is tradable and that the numeraire instrument is of the form 13.34.

13.8 Applications in the Black-Scholes World

Special cases of the general material discussed above have been encountered in Sects. 7.1.3, 7.2.1 and 9.2.3 in various "disguises". The irrelevance of the real world drift became most apparent in Sect. 9.3. There, the requirement that the expectation of the underlying, computed using its stochastic process, must equal the expectation of the underlying resulting from the martingale property[16] led to Eq. 9.25, i.e., an explicit specification of the drift to be used in the valuation. The volatility σ, in contrast, was subject to no such condition and remained the same.

Equation 9.25 holds for a very special case, namely when *(a)* the underlying behaves as in Eq. 2.23 and *(b)* is tradable, *(c)* the numeraire is given by $Y(t) = B(t, T)$ and *(d)* interest rates and volatilities are constant; in short, in the Black-Scholes world. Let us therefore apply the general results of the above sections to this special situation as an example.

Firstly, a zero bond $B(t, T)$ is chosen as the numeraire instrument maturing at some arbitrary future time T. The process for B in the real world is of the form 13.34. For continuous compounding, we have explicitly

$$dY \equiv dB(t, T) = \frac{dB(t, T)}{dt}dt = rB(t, T)dt \implies \frac{d\ln Y(t)}{dt} = r \,.$$
$$(13.47)$$

The special process, Eq. 2.23 (or Eq. 2.24) corresponds to the general process, Eq. 13.17 with the parameters

$$a(S, t) = \widetilde{\mu}S(t) = \left(\mu + \frac{\sigma^2}{2}\right)S(t)$$
$$b(S, t) = \sigma S(t) \,.$$

Since the underlying is tradable, Eq. 13.44 can be applied directly to obtain the underlying process in the martingale measure

$$dS(t) = rS(t)\,dt + \sigma S(t)d\widetilde{W} \,.$$

---

[16]There, instead of the general formulation "expectation with respect to the martingale measure", the expression "risk neutral expectation" was used.

This process and not 2.23 is to be used in the valuation of the financial instrument. Comparison of this process with Eq. 2.23 shows that the choice

$$\mu = r - \sigma^2/2 \quad \text{or equivalently} \quad \widetilde{\mu} = r \tag{13.48}$$

for the drift transforms the real world process directly to the process to be used for pricing, in agreement with Eq. 9.25 (dividend yield equal to zero). The market price of risk of the underlying is simply

$$\gamma_S(t) = \frac{a(S,t) - rS(t)}{b(S,t)} = \frac{\widetilde{\mu} - r}{\sigma} \tag{13.49}$$

in this special case and is exactly equal to zero for $\widetilde{\mu} = r$. In fact "setting the market price of risk equal to zero" is equivalent to "choosing the correct drift for pricing".

Equation 13.49 provides the motivation for the name "market price of risk". In the special case considered here the underlying drift in the real world is simply $a(S,t) = \widetilde{\mu}S(t)$. The expectation of the underlying-*yield* is thus $a(S,t)/S(t) = \widetilde{\mu}$. This implies that $\widetilde{\mu} - r$ represents the excess yield above the risk-free rate, which is expected from the underlying in the real world. If the volatility σ is viewed as a measure of the risk of the underlying, then the market price of risk γ_S (at least in the context of this special case) can be interpreted as the excess yield above the risk-free rate *per risk unit σ* which the market expects from the underlying. This is, so to speak, the price (in the form of an excess yield above the risk-free rate) which the market demands for the risk of investing in the underlying. In the real world, the market is by no means risk neutral, but rather expects higher yields for higher risks; the market price of risk illustrates this clearly.

Note that the market price of risk in Eq. 13.49 is identical to the so called *Sharpe Ratio* [138, 139] heavily used in asset management and portfolio optimization since more than 50 years. Very generally, the Sharpe Ratio is defined as the expected excess return (above the risk free rate) of an investment divided by the investment risk (measured as its volatility).

$$\text{Sharpe Ratio} \equiv \frac{R - r_f}{\sigma}$$

Our case above corresponds to an investment in a single risky asset, namely in the risk factor S. As we have shown in Eq. 2.32, the drift $\widetilde{\mu}$ appearing in Eq. 13.49 is the expected return for linear compounding. Since in asset

management returns are usually defined as relative (as opposed to logarithmic) price changes, linear compounding is indeed applicable (see Eq. 2.31). Thus, $\widetilde{\mu}$ in Eq. 13.49 exactly corresponds to the expected return R used in portfolio management for an investment in S. And therefore the Sharpe Ratio and the market price of risk are the same thing, see also [49].

To conclude this section we will now show that we can actually use all of this information about martingales and drifts to really calculate something. We consider below only payoff profiles of path-independent instruments. These are payoff profiles which depend solely on the value of the underlying at maturity T and not on the path taken by the process S between t and T. For such processes

$$V(S, T) = V(S(T), T) \; .$$

holds. Therefore we only need the distribution of S at time T (and not the distribution of all paths of S between t and T). Choosing the zero bond as the numeraire instrument, $Y(t) = B(t, T)$, Eq. 13.31 for the price V of a financial instrument becomes

$$V(S, t) = Y(t)\mathrm{E}_t^Y \left[\frac{V(S, T)}{Y(T)} \right] = B(t, T)\mathrm{E}_t^Y \left[V(S(T), T) \right] \; .$$

since $B(T, T) = 1$. We again model the underlying with the simple process in Eq. 2.23. The associated underlying process over a *finite* time interval of length $\delta t = T - t$, i.e., the solution of the stochastic differential equation 2.23 for S, has already been given in Eq. 2.28 for an arbitrary drift μ, and thus for an arbitrary probability measure, namely

$$S(T) = S(t) \exp\left[\mu(T - t) + \sigma W_{T-t}\right] \quad \text{mit} \;\; W_{T-t} \sim \mathrm{N}(0, T - t)$$

$$= S(t) \exp\left[x\right] \quad \text{with} \;\; x \sim \mathrm{N}\left(\mu(T - t), \sigma^2(T - t)\right) \; . \qquad (13.50)$$

The distribution of $S(T)$ is therefore $S(t)$ multiplied by the exponential of the normal distribution with expectation $\mu(T - t)$ and variance $\sigma^2(T - t)$. Using this, the expectation of the function

$$V(S(T), T) = V(S(t)e^x, T) =: g(x)$$

can be computed explicitly:

$$E_t [V(S(T), T)] =: E_t [g(x)] = \int_{-\infty}^{\infty} g(x) p(x) dx$$

with the probability density of the normal distribution (see Eq. A.46)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2(T-t)}} \exp\left[-\frac{(x - \mu(T-t))^2}{2\sigma^2(T-t)}\right].$$

Thus

$$E_t [V(S(T), T)] = \frac{1}{\sqrt{2\pi\sigma^2(T-t)}} \int_{-\infty}^{\infty} V\left(S(t)e^x, T\right) \exp\left[-\frac{(x - \mu(T-t))^2}{2\sigma^2(T-t)}\right] dx.$$

This is the expectation of the payoff profile with respect to a probability measure associated with μ, for instance with respect to the probability measure in the real world if μ represents the drift in the real world. To allow the use of this expectation for the *valuation* of the instrument V, it must be computed with respect to the (in a complete market, unique) *martingale* measure. This is accomplished through the choice of drift in accordance with Eq. 13.48. The price of V then becomes

$$V(S, t) = B(t, T)E_t^Y [V(S(T), T)]$$

$$= \frac{B(t, T)}{\sqrt{2\pi\sigma^2(T-t)}} \int_{-\infty}^{\infty} V\left(S(t)e^x, T\right) \exp\left[-\frac{\left[x - \left(r - \frac{1}{2}\sigma^2\right)(T-t)\right]^2}{2\sigma^2(T-t)}\right] dx.$$

$$(13.51)$$

This equation holds in complete generality for every financial instrument with a non-path dependent payoff profile $V(S(T), T)$ on an underlying S of the form specified in Eq. 2.23.

To be more specific, we now use the payoff profile of a plain vanilla call as an example:

$$V(S(T), T) = \max\{S(T) - K, 0\} = \max\{e^x S(t) - K, 0\}.$$

with x as defined in Eq. 13.50. This payoff profile is only non-zero when $e^x S(t) \geq K$, or equivalently when $x \geq \ln(K/S(t))$. Equation 13.51 for the

price of the call is then

$$
V(S,t) = \frac{B(t,T)}{\sqrt{2\pi\sigma^2(T-t)}} \int_{\ln(K/S(t))}^{\infty} \left[e^x S(t) - K\right] \exp\left[-\frac{\left(x - \left(r - \frac{\sigma^2}{2}\right)(T-t)\right)^2}{2\sigma^2(T-t)}\right] dx .
$$

This (with $\mu = r - \sigma^2/2$) is in complete agreement with Eq. 8.4, for example. In connection with Eq. 8.4 it has been demonstrated how this integral can be computed explicitly. The result is the famous Black-Scholes equation 8.6.

14

Check for
updates

Interest Rates and Term Structure Models

So far, with only few exceptions (e.g. Sect. 8.3.3), we have considered interest rates as being deterministic or even constant. This directly contradicts to the simple existence of interest rate options. If interest rates were deterministic and hence predictable with certainty for all future times, we would know for certain if a given interest rate option is either worthless (because it is out-of-the-money) or otherwise would be a simple forward contract (if it is in-the-money) and could be priced by simple cash flow discounting.

Nevertheless, interest rates may assumed to be deterministic, if the derivative underlying are shares (or some other asset class like FX or commodities) and if the term to expiry is rather short (e.g. less than 3 years), since in such a case, the value of an equity derivative is much more sensitive to changes of the underlying share price than to changes of the interest rate level. Also, the volatility of share prices is often much higher than interest rate volatilities.[1] Empirical studies (e.g. see [109, 145]) also demonstrate that different effects of "false" assumptions (in particular, the assumed equality of forward and futures prices) tend to cancel out each other.

Methods for interest rate modeling and valuation of interest rate options have been for more than 20 years and area of active research in financial mathematics. *Term structure models* are used to model the stochastic changes observed in interest rate curves, similar to the way stock prices and exchange rates are modeled with an underlying stochastic process $S(t)$.[2]

---

[1]As we will see later, e.g. in Sect. 30.1 ff, there is more beyond Black-Scholes.

[2]The Black-76 model for simple interest rate options can be derived as a special case of the more complex Heath-Jarrow-Morton term structure model.

In contrast to a stock price or exchange rate $S(t)$, an interest rate $R(t, T)$ depends on *two* time variables, namely on the time t at which the interest rate is being considered *and* on the maturity T of the term over which the interest is to be paid. Usually, interest is paid at a higher rate for longer terms $\tau := T - t$ (i.e., later maturities) than for shorter ones. For a fixed time t (today), the set of all interest rates for the various maturity dates T form a curve called the *term structure*. Theoretically, each point in this curve is a stochastic variable associated with a single maturity. The term structure is thus theoretically a continuum of infinitely many stochastic variables. From a practical point of view, these processes are naturally very strongly correlated; the interest rate over a term of 3 years and 1 day is (almost) the same as that over a term of 3 years and 0 days and so on. In practice, market participants therefore consider only finitely many terms whose lengths lie well distinct from one another (for example, terms $\tau = T - t$ of 1 day, 1 month, 3 months, 6 months, 9 months, 1 year, 2 years, 3 years, 5 years, 10 years and so on), depending on which liquid market quotes have been used to build up the curve. Motivated by the results of principle component analysis (see for instance Sect. 34.2) which show that well over 90% of the dynamics of the yield curve can be explained by just one or two stochastic factors, most term structure models go a step further and reduce the number of factors driving the stochastic evolution of the entire term structure to just a few stochastic variables (e.g. 1 or 2 for simple term structure models). These models are referred to as 1-factor or 2-factor, for example, depending on the number of stochastic variables.

14.1 Instantaneous Spot Rates and Instantaneous Forward Rates

Many simple 1-factor models are based on a stochastic process of the form specified in Eq. 2.19, where the stochastic factor is usually assumed to be a very short term rate, called *instantaneous* interest rate. These rates take the form of either an *instantaneous spot rate* (also called the *instantaneous short rate*) or an *instantaneous forward rate*.[3] The terms $\tau = T - t$ belonging to these rates are infinitesimally short, i.e., we consider the limit $\tau \to 0$, or equivalently $T \to t$.

---

[3]This approach is to be differentiated from *market rate models* (which are also used to be called *Brace-Gatarek-Musiela models* or (*BGM models* for short). For such models, forward interest rates over longer periods (e.g. the 3-months LIBOR rate), but distinct start dates, are modeled (see Sect. 14.13 and [19]).

In the following discussion, we adopt the convention of continuous compounding usually observed in the literature. This allows the instantaneous rates to be defined quite easily. The *instantaneous spot rate $r(t)$* is defined by

$$e^{-r(t)dt} := \lim_{dt \to 0} B(t, t + dt)$$

$$r(t) = -\lim_{dt \to 0} \frac{\ln B(t, t + dt)}{dt} .$$
(14.1)

The *instantaneous forward rate $f(t, T)$* is defined by

$$e^{-f(t,T)dT} := \lim_{dT \to 0} B(T, T + dT \mid t) = \lim_{dT \to 0} \frac{B(t, T + dT)}{B(t, T)}$$

$$f(t, T) = -\lim_{dT \to 0} \frac{1}{dT} \ln \frac{B(t, T + dT)}{B(t, T)} = -\lim_{dT \to 0} \frac{\ln B(t, T + dT) - \ln B(t, T)}{dT} ,$$

where Eq. 2.7 has been used. Thus

$$f(t, T) = -\frac{\partial \ln B(t, T)}{\partial T} .$$
(14.2)

Integrating this equation over dT and making use of the fact that $B(t, t) = 1$ yields

$$\int_t^T f(t, s)ds = -\int_t^T \frac{\partial \ln B(t, s)}{\partial s} ds = -\ln B(t, T) + \ln \underbrace{B(t, t)}_{1} = -\ln B(t, T) ,$$

and therefore

$$B(t, T) = \exp\left[-\int_t^T f(t, s)ds \right] .$$
(14.3)

This, together with Eq. 2.7, implies that the forward rate over a finite time interval of length $T' - T$ is the *average* of the instantaneous forward rates over this interval:

$$R(T, T' \mid t) = -\frac{1}{T' - T} \ln \frac{B(t, T')}{B(t, T)} = \frac{1}{T' - T} \int_T^{T'} f(t, s)ds .$$
(14.4)

The definitions of instantaneous spot and forward rate imply that both are interest rates with continuous compounding (see also Table 2.5).

The relation between the forward rates and zero bond yields (spot rates over finite time intervals of length $T - t$) can be established as well by inserting the explicit form of the discount factor for continuous compounding into Eq. 14.2 and taking the derivative with respect to T:

$$f(t, T) = -\frac{\partial \left(\ln \exp \left[-R(t, T)(T - t) \right] \right)}{\partial T} = \frac{\partial \left[R(t, T)(T - t) \right]}{\partial T} \, ,$$

And after application of the product rule

$$f(t, T) = R(t, T) + (T - t)\frac{\partial R(t, T)}{\partial T} \, . \tag{14.5}$$

The forward rates are greater than the spot rates for $\partial R(t, T)/\partial T > 0$, i.e., for term structures (interest rate term structure = spot rates $R(t, T)$ as a function of T) whose values increase with T.

Finally, we make note of the relationship between the instantaneous forward rate and the *instantaneous* spot rate. This is simply

$$r(t) = \lim_{T \to t} f(t, T) \, . \tag{14.6}$$

In anticipation of the following sections we stress here that all of the above equations hold in any arbitrary probability measure, i.e., irrespective of any choice of numeraire, since they have been derived directly from the definitions of the instantaneous interest rates.

14.2 Important Numeraire Instruments

As was shown in Chap. 13 in great detail, the value V of an arbitrary, tradable interest rate instrument[4] normalized with an arbitrary, tradable financial instrument Y is a process $Z = V/Y$ which, according to Eq. 13.15, is a martingale

$$Z(t) = \mathrm{E}_t^Y \left[Z(u) \right] \qquad \forall u \geq t \, .$$

---

[4]In spot rate models, every interest rate instrument can be interpreted as a derivative V on the underlying $S(t) = r(t)$ or $S(t) = \ln r(t)$.

The martingale measure at time t with respect to which the expectation E_t^Y is calculated depends on the choice of the numeraire instrument Y. In principle, any arbitrary tradable instrument can be used as the *numeraire*. An appropriate choice of numeraire, however, is a deciding factor in enabling an elegant solution of specific problems to be found. Two numeraire instruments are particularly popular (not only for term structure models but for other models as well). These will be introduced in the following sections.

14.2.1 The Risk-Neutral Measure

We define $\beta(t_0, t)$ as the value of a *bank account* or *money market account*. This is the value one monetary unit (for example, 1 euro) has at time t, if it was invested at a time $t_0 < t$ and was subsequently always compounded at the current spot rate, with the interest earnings being immediately reinvested in the same account at the current spot rate. Intuitively, we could imagine reinvesting ever decreasing interest payments earned over ever shorter interest periods, the number of interest periods finally approaching infinity. The value of such an account can be written in terms of the instantaneous spot rate as

$$\beta(t_0, t) = \exp\left[\int_{t_0}^{t} r(s)ds\right] . \tag{14.7}$$

In the *risk-neutral measure* this bank account, Eq. 14.7, is used as the numeraire.

$$Y(t) = \beta(t_0, t) = \exp\left[\int_{t_0}^{t} r(s)ds\right] \qquad \text{for arbitrary } t_0 \leq t .$$

This numeraire has the advantage of satisfying the important property specified in Eq. 13.34. Since

$$\frac{dY(t)}{dt} = r(t)Y(t) .$$

we have

$$dY(t) = Y(t + dt) - Y(t) = \left(e^{r(t)dt} - 1\right)Y(t) \approx r(t)Y(t)dt$$

Both $r(t)$ and $Y(t)$ are known at time t (even if they are stochastic variables and as such not yet known for the future time $t + dt$). This implies that the process $m(t) = r(t)Y(t)$ is previsible as required in Eq. 13.34.

With this numeraire the martingale property Eq. 13.15 becomes

$$\frac{V(t)}{\beta(t_0, t)} = E_t^\beta \left[\frac{V(u)}{\beta(t_0, u)} \right] \quad \forall u \geq t \geq t_0 .$$

Note that the initial time point t_0 in the money account can be chosen at will. For every t_0 we obtain another, different risk-neutral measure. Setting $t_0 = t$ and using $\beta(t, t) = 1$ reduces the above expression to

$$V(t) = E_t^\beta \left[\frac{V(u)}{\beta(t, u)} \right] = E_t^\beta \left[e^{-\int_t^u r(s)ds} V(u) \right] \quad \forall u \geq t , \tag{14.8}$$

which directly yields the price of the financial instrument. In words:

> With respect to the risk-neutral measure, today's value of a financial instrument is equal to the expectation of the discounted future value.

This is *not* the same as the discounted future expectation. In this measure, the discounting is performed *first* and *then* the expectation is calculated. Discounting means *division* by the numeraire.

We take as an example, the value at time t of a zero-coupon bond with maturity u, i.e., we set $V(t) = B(t, u)$ and consequently $V(u) = B(u, u) = 1$, in Eq. 14.8 to obtain

$$B(t, u) = E_t^\beta \left[\frac{1}{\beta(t, u)} \right] = E_t^\beta \left[e^{-\int_t^u r(s)ds} \right] . \tag{14.9}$$

The bond price is the expectation with respect to the risk neutral measure of the reciprocal of the bank account. Comparing this with Eq. 14.3, which always holds, yields the relationship between the instantaneous forward rates and the future instantaneous spot rates with respect to the risk-neutral measure

$$e^{-\int_t^u f(t,s)ds} = E_t^\beta \left[e^{-\int_t^u r(s)ds} \right] .$$

The *future price* of an instrument V is $V(u)$ for $u > t$ and is unknown at time t. It is well known that this must be distinguished from the *forward price* $V(t, u)$, which is known at time t since it follows from arbitrage arguments (see Eq. 6.1). If the instrument under consideration pays no dividends between t and u, the forward price is

$$V(t, u) = \frac{V(t)}{B(t, u)} \quad \text{with } u > t .\tag{14.10}$$

The forward price, Eq. 14.10, of an instrument with respect to the risk-neutral measure is given by

$$\frac{V(t)}{B(t, u)} = \frac{1}{B(t, u)} \mathrm{E}_t^{\beta}\left[\frac{V(u)}{\beta(t, u)}\right] = \mathrm{E}_t^{\beta}\left[e^{-\int_t^u (r(s)-f(t,s))ds} V(u)\right] \quad \forall u \geq t ,$$
$$\tag{14.11}$$

where Eq. 14.3 was used, exploiting the fact the instantaneous forward rates are by definition known at time t and thus can be included in or taken out of the expectation operator as desired.

14.2.2 The Forward-Neutral Measure

For the *forward-neutral measure* or T-emphterminal measure (to emphasize the fixed end date T), a zero bond is used as the numeraire (see Eq. 14.3)

$$Y(t) = B(t, T) = \exp\left[-\int_t^T f(t, s)ds\right] \quad \text{for arbitrary } T > t .$$

This numeraire has the property prescribed in Eq. 13.34 as well:

$$\frac{dY(t)}{dt} = f(t, t)Y(t) = r(t)Y(t) .$$

Both $r(t)$ and $Y(t)$ are known values at time t, implying the previsibility of $m(t) = r(t)Y(t)$ as required in Eq. 13.34.

With this choice, the martingale property equation 13.15 becomes

$$\frac{V(t)}{B(t, T)} = \mathrm{E}_t^T\left[\frac{V(u)}{B(u, T)}\right] \quad \forall T \geq u \geq t .$$

The index T of E represents here always the maturity of the zero bond numeraire. Observe that T can be selected arbitrarily. For each choice of T, we obtain a different normalizing factor and thus another forward-neutral measure. Setting $T = u$ and using the fact that $B(u, u) = 1$ we obtain the following price for a financial instrument

$$V(t) = B(t, u) \mathrm{E}_t^u [V(u)] = e^{-\int_t^u f(t,s)ds} \mathrm{E}_t^u [V(u)] \quad \forall u \geq t . \qquad (14.12)$$

> Today's value of a financial instrument is equal to the discounted expectation of its future value taken with respect to the forward neutral measure.

The expectation is *first* taken with respect to this measure and *then* discounted. Discounting means *multiplication* by the numeraire.

The forward price, Eq. 14.10, for an interest rate instrument is, with respect to this measure, exactly equal to the expected future price, hence the name "forward-neutral"

$$\frac{V(t)}{B(t, u)} = \mathrm{E}_t^u [V(u)] \quad \forall u \geq t . \qquad (14.13)$$

14.3 The Special Case of Deterministic Interest Rates

The reader may be somewhat confused, since in earlier chapters (see in particular Eq. 9.20), prices of financial instruments were always calculated by discounting the future expectation using $B(t, T)$, which, from the above discussion, would indicate that the *forward*-neutral measure was used in the calculations. The measure was, however, always referred to as the *risk*-neutral measure. In those chapters however, interest rates were always assumed to be completely deterministic (or in many cases even constant). We will now show that for deterministic interest rates, the forward-neutral and the risk-neutral measures are identical.

Consider a portfolio consisting of a zero bond $B(t, T)$ and a loan made at time t to finance the purchase of the zero bond. The interest rate on this loan is floating and is always equal to the current spot rate for each interest period δt. After one such period δt, the loan debt will have grown to $B(t, T) \exp(r(t)\delta t)$. The portfolio thus constructed has no value at time t

and the evolution of the entire portfolio over one time step δt is known exactly at time t, and therefore involves no risk. Because arbitrage is not possible and no risk is taken, the portfolio must have no value for all later times as well. Thus, at time $t + \delta t$ the value of the zero bond must be equal to the value of the loan

$$B(t + \delta t, T) = B(t, T) \exp(r(t)\delta t) \, .$$

So far, everything is as it was as for stochastic interest rates. On the right-hand side of the above equation there appear only terms which are known at time t. The difference comes in the next time step: for deterministic interest rates, the spot rates at all later times $u > t$ are known at time t as well. After *two* interest periods, the credit debt will have grown to $B(t, T) \exp(r(t)\delta t) \exp(r(t + \delta t)\delta t)$ with a *known* interest rate $r(t + \delta t)$. Again, no risk is taken since the rate $r(t + \delta t)$ is already known at time t. Therefore, because the market is arbitrage free, the portfolio must still be worthless at time $t + 2\delta t$ and the loan must therefore still equal the value of the zero bond. Proceeding analogously over n time steps, we obtain the value of the zero bond as

$$B(t + n\delta t, T) = B(t, T) \exp\left[\sum_{i=0}^{n-1} (r(t + i\delta t)\delta t) \right] \, .$$

Taking the limit as $\delta t \to 0$, the value of the bond at time $u := t + n\delta t$ is

$$B(u, T) = B(t, T) \exp\left[\int_t^u r(s)ds \right] \quad \text{with} \quad u \geq t \, .$$

This holds for every $u \geq t$, in particular for $u = T$. Thus, observing that $B(T, T) = 1$, we obtain the price of a zero bond for deterministic interest rates:

$$B(t, T) = \exp\left[-\int_t^T r(s)ds \right] = \frac{1}{\beta(t, T)} \, , \tag{14.14}$$

where in the last step the definition of a bank account, Eq. 14.7, was used. Hence, for deterministic interest rates, the numeraire $B(t, T)$ associated to the forward-neutral measure is equal to the reciprocal $1/\beta(t, T)$ of the numeraire of the risk-neutral measure. Substituting this into Eq. 14.8, we obtain

$$V(t) = E_t^\beta \left[\frac{V(u)}{\beta(t, u)} \right] = E_t^\beta [B(t, u)V(u)] = B(t, u)E_t^\beta [V(u)] \, ,$$

where in the last step we have made use of the fact that $B(t, u)$ is known at time t, is therefore not stochastic and can be factored out of the expectation. The price of a financial instrument must, however, be independent of the measure used in its computation. In consequence, comparison of this equation with Eq. 14.12 immediately yields the equation

$$B(t, u)\mathrm{E}_t^u[V(u)] = V(t) = B(t, u)\mathrm{E}_t^\beta[V(u)] \ ,$$

and thus $\mathrm{E}_t^u[V(u)] = \mathrm{E}_t^\beta[V(u)]$. This implies that both measures are identical if interest rates are deterministic.

The fundamental difference between the general case and deterministic interest rates is that the price of a zero bond in Eq. 14.14 is given by the *future spot rates*. Of course, the general equation 14.3 stating that the price of a zero bond is given by the *current forward rates* continues to hold. We might suspect that this is closely related to the fact that the future instantaneous spot rates must equal the current instantaneous forward rates if interest rates are deterministic. This is in fact the case since the derivative of Eq. 14.14 with respect to T gives

$$r(T) = -\frac{\partial}{\partial T} \ln B(t, T) \ .$$

Comparing this with Eq. 14.2, which holds in general, yields

$$f(t, T) = r(T) \quad \forall t, T \quad \text{with } t \leq T \ .$$

Thus, if interest rates are deterministic, the instantaneous forward rates are indeed equal to the (known) future instantaneous spot rates. Since the right-hand side of this equation is not dependent on t, this must be true for the left-hand side as well. Hence, if interest rates are deterministic, the instantaneous forward rates are independent of the present time t.

14.4 Tradable and Non-tradable Variables

As was emphasized at the end of the Sect. 13.2, both the numeraire Y and the financial instrument V *must* be **tradable** instruments in order for the martingale property Eq. 13.1 to hold. Otherwise, potential arbitrage opportunities cannot be exploited by trading and in this way fail to violate the assumption of an arbitrage-free market. This realization, which may seem trivial from our

modern perspective, was in the past by no means trivial. In fact, there exist pricing methods which assume that non-tradable variables have the martingale property. The mistake made in doing so is then (approximately) corrected after the fact by making a so-called *convexity adjustment* (see Sect. 14.5).

One example of a non-tradable variable is the *yield* of a tradable instrument whose price is a non-linear function of its yield. Take, for example, a zero bond B with lifetime τ. The yield r of the zero bond depends non-linearly on its price (except if the linear compounding convention has been adopted; see Table 2.5). For example, for simple compounding the price of the zero bond is $B = (1 + r\tau)^{-1}$. The zero bond is obviously a tradable instrument. Therefore its future expectation taken with respect to the forward-neutral measure as in Eq. 14.13 must be equal to its forward price.

$$B(T, T + \tau | t) = \mathrm{E}_t^T [B(T, T + \tau)]$$

$$= \mathrm{E}_t^T \left[\frac{1}{1 + r(T, T + \tau)\tau} \right] \geq \frac{1}{1 + \mathrm{E}_t^T [r(T, T + \tau)]\, \tau} \ .$$

The lower-equal sign in the above expression follows from Jensen's inequality.[5] In general, the expectation of the price is *not* equal to the price calculated with the expectation of the yield because of the non-linearity in the relation between the price and the yield. How big the difference is depends on the applied interest rate term structure model.

On the other hand, the forward rate for the time period between T and $T + \tau$ is by definition equal to the yield of a forward zero bond over this period; for linear compounding explicitly:

$$\frac{1}{1 + r_f(T, T + \tau | t)\tau} = \frac{B(t, T + \tau)}{B(t, T)} \equiv B(T, T + \tau | t) \ . \qquad (14.15)$$

Comparing this with the above expression for the forward price of the bond gives

$$\frac{1}{1 + r_f(T, T + \tau | t)\tau} \geq \frac{1}{1 + \mathrm{E}_t^T [r(T, T + \tau)]\, \tau}$$

$$\implies \quad r_f(T, T + \tau | t) \leq \mathrm{E}_t^T [r(T, T + \tau)] \ .$$

---

[5]For convex function f and stochastic variable X, Jensen's inequality states that $f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$. The above equation follows from the convexity (all points $f(x)$ with $a < x < b$ lie below a straight line through points $f(a)$ and $f(b)$) of function $f(x) = 1/(1 + ax)$.

Therefore, the forward rate can *not* simply (i.e. model independently) be equalized with the future expectation of the spot rate taken with respect to the forward-neutral measure. Because of Eq. 14.12, zero bond yields are no martingales with respect to the forward-neutral measure. Since *all tradable* instruments must be martingales with respect to the forward-neutral measure, zero bond yields can *not* be tradable.

What about products depending linearly on the forward rate? With Eq. 14.15 we have

$$r_f(T, T + \tau | t) = \frac{1}{\tau} \left(\frac{B(t, T)}{B(t, T + \tau)} - 1 \right) . \tag{14.16}$$

On the other hand, we could express the zero bond $B(t, T)$ in terms of the expectation in the $(T + \tau)$-forward measure

$$\frac{B(t, T)}{B(t, T + \tau)} = \mathrm{E}_t^{T+\tau} \left[\frac{B(T, T)}{B(T, T + \tau)} \right] . \tag{14.17}$$

Comparison of Eq. 14.16 with Eq. 14.17 yields directly

$$r_f(T, T + \tau | t) = \mathrm{E}_t^{T+\tau} \left[r_f(T, T + \tau | t) \right] = \mathrm{E}_t^{T+\tau} \left[r(T, T + \tau) \right] . \tag{14.18}$$

That's the fundamental reason why, for example, in *forward rate agreements* or *caplets* and *floorlets* (which, of course, are tradable), the difference between the future LIBOR and the strike is not paid out at the LIBOR fixing at the beginning of the relevant interest period but at the *end* of that period (see Sects. 15.2 and 18.6.3,[6] respectively). This has the effect that we are not really dealing with a forward contract (or an option) on a future interest rate, but rather with a forward contract (or an option) on future *zero bonds,* i.e., on *tradable* instruments. The same holds for *swaps, caps* and *floors,* which are nothing other than a series of forward rate agreements, caplets and floorlets, respectively, strung together.

---

[6]In practice, a FRA pays out at maturity (i.e. at the beginning of the interest rate period) the present value of a virtual future payment at the end of the period. This reduces the credit default risk. If, as we did here, credit default risk is neglected, or if the trade is sufficiently collateralized, this does not have a significant impact.

14.5 Convexity Adjustments

According to Eq. 14.12, today's price $V(t)$ of a financial instrument normalized with respect to the forward-neutral numeraire (or if interest rates are considered deterministic) is equal to its discounted future expectation. If the future time for which the expectation is calculated is chosen to be the maturity date T of the instrument, then the price of the instrument is equal to the discounted future expectation of its *payoff profile* $V(T)$. For instruments whose payoff profiles are *linear* functions of the underlying S, i.e., $V(S, T) = a + bS$, the expectation of the *payoff profile* is equal to the payoff profile of the expectation of the *underlying*:

$$E[V(S, T)] = E[a + bS] = \int_{-\infty}^{\infty} [a + bS]\, p(S)dS = a \int_{-\infty}^{\infty} p(S)dS + b \int_{-\infty}^{\infty} S\, p(S)dS$$

$$= a + bE_t^T[S] , \tag{14.19}$$

where p denotes the probability density of the pertinent martingale measure.

Although the prices of most instruments (for example, the zero bond) are non-linear functions of their underlyings, there do exist transactions with linear payoff profiles for which Eq. 14.19 holds, for example forward contracts. The payoff profile of a forward contract with maturity T and delivery price K is known to be $S(T) - K$. The expectation of this payoff profile is simply the expectation of the underlying less the delivery price:

$$E[V(S, T)] = E[S(T) - K] = E[S(T)] - K .$$

If the underlying S of the forward contract is itself a *tradable* instrument (for example, a stock), we can now go a step further. With respect to the forward-neutral normalization (or in the case of deterministic interest rates), the future expectation of any tradable instrument is equal to its current forward price in accordance with Eq. 14.13, i.e., $E_t^T[S(T)] = S(t, T)$. Furthermore, Eq. 14.12 states that with respect to this measure, today's price is equal to the discounted future expectation. So in summary (and in agreement with Eq. 6.5):

$$V(S, t) = B(t, T)E_t^T[V(S, T)]$$
$$= B(t, T)\left(E_t^T[S(T)] - K\right)$$
$$= B(t, T)(S(t, T) - K) .$$

In the first of the above equations, we used Eq. 14.12. In the second line we made use of the fact that the payoff profile is a linear function of the underlying and hence, that Eq. 14.19 holds, while in the third equation, we finally used of the namesake property of the forward-neutral measure, Eq. 14.13.

The mistake made in some traditional pricing methods corresponds to precisely this last step, i.e., simply replacing the future expectation of the underlying as in Eq. 14.13 with the forward price of the underlying even if the underlying is *not* a tradable instrument (and as such, is not a martingale with respect to the forward-neutral measure and does not satisfy Eq. 14.13). This mistake is then corrected (approximately) after the fact by a *convexity adjustment*.

The convexity adjustment is defined as the difference between the future expectation of the underlying (with respect to the forward-neutral measure) and the forward price of the underlying

$$\text{Convexity Adjustment} \equiv \mathrm{E}_t^T[S(T)] - S(t, T) \ . \tag{14.20}$$

For interest rate underlyings, it is often the case that convexity adjustment need to be taken into account. As shown in Sect. 14.4, zero bond yields are no tradable instruments, because the relation between those yields and related prices of the (tradable) instruments are non-linear, according to Table 2.5:

$$B(t, t + \tau) = \begin{cases} \exp(-r\tau) & \text{continuous} \\ (1 + r)^{-\tau} & \text{discrete} \\ (1 + r\tau)^{-1} & \text{simple} \ . \end{cases} \tag{14.21}$$

We now want to determine an approximation for the convexity adjustment of the yield of such a zero bond, i.e., we want to determine

$$\text{Convexity Adjustment} = \mathrm{E}_t^T[r(T, T + \tau)] - r_f(T, T + \tau \,|\, t) \ .$$

Here we have written out explicitly the time dependence of the zero bond yields: $r(T, T + \tau)$ is the (unknown future) spot rate at time T for a period of length τ starting at time T. The forward rate for that same period, as known at time t, is denoted by $r_f(T, T + \tau \,|\, t)$. To find the expectation of the unknown future spot rate we expand the (also unknown) future bond price formally written as function of r as $B(T, T + \tau) = B(r, T, T + \tau)$ as a Taylor series

up to second order around the (known) forward rate $r_f = r(T, T + \tau \,|\, t)$:

$$B(r, T, T + \tau) = \sum_{n=0}^{\infty} \frac{1}{n!} \left[r(T, T + \tau) - r_f(T, T + \tau \,|\, t) \right]^n \left. \frac{\partial^n B(r, T, T + \tau)}{\partial r^n} \right|_{r=r_f}$$

$$\approx B_{r_f} + \left[r - r_f \right] B' \big|_{r=r_f} + \frac{[r - r_f]^2}{2} B'' \big|_{r=r_f} \,,$$

where in the last line we have dropped all time arguments for ease of notation. We now calculate the forward-neutral expectation of this bond price:

$$\underbrace{E_t^T [B]}_{B_r} \approx B_{r_f} + B' \big|_{r=r_f} E_t^T \left[r - r_f \right] + \frac{1}{2} B'' \big|_{r=r_f} \underbrace{E_t^T \left[(r - r_f)^2 \right]}_{\approx \text{var}[r]} .$$

(14.22)

Here, we used the fact that B and its derivatives are evaluated at r_f which is known at time t. Therefore, B and its derivative can be factored out of the expectation.

On the left-hand side appears the expectation of the bond price with respect to the forward-neutral measure. This is, since bonds are tradable, exactly the forward bond price and, by the definition of the forward rate r_f, is identical to $B(r_f, T, T + \tau)$.

The expectation $E_t^T \left[(r - r_f)^2 \right]$ is approximately equal to the variance of $r(T, T + \tau)$ (it would be exactly this variance if $r_f = E_t^T [r]$). To express this variance in terms of values known at time t, the variance is approximated by the variance of the *forward rate*, i.e.:

$$E_t^T \left[(r - r_f)^2 \right] \approx \text{Var}\,[r] \approx \text{Var}\,[r_f] = r_f^2 \sigma_f^2 \,(T - t) \,. \qquad (14.23)$$

Here the volatility σ_f of the forward rate r_f, called the *forward volatility*, appears. This is (at least in principle) known at time t.

Substituting all of the above into Eq. 14.22 yields

$$0 \approx B' \big|_{r=r_f} \left(E_t^T [r] - r_f \right) + \frac{1}{2} B'' \big|_{r=r_f} r_f^2 \sigma_f^2 \,(T - t) \,.$$

This can now be solved for the desired expectation of the future spot rate:

$$E_t^T [r] \approx r_f \underbrace{- \frac{1}{2} r_f^2 \sigma_f^2 \,(T - t) \frac{B'' \big|_{r=r_f}}{B' \big|_{r=r_f}}}_{\text{convexity adjustment}} . \qquad (14.24)$$

Within these approximations, i.e., by expanding the bond price up to second order (see Eq. 14.22) and with the approximations in Eq. 14.23, the future expectation of the interest rate can thus be approximated by the forward rate adjusted by the amount:

$$
\begin{aligned}
&\mathrm{E}_t^T\left[r(T, T+\tau)\right] - r_f(T, T+\tau\,|\,t) \\
&\approx -\frac{1}{2} r_f(T, T+\tau\,|\,t)^2 \sigma_f^2\,(T-t)\frac{B''(r, T, T+\tau)\big|_{r=r_f}}{B'(r, T, T+\tau)\big|_{r=r_f}} \qquad (14.25)\\
&= \begin{cases}
\frac{1}{2} r_f(T, T+\tau\,|\,t)^2 \sigma_f^2\,(T-t)\tau & \text{continuous}\\
\frac{1}{2} r_f(T, T+\tau\,|\,t)^2 \sigma_f^2\,(T-t)\frac{\tau(\tau+1)}{1+r_f} & \text{discrete}\\
r_f(T, T+\tau\,|\,t)^2 \sigma_f^2\,(T-t)\frac{\tau}{1+\tau r_f} & \text{simple.}
\end{cases}
\end{aligned}
$$

Here, the convexity adjustments for all compounding conventions listed in Eq. 14.21 have been explicitly calculated.

14.5.1   In-Arrears Swaps

As an example, we consider a forward contract on an interest rate index (zero bond yield). As was mentioned at the end of Sect. 14.4, no convexity correction is required for a standard forward rate agreement, if the payment of the interest rate (difference), which is fixed at the start of the period, is at the *end* of the period (resp. if the present value of the virtual cash flow at the end of the period is paid at the start of the period). Effectively, such contracts are forward contracts on (tradable) zero bonds. The same holds for plain vanilla swaps since such a swap can be interpreted as a portfolio of forward rate agreements.

But there are swaps for which the difference between the future interest rate index, e.g. LIBOR, and the fixed side is paid at the same time when the LIBOR rate is fixed. Or, to put it the other way round: the interest payable is determined only at the time when payment is to be made, i.e., at the *end* of the corresponding interest period. Such instruments are called *In-Arrears Swaps*. Here indeed the underlying is directly the LIBOR rate which is *not* a tradable instrument. If we nevertheless wish to price such an instrument as if we could replace the future expectation of the underlying (with respect to the forward-neutral measure) with the forward rate, the resulting error must be corrected, at least approximately, by the convexity adjustment.

For the sake of simplicity, we consider only one period of an In-Arrears Swap, an *In-Arrears Forward Rate Agreement (FRA)*[7] so to speak. This has a principal N, a fixed interest rate K and extends over a period from T to $T + \tau$. A potential compensation payment is calculated by means of simple compounding and flows at the beginning of this period, i.e., directly at time T when the LIBOR rate $r(T, T + \tau)$ is fixed. Such an FRA has a payoff profile given by

$$V(r(T, T + \tau), T) = N\tau \left[r(T, T + \tau) - K\right].\qquad(14.26)$$

This is a *linear* function of the underlying r. The expectation of the payoff profile is thus simply

$$\mathrm{E}\left[V(r(T, T + \tau), T)\right] = N\tau \mathrm{E}\left[r(T, T + \tau)\right] - N\tau K,\qquad(14.27)$$

and its value today is this expectation, taken with respect to the forward-neutral measure, discounted back to today

$$V(r(T, T + \tau), t) = B(r, t, T)N\tau \left(\mathrm{E}_t^T \left[r(T, T + \tau)\right] - K\right).$$

Up to now, all the equations are exact. The calculation of the future expectation is now performed either using a term structure model or an approximation by means of the convexity adjustments given in Eq. 14.24:

$$
\begin{aligned}
&V(r(T, T + \tau), t)\\[4pt]
&= B(r, t, T)N\tau \left(\mathrm{E}_t^T \left[r(T, T + \tau)\right] - K\right)\\[4pt]
&\approx B(r, t, T)N\tau \left(r_f(T, T + \tau\,|\,t)\right.\\[4pt]
&\qquad\qquad \left. -\frac{1}{2}r_f^2(T, T + \tau\,|\,t)\sigma_f^2\,(T - t)\frac{B''(r, T, T + \tau)\big|_{r=r_f}}{B'(r, T, T + \tau)\big|_{r=r_f}} - K\right)\\[4pt]
&= B(r, t, T)N\tau \left(r_f(T, T + \tau\,|\,t) + r_f^2(T, T + \tau\,|\,t)\sigma_f^2\,(T - t)\frac{\tau}{1 + \tau r_f} - K\right).
\end{aligned}
$$
$$(14.28)$$

---

[7] An In-Arrears Swap is simply a portfolio consisting of such FRAs.

The convexity adjustment for simple compounding has been used in the last step, since the specified instrument prescribes this compounding convention. Thus for r_f, we must use the forward rate *with respect to simple compounding* as well.

Again, we emphasize that above considerations hold only for those instruments whose payoff profiles are *linear* functions of the underlying. Equation 14.19 holds only in this case and only than does the expectation of the *underlying* come into play. For instruments with *non*-linear payoff profiles, on the other hand, the expectation of the *payoff profile* must be calculated directly. For example, for a plain vanilla call the expectation of the payoff profile is, in contrast to Eq. 14.19

$$E\left[\max\left\{S(T) - K, 0\right\}\right] = \int_{-\infty}^{\infty} \max\left\{S - K, 0\right\} p(S) dS$$

$$= \int_{K}^{\infty} (S - K) p(S) dS \neq \max\left\{E\left[S(T)\right] - K, 0\right\} .$$

14.5.2 Money Market Futures

The above example of a In-Arrears FRA may appear a bit academic to the reader. In reality though, *money market futures* are among the most actively traded interest derivatives and are nothing other than In-Arrears FRAs traded on an exchange, most commonly based on the 3-month LIBOR or the EURIBOR.

A money market future with a nominal N, a fixed rate K over a period from T to $T + \tau$ yields (theoretically) at maturity T a compensation payment calculated using simple compounding as in Eq. 14.26, thus

$$V(T) = N\tau\left[r(T, T + \tau) - K\right] .$$

Since this instrument has a *future-styled* payment mode, the changes in the position's value do not remain unrealized until maturity T, but are *immediately* realized on a margin account. As explained in Sect. 6.1.4, this has the effect that today's value of a futures position is equal to the future expectation of the payoff profile *without discounting*. This expectation of the future payoff profile is given by Eq. 14.27. The value of a money market future at time t is then, in contrast to Eq. 14.28, simply:

$$V(t) = N\tau\left(E_t^T\left[r(T, T + \tau)\right] - K\right) . \tag{14.29}$$

This is obviously equal to zero when $K = \mathrm{E}[r(T, T + \tau)]$. The fixed rate K of a money market future is in fact always chosen so that the value of the contract is zero at the time when the contract is entered into. The fixed rate K of a money market futures contracted at time t thus gives directly the information on the opinion of the market on the value of $\mathrm{E}[r(T, T + \tau)]$, in other words, on the *future expectation* of the interest rate. Since the interest rate is not a tradable instrument, this is *not* equal to the forward rate.

Money market futures are often used in constructing spot rate curves for maturities ranging from approximately three months to two years (we refer the reader to Part VI). To construct spot rate curves from such contracts, the *forward* rates associated with these transactions are needed, since we can use these to calculate the spot rates quite easily by utilizing Eq. 2.6, for example. To determine the forward rate $r_f(T, T + \tau \,|\, t)$ from the market's opinion on the future LIBOR *expectation* obtained from quotes on money market futures, the convexity adjustment must be *subtracted* from the expectation as in Eq. 14.20 to obtain

$$
\begin{aligned}
r_f(T, T + \tau \,|\, t) &= \mathrm{E}_t^T \left[r(T, T + \tau) \right] - \text{Convexity Adjustment} \\
&\approx \mathrm{E}_t^T \left[r(T, T + \tau) \right] \\
&\quad + \frac{1}{2} r_f^2(T, T + \tau \,|\, t)\sigma_f^2 (T - t) \frac{B''(r, T, T + \tau)\big|_{r=r_f}}{B'(r, T, T + \tau)|_{r=r_f}} \\
&= \mathrm{E}_t^T \left[r(T, T + \tau) \right] - r_f^2(T, T + \tau \,|\, t)\sigma_f^2 (T - t) \frac{\tau}{1 + \tau r_f} ,
\end{aligned}
$$
$$(14.30)$$

where in the last step the approximation in Eq. 14.25 for the convexity adjustment for the linear compounding convention is used. This is a non-linear equation which can be solved numerically for the unknown r_f.

As has been often emphasized, the convexity adjustments presented above are only approximations since they derive from an approximation of the theoretical value of $\mathrm{E}_t^T \left[r(T, T + \tau) \right]$, see Eq. 14.24). The future expectation $\mathrm{E}_t^T \left[r(T, T + \tau) \right]$ can, however, be calculated by other means, for example, using a term structure model. We then obtain *another* expression for $\mathrm{E}_t^T \left[r(T, T + \tau) \right]$, and not Eq. 14.24. Since the forward rates are determined solely from arbitrage considerations (independent of any model) as in Eq. 2.6, this implies that also another expression for the convexity adjustment is obtained. The convexity adjustment is thus dependent on the method (the term structure model) being used. To ensure consistency, the convexity

adjustment for the money market futures used in constructing the spot rate curves should be consistent with the term structure model applied for pricing.

Quotation for Money Market Futures

In Europe, money market futures are traded primarily on the *LIFFE*. Futures on the 3-month LIBOR in pound sterling (*short sterling future*), in US dollars (*euro dollar future*), in euros (*euro EUR future*) and in Swiss franc (*euro Swiss future*) are available for trade on this exchange. Futures in euros on the 3-month EURIBOR and on the 1-month EURIBOR are available for trade on the *EUREX*.

Money market futures are quoted in a way which takes some getting used to. Not the delivery price $E_t^T [r(T, T + \tau)]$ is quoted, i.e., the fixed interest rate for which the future has no value, but rather

$$\text{Quote}_{\text{Money Market Future}} = 100\% - E_t^T [r(T, T + \tau)] .$$

A quote of 96.52%, for example, for a money market future means that in the opinion of the market, the expectation (in the forward neutral measure) for the future 3-month rate is $E[r(T, T + \tau)] = 3.48\%$. The value at time t of a futures position contracted at time $t = 0$ with $K = E_0^B [r(T, T + \tau)]$ is given, according to Eq. 14.29, by

$$\begin{aligned}
V(t) &= N\tau \left(E_t^T [r(T, T + \tau)] - E_0^T [r(T, T + \tau)] \right) \\
&= N\tau \underbrace{\left(1 - E_0^T [r(T, T + \tau)] \right)}_{\text{Quote at Time } t=0} - N\tau \underbrace{\left(1 - E_t^T [r(T, T + \tau)] \right)}_{\text{Quote at Time } t} .
\end{aligned}$$

A money market future on a 3-month LIBOR (i.e., $\tau = 1/4$) with a nominal amount of $N = 1,000,000$ euros which was agreed to at a quoted price of 96.52% and which is currently quoted as 95.95% is thus valued at

$$\begin{aligned}
V(t) &= 1.000.000 \text{ EUR} \times \tfrac{1}{4} (96, 52\% - 95, 95\%) \\
&= 1.425 \text{ EUR} .
\end{aligned}$$

This amount is deposited in a margin amount. A change in the value of this position is directly reflected by a corresponding daily adjustment in the balance of the margin account.

14.6 Arbitrage-Free Interest Rate Trees Grid (Tree) Models

Construction of arbitrage-free *tree models* begins with the assumption that the martingale property Eq. 13.15 holds, justifying the name arbitrage free.[8] An appropriate normalizing factor (numeraire) is selected and the integrals necessary for the computation of the expectations (with respect to the chosen measure) are discretized in a tree-structure. This procedure will be demonstrated explicitly in this section for *1-factor short rate models*, i.e., for models which have the instantaneous short rate, defined in Eq. 14.1, as their one and only stochastic driver.

We will use the risk-neutral measure and the associated numeraire, which is the bank account. The price of every interest rate instrument is then given by Eq. 14.8. We discretize first with respect to time by partitioning the time axis in intervals of length δt taking this length to be so small that the (stochastic) short rate can be assumed to be constant over this interval. Then Eq. 14.8 for $u = t + \delta t$ becomes

$$V(t) = \mathrm{E}_t^\beta \left[e^{-\int_t^{t+\delta t} r(s)ds} V(t+\delta t) \right] \qquad (14.31)$$

$$\approx \mathrm{E}_t^\beta \left[e^{-r(t)\delta t} V(t+\delta t) \right]$$

$$= e^{-r(t)\delta t} \mathrm{E}_t^\beta \left[V(t+\delta t) \right]$$

$$= B(t, t+\delta t)\, \mathrm{E}_t^\beta \left[V(t+\delta t) \right] \; .$$

In calculating the integral, we have made use of the assumption that r is approximately constant on the interval of integration. Variables which are known at time t can be factored out of the expectation. According to the last equation, the risk-neutral price and the forward-neutral price (see Eq. 14.12) cannot be distinguished from one another over the very *short* time interval δt. This is in agreement with Sect. 14.3, since r is taken to be constant (in particular, deterministic) over this short time interval. The equality, however, does not hold for longer time spans, since r changes randomly from one time interval δt to the next. Thus, globally, we are still within the framework of the risk-neutral measure even though the local equations may look "forward neutral".

---

[8]More traditional models, known as *equilibrium models*, will not be investigated here. Our discussion will be restricted to arbitrage-free pricing methods.

14.6.1   Backward Induction

What we have accomplished up to this point is to factor the numeraire (the bank account) out of the expectation. It now remains to calculate the expectation of the instrument's price. To this end, we discretize the continuous range of V after a time step of length δt into finitely many values, i.e., starting from the value of V at time t, the value should be allowed to take on only finitely many different values after the next time step. Allowing *two* different values generates a *bi*nomial tree, *three* a *tri*nomial tree, etc. Since the financial instrument V under consideration is an interest rate instrument, the different values potentially taken on by V at time $t + \delta t$ result directly from the different possible interest rate term structures which might exist at $t + \delta t$. Since one of the model assumptions was that the evolution of the entire interest rate curve is driven by the instantaneous short rate, it follows that the different values attained by V are ultimately determined by the values this short rate can take on.

We will work below with binomial trees. We assume that the short rate increases to the value r_u or decreases to the value r_d with a probability p and $1 - p$, respectively. We have more than one possibility at our disposal to ensure that the martingale property is satisfied (and thus eliminating arbitrage). Either we fix the values r_u and r_d and selects the probability p accordingly so that the market is governed by an arbitrage-free measure (this, for example, is done in finite difference methods where the grid is given at the onset of the analysis), or we specify the probability p first and subsequently select appropriate values for r_u and r_d. We will take the second path in our discussion here. We set

$$p = 1/2 \tag{14.32}$$

and determine the value of the short rate (i.e., the discount factors) on all nodes of the tree so that the short rate process as described by the tree guarantees arbitrage freedom at time t. The binomial tree with $p = 1/2$ allows the expectation in Eq. 14.31 to be written as

$$
\begin{aligned}
V(t) &\approx B(t, t + \delta t)\, \mathrm{E}_t^\beta\, [V(t + \delta t)] \\
&\approx B(t, t + \delta t)\, [p\, V(r_u, t + \delta t) + (1 - p)\, V(r_d, t + \delta t)] \\
&= B(t, t + \delta t) \left[\frac{1}{2} V(r_u, t + \delta t) + \frac{1}{2} V(r_d, t + \delta t) \right] \\
&= B(t, t + \delta t) \left[\frac{1}{2} V_u + \frac{1}{2} V_d \right] ,
\end{aligned}
\tag{14.33}
$$

where in the last step, the short form notation $V_u := V(r_u, t + \delta t)$, etc. has been introduced.

One time step later, the short rate branches again (and in consequence, the price of the financial instrument does as well): $r_u \rightarrow r_{uu}$ with probability p, and $r_u \rightarrow r_{ud}$ with probability $1 - p$, and similarly for r_d. In addition, note that at time $t + \delta t$, two *different* discount factors $B(t + \delta t, t + 2\delta t)$ appear, according to whether the short rate rose to r_u or fell to r_d in the previous step. To emphasize the difference, the discount factors are indexed with the associated short rate. V_u, for example, is then expressed as

$$V_u \approx B_{r_u}(t + \delta t, t + 2\delta t) \, \mathrm{E}_t^\beta \left[V_u(t + 2\delta t) \right]$$

$$\approx B_{r_u}(t + 1\delta t, t + 2\delta t) \left[\frac{1}{2} V(r_{uu}, t + 2\delta t) + \frac{1}{2} V(r_{ud}, t + 2\delta t) \right]$$

$$= B_u \left[\frac{1}{2} V_{uu} + \frac{1}{2} V_{ud} \right] ,$$

where in the last step the short form notation $V_{uu} := V(r_{uu}, t + 2\delta t)$, analogous to that introduced above, has been used, and also the short form notation $B_u := B_{r_u}(t + 1\delta t, t + 2\delta t)$. Analogously,

$$V_d \approx B_d \left[\frac{1}{2} V_{du} + \frac{1}{2} V_{dd} \right] .$$

Substituting this into Eq. 14.33, the price $V(t)$ given by a binomial tree with two steps can be calculated as

$$V(t) \approx B(t, t + \delta t) \left[\frac{1}{2} B_u \left[\frac{1}{2} V_{uu} + \frac{1}{2} V_{ud} \right] + \frac{1}{2} B_d \left[\frac{1}{2} V_{du} + \frac{1}{2} V_{dd} \right] \right] . \tag{14.34}$$

Proceeding analogously, $V_{dd} \approx B_{dd} \left[\frac{1}{2} V_{ddu} + \frac{1}{2} V_{ddd} \right]$, etc., the price at time t calculated from the prices three time steps later is given by

$$V(t) \approx B(t, t + \delta t) \left[\frac{1}{2} B_u \left[\frac{1}{2} B_{uu} \left[\frac{1}{2} V_{uuu} + \frac{1}{2} V_{uud} \right] + \frac{1}{2} B_{ud} \left[\frac{1}{2} V_{udu} + \frac{1}{2} V_{udd} \right] \right] \right.$$

$$\left. + \frac{1}{2} B_d \left[\frac{1}{2} B_{du} \left[\frac{1}{2} V_{duu} + \frac{1}{2} V_{dud} \right] + \frac{1}{2} B_{dd} \left[\frac{1}{2} V_{ddu} + \frac{1}{2} V_{ddd} \right] \right] \right] , \tag{14.35}$$

and so on. Calculating backwards through the tree (*backward induction*) is completely analogous to the treatment of options on stocks and exchange rates with binomial trees as described in Chap. 9.

In order to ensure that any arbitrage opportunity has been eliminated, the short rates r_u, r_d, r_{uud} (and in consequence, the discount factors B_u, B_d, B_{uud}), etc. must be chosen so that the *prices calculated* using the tree actually agree with the *market prices* of traded instruments. In particular, *zero bonds* (and thus today's term structure) must be exactly reproduced. However, at each time point $t + i\delta t$, there are just as many unknowns as a (non-recombinant) binomial tree has nodes, namely 2^i starting with $r_{uu...u}$ continuing through all permutations of up and down moves until $r_{dd...d}$. The number of unknowns increases exponentially with the number of time steps! So many conditions cannot conceivably be generated by the market prices of tradable instruments. In particular, no such functional relation between the number of interest rate instruments on the market and the number of time steps (depending only on the numerical implementation) in a tree could possibly exist. Therefore, based on this principle, another condition must first be established preventing the exponential growth of the number of unknowns with respect to increasing i.

Requiring that the tree be *recombinant* for all financial instruments with path-independent payoff profiles presents itself as a good candidate for the above mentioned condition, i.e., that $V_{ud} = V_{du}$, $V_{uud} = V_{udu} = V_{duu}$, etc. This can only happen if the tree for the underlying also recombines. A recombinant binomial tree is known to have only $i + 1$ nodes after i steps. This fact will be accounted for in our notation. The nodes of the tree at which we arrive having traveled upwards i times and downwards j times is uniquely determined by the ordered pair (i, j), irrespective of the order in which these upward and downward steps were taken. This is because the tree is recombinant. Therefore we denote the value of the financial instrument at this node with $V(i, j)$ and the nodes themselves with the ordered pairs (i, j) as presented in Fig. 14.1. For example, $V(1, 2) = V_{udd} = V_{dud} = V_{ddu}$, etc.

We use the same notation for the short rate and for the zero bonds evaluated at the nodes:

$$(i, j) = \text{Node after } i \text{ up moves and } j \text{ down moves}$$

$$r(i, j) = \text{Instantaneous short rate at the node } (i, j)$$

$$V(i, j) = \text{Value of an interest rate instrument at the node } (i, j)$$

$$B(i, j) = \exp\{-r(i, j)\delta t\} \ . \tag{14.36}$$
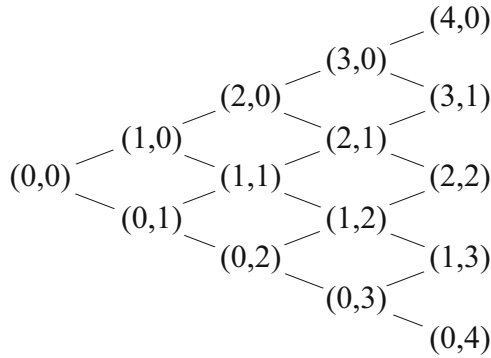
Fig. 14.1   The binomial tree with the indexing showing the number of up and down moves required to get to the nodes starting from node $(0, 0)$. For instance it takes two up moves and one down move to get to the node $(2, 1)$

In particular, $B(t, t + \delta t) = B(0, 0)$. In this notation, the prices of (path independent) financial instruments can be written after one binomial step in a recombinant tree as

$$V(t) = V(0, 0)$$

$$\approx V(1, 0)\frac{1}{2}B(0, 0)$$

$$+ V(0, 1)\frac{1}{2}B(0, 0) \tag{14.37}$$

Of course, this equation holds not only for the node $(0, 0)$ but for any arbitrary node (i, j) in the tree

$$V(i, j) \approx B(i, j)\left[\frac{1}{2}V(i + 1, j) + \frac{1}{2}V(i, j + 1)\right] \tag{14.38}$$

This equation applied to $V(1, 0)$ and $V(0, 1)$ yields the price after two binomial steps

$$V(t) \approx V(2, 0)\frac{1}{4}B(0, 0)\ B(1, 0)$$

$$+ V(1, 1)\frac{1}{4}B(0, 0)\ [B(1, 0) + B(0, 1)]$$

$$+ V(0, 2)\frac{1}{4}B(0, 0)\ B(0, 1) \tag{14.39}$$

and after three

$$V(t) \approx V(3, 0)\frac{1}{8}B(0, 0)B(1, 0)B(2, 0)$$

$$+ V(2, 1)\frac{1}{8}B(0, 0)\left[B(1, 0)B(2, 0) + B(1, 1)\left[B(1, 0) + B(0, 1)\right]\right]$$

$$+ V(1, 2)\frac{1}{8}B(0, 0)\left[B(0, 1)B(0, 2) + B(1, 1)\left[B(1, 0) + B(0, 1)\right]\right]$$

$$+ V(0, 3)\frac{1}{8}B(0, 0)B(0, 1)B(0, 2) \tag{14.40}$$

and so on. The expressions were purposely written in terms of the prices at the last node in each branch.

14.6.2   Forward Induction and Green's Functions

Although backward induction was simplified greatly by the requirement that the tree be recombinant, we are still not yet prepared to calculate an instrument's price since the instantaneous discounting factors $B(i, j)$ at the nodes $(i, j) \neq (0, 0)$ remain unknown. Before we construct a tree for the instantaneous discount factors (i.e., for the short rate, see Eq. 14.36) via a procedure, known as *forward induction*, on the basis of arbitrage considerations, we introduce a class of extremely useful "artificial" instruments. One such artificial instrument whose value at time t is denoted by $G(i, j)$ pays by definition one monetary unit if and only if the underlying (the short rate) attains the tree node (i, j) at time $t + (i + j)\delta t$. This is the node at which we arrive having traveled upwards i times and downward j times, regardless of the order in which the upward and downward moves occurred as the tree is recombinant. $G(i, j)$ then is the value at time t of a single monetary unit paid out at one single node of the tree, namely at node (i, j). In this sense, G is the system's reaction to a perturbation of magnitude one at a single point in the system. This is analogous to the Green's functions in physics. We will therefore refer to G as a *Green's function*. By definition,

$$G(0, 0) \equiv 1 . \tag{14.41}$$

In order to establish further values of the Green's function, we just have to set the value $V(i, j)$ in the above Eqs. 14.37, 14.39 and 14.40 equal to one at exactly one node and zero on all the other nodes. Equation 14.37 then yields

$$G(1, 0) = \frac{1}{2}B(0, 0) = G(0, 1) \qquad (14.42)$$

From Eq. 14.39 we obtain

$$G(2, 0) = \frac{1}{4}B(0, 0) \ B(1, 0) = \frac{1}{2}G(1, 0)B(1, 0)$$

$$G(0, 2) = \frac{1}{4}B(0, 0) \ B(0, 1) = \frac{1}{2}G(0, 1)B(0, 1)$$

$$G(1, 1) = \frac{1}{4}B(0, 0) \ [B(1, 0) + B(0, 1)] = \frac{1}{2}G(1, 0)B(1, 0) + \frac{1}{2}G(0, 1)B(0, 1)$$

and finally Eq. 14.40 gives

$$G(3, 0) = \frac{1}{2}G(2, 0)B(2, 0)$$

$$G(0, 3) = \frac{1}{2}G(0, 2)B(0, 2)$$

$$G(2, 1) = \frac{1}{2}G(2, 0)B(2.0) + \frac{1}{2}G(1, 1)B1, 1)$$

$$G(1, 2) = \frac{1}{2}G(0, 2)B(0, 2) + \frac{1}{2}G(1, 1)B(1, 1) \ .$$

The following general recursion relation can be easily verified (this will be proven for an even more general case later)

$$G(i, j) = \frac{1}{2}G(i, j-1)B(i, j-1) + \frac{1}{2}G(i-1, j)B(i-1, j) \quad \text{for } i > 0, j > 0$$

$$G(i, 0) = \frac{1}{2}G(i-1, 0)B(i-1, 0)$$

$$G(0, j) = \frac{1}{2}G(0, j-1)B(0, j-1) \qquad (14.43)$$

The prices of all path-independent interest rate instruments can be represented as linear combinations of the Green's function evaluated at diverse nodes on the tree since each payment profile of the form $f(r, T)$ can be distributed on

the nodes (i, j) as appropriate:

$$f(r, T) \to f_T(i, j) := f(r(i, j), T) \quad \text{with} \quad t + (i + j)\delta t = T \quad \forall i, j .$$

The value at time t of each individual payment $f_T(i, j)$ at node (i, j) is naturally equal to the Green's function belonging to this node (which has a value of exactly one monetary unit) multiplied by the number of monetary units that are to be paid, i.e., multiplied by $f_T(i, j)$. The total value of the financial instrument $V(t)$ with this payoff profile is then simply

$$V(t, T) = \sum_{(i,j)} f_T(i, j) \, G(i, j) \quad \text{with} \quad t + (i + j)\delta t = T \quad \forall i, j ,$$

where $\sum_{(i,j)}$ denotes "the sum over the nodes (i, j)". This can be immediately generalized to instruments with payoff profiles defined on *arbitrary* nodes (which need not all lie in the set of nodes corresponding to a time T)

$$V(t) = \sum_{(i,j)} f(i, j) \, G(i, j) \quad \text{for arbitrary payoff profiles } f(i, j) .$$

$$(14.44)$$

Path independence here is therefore *not* the restriction that the payoff profile depend only on a time point T. The payoff profile can depend on (the interest rate at) all possible nodes at all times, not however, on the path taken to arrive at these nodes (since this information is not available in the recombinant tree).

As a simple example, we consider the value of a zero bond $B(t, T)$, an instrument that pays one monetary unit at time T, regardless of the state of the underlying

$$B(t, T) = \sum_{(i,j)} G(i, j) \quad \text{for} \quad t + (i + j)\delta t = T \qquad (14.45)$$

$$= \sum_{i=0}^{n} G(i, n - i) \quad \text{for} \quad n = \frac{T - t}{\delta t} .$$

This *theoretical* price has to exactly match the *market* price of the zero bond to prevent arbitrage. If the current term structure is available (for example, because it has been constructed on the basis of traded benchmark bonds for some maturities and interpolations in between) this term structure yields the

market price for all zero bonds $B(t, T)$. In particular it yields the market prices for all those zero bonds maturing at times $t + i\delta t$ corresponding to the time steps of the binomial tree. Those have to be matched by Eq. 14.45. Hence, we obtain one single condition for each time step $t + i\delta t$. For example, for $i = 1$

$$B(t, t + 1\delta t) = \sum_{(i,j)} G(i, j) \text{ with } i + j = 1$$

$$= G(1, 0) + G(0, 1)$$

$$= B(0, 0) ,$$

where in the last step, the Green's function as in Eq. 14.42 is introduced into the equation. This simply checks for consistency: the discount factor at the first node of our tree (the right-hand side) must be equal to the market price for the zero bond with maturity $T = t + \delta t$ (left-hand side). It becomes more interesting at the next maturity date:

$$B(t, t + 2\delta t) = \sum_{(i,j)} G(i, j) \text{ with } i + j = 2$$

$$= G(2, 0) + G(1, 1) + G(0, 2)$$

$$= \frac{1}{2}G(1, 0)B(1, 0) + \frac{1}{2}G(1, 0)B(1, 0)$$

$$+ \frac{1}{2}G(0, 1)B(0, 1) + \frac{1}{2}G(0, 1)B(0, 1)$$

$$= G(1, 0)B(1, 0) + G(0, 1)B(0, 1) .$$

We have used Eq. 14.43 to calculate backwards from the Green's function at the time corresponding to $i + j = 2$ to the previous time step, i.e., to the time for which $i + j = 1$. The Green's function $G(1, 0)$ at the earlier time step $t + 1\delta t$ is already known (see Eq. 14.42). The resulting equation is the arbitrage condition for the instantaneous discount factors $B(1, 0)$ and $B(0, 1)$ at the tree nodes $(1, 0)$ and $(0, 1)$. This procedure can be generalized to n time steps. In accordance with Eq. 14.45 we write

$$B(t, t + n\delta t) = \sum_{i=0}^{n} G(i, n - i) .$$

At this point we separate the boundary terms from the rest of the sum, since these obey another recursion in Eq. 14.43. Now applying Eq. 14.43, we see that

$$B(t, t + n\delta t) = G(0, n) + \sum_{i=1}^{n-1} G(i, n - i) + G(n, 0)$$

$$= \frac{1}{2} G(0, n - 1) B(0, n - 1) + \frac{1}{2} \sum_{i=1}^{n-1} G(i, n - i - 1) B(i, n - i - 1)$$

$$+ \frac{1}{2} \sum_{i=1}^{n-1} G(i - 1, n - i) B(i - 1, n - i) + \frac{1}{2} G(n - 1, 0) B(n - 1, 0) .$$

The boundary terms combine this expression into a simple sum. To see this, we use the index $k = i - 1$ in the second sum

$$B(t, t + n\delta t) = \frac{1}{2} G(0, n - 1) B(0, n - 1) + \frac{1}{2} \sum_{i=1}^{n-1} G(i, n - i - 1) B(i, n - i - 1)$$

$$+ \frac{1}{2} \sum_{k=0}^{n-2} G(k, n - k - 1) B(k, n - k - 1) + \frac{1}{2} G(n - 1, 0) B(n - 1, 0) .$$

Both boundary terms now have exactly the form needed to extend the index range in both sums to 0 through $n - 1$. This means that the no arbitrage requirement can be represented as a simple recursion formula by means of the Green's function:

$$B(t, t + n\delta t) = \sum_{i=0}^{n-1} G(i, n - i - 1) B(i, n - i - 1) . \qquad (14.46)$$

Remember: on the left we have the market price of a zero bond, while on the right we have the values to be determined for the interest rate tree, i.e., the instantaneous discount factors at the nodes. The values on the right-hand side are all in terms of nodes corresponding to the time point

$$t + [i + (n - i - 1)] \, \delta t = t + (n - 1)\delta t ,$$

i.e., the time step prior to the maturity $t + n\delta t$ of the zero bond on the left. The *Green's function* evaluated at this earlier time point has already

been calculated in the previous iteration step so that Eq. 14.46 represents an arbitrage condition for the instantaneous *discount factors* at the nodes at time point $t + (n-1)\delta t$. Having determined the discount factors, we can use the recursion for the Green's function, Eq. 14.43, to determine the values of the Green's function at the time point $t + n\delta t$. These are subsequently used again in the arbitrage condition, Eq. 14.46, to calculate the next discount factors for the time step $t + n\delta t$. These are then used to calculate the next Green's function values for $t + (n+1)\delta t$ utilizing Eq. 14.43 and so on. In this way, an arbitrage-free interest rate tree is constructed using *forward* induction from today into the future. Nevertheless, Eq. 14.46 contains several (to be precise n) unknown discount factors which cannot be uniquely determined by only *one* arbitrage condition (except, of course, for the case $n = 1$). In other words, the exact reproduction of the term structure at time t is already attained singly from the fact the instantaneous discount factors and the Green's function at each node satisfy Eqs. 14.46 and 14.43. But this does not fix the *numerical* values of all instantaneous discount factors—or short rates $r(i, j)$. Additional information (as we will see below, the volatility) extracted from the market as well as an explicit specification of a stochastic process of the form 2.19 for the short rate is needed to fix the numerical values of $r(i, j)$. Before we start on this point however, we will first introduce a few more concepts which hold in general, i.e., for every interest rate tree, irrespective of the specification of a specific stochastic process.

14.7 Market Rates vs. Instantaneous Rates

The valuation of financial instruments using the Green's function as in Eq. 14.44 above is only possible for those instruments whose payoff profiles are functions of the *instantaneous* short rate.[9] This is not usually the case for *traded* instruments, however. Typical interest rate underlyings for traded instruments are 3-month or 6-month LIBOR rates, which belong to longer time periods. Choosing such a long period as time distant in the tree, the calculation would no longer be accurate enough for many applications. In contrast to the chapters on stock or FX options, the stochastic process being simulated (the short rate) does *not* describe the evolution of the underlying (the 3-month LIBOR, for example) of the instrument to be priced! This problem can be overcome in

---

[9]Or for financial instruments whose payoff profile is independent of the short rate, such as a zero bond, for which $f(i, j) = 1$ for $t + (i + j)\Delta t = T$ and $f(i, j) = 0$ otherwise.

some cases, as for floaters and forward rate agreements, since, as will be seen in Eqs. 15.21 or 15.4, their prices at time t can be interpreted as a combination of zero bond prices and as such can be priced *exactly* using Eq. 14.46. However, the decomposition into zero bonds will not be possible for options such as caps and floors on the 3-month LIBOR whose general payoff functions are given by Eq. 18.2 and, adopting the usual market practice of using simple compounding over the cap period, through Eq. 18.2 (or when interpreted as bond options by Eq. 18.6). In such cases, before the payoff profile at nodes (m, n) with $t + (m + n)\delta t = T$ can be calculated, the value of the *underlying* on each of these nodes must first be determined. The question now is: how do we calculate the underlying of interest (for example, a 3-month rate, a 6-month rate, a swap rate, etc.) at all nodes corresponding to the exercise date T from the stochastic process (the tree) of the instantaneous short rate?

As soon as the payoff profile at all nodes is known, the value of all (path independent) financials instruments at time t, i.e., at the node $(0, 0)$, are directly given by Eq. 14.44. As with the node $(0, 0)$, the value of all instruments (in particular, of all zero bonds and thus all interest rates over arbitrary interest periods) would be known at an *arbitrary* node (m, n) if the "Green's functions" were known for *that* node.

14.7.1   Arrow-Debreu Prices

Arrow-Debreu prices (*ADP*s for short) are generalized Green's functions whose reference point is a fixed but arbitrary node (m, n) rather than the origin node $(0, 0)$. The Arrow-Debreu price $G_{m,n}(i, j)$ is the value at node (m, n) of an instrument paying one monetary unit at node (i, j). The Arrow-Debreu prices at the node $(m = 0, n = 0)$ are, of course, simply the values of the Green's function introduced above:

$$G(i, j) = G_{0,0}(i, j) .$$

It follows immediately from the geometry of the tree that a monetary unit at node (i, j) can generate non-zero prices at a node (m, n) only if (i, j) is attainable when starting from the node (m, n). We know that m up moves have already occurred at node (m, n). These up moves cannot be undone, even if the following steps consist only of down moves since the index m merely counts the number of up moves having been made up to this point. This implies that for all nodes (i, j) attainable from the starting node (m, n), the condition $i \geq m$ must hold. Likewise, n down moves have already occurred at node

(m, n) which cannot be reversed. As before, this means that the condition $j \geq n$ must hold for all nodes (i, j) which are attainable from the starting point (m, n). All other Arrow-Debreu prices must be zero:

$$
\begin{aligned}
G_{m,n}(i, j) &= 0 \quad \forall m > i \\
G_{m,n}(i, j) &= 0 \quad \forall n > j \\
G_{i,j}(i, j) &= 1 \quad \forall i, j \ ,
\end{aligned}
\tag{14.47}
$$

where the last property is included solely for the sake of completeness, being in itself trivial: one monetary unit at node (i, j) is, of course, worth exactly one monetary unit at this node. This corresponds to property Eq. 14.41 of the Green's function.

A further fundamental property of Arrow-Debreu prices follows from Eq. 14.38, which holds for the value of any instrument at any node in the tree. Setting either $V(i + 1, j) = 1$ and $V(i, j + 1) = 0$ or $V(i + 1, j) = 0$ and $V(i, j + 1) = 1$, yields the *instantaneous* Arrow-Debreu prices, i.e., the Arrow-Debreu prices over a time step of length δt

$$
G_{i,j}(i + 1, j) = \frac{1}{2} B(i, j) = G_{i,j}(i, j + 1) \ .
\tag{14.48}
$$

The instantaneous Arrow-Debreu prices are thus half the instantaneous discount factors. This accomplishes the first step in the calculation of the Arrow-Debreu prices. We now merely require a generalization of the recursion relation given in Eq. 14.43 in order to determine all subsequent Arrow-Debreu prices. To this end, consider one monetary unit at node (i, j) at time $t + (i + j)\delta t$. Because of the binomial structure of the tree, non-zero Arrow-Debreu prices are generated by this monetary unit at two nodes in the previous time slice, that is at nodes $(i, j - 1)$ and $(i - 1, j)$. The *sum* of the ADPs of *both* of these prices at a still earlier node (m, n) must then be equal to the Arrow-Debreu price at node (m, n) of the whole, original monetary unit at node (i, j). This is illustrated in Fig. 14.2. This means that the Arrow-Debreu prices obey the following recursion:

$$
G_{m,n}(i, j) = G_{m,n}(i, j - 1)G_{i,j-1}(i, j) + G_{m,n}(i - 1, j)G_{i-1,j}(i, j) \ .
$$
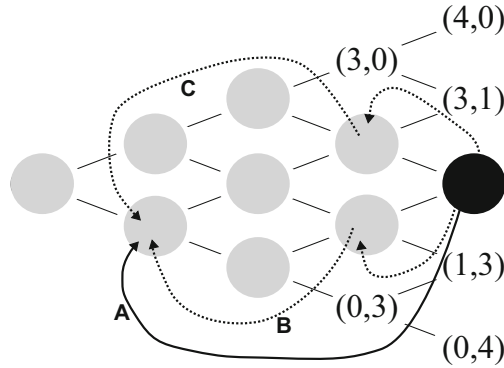
Fig. 14.2   A monetary unit (black dot) at node $(2, 2)$ generates ADPs at nodes $(2, 1)$ and $(1, 2)$ and also at all the earlier 'striped' nodes. The monetary unit must have the same influence (shown as line **A**) on a striped node, e.g. on node $(0, 1)$, as both ADPs it has induced at nodes $(2, 1)$ and $(1, 2)$ together (shown as lines **B** and **C**)

Substituting the corresponding discount factors for both of the instantaneous Arrow-Debreu prices as given in Eq. 14.48, the recursion relation becomes

$$G_{m,n}(i, j) = G_{m,n}(i, j-1)\frac{1}{2}B(i, j-1) + G_{m,n}(i-1, j)\frac{1}{2}B(i-1, j) \text{ for } i \geq m, \ j \geq n \ .$$

$$(14.49)$$

Together with the fundamental properties in Eq. 14.47, this recursion determines uniquely all Arrow-Debreu prices. For example, for $m = i$ or $n = j$ Eq. 14.47 implies that one of the Arrow-Debreu prices on the right-hand side is equal to zero.[10] This allows the recursion to be carried out explicitly for these cases:

$$G_{i,n}(i, j) = G_{i,n}(i, j-1)\frac{1}{2}B(i, j-1) = \frac{1}{2^{j-n}}\prod_{k=1}^{j-n} B(i, j-k)$$

$$G_{m,j}(i, j) = G_{m,j}(i-1, j)\frac{1}{2}B(i-1, j) = \frac{1}{2^{i-m}}\prod_{k=1}^{i-m} B(i-k, j) \ .$$

$$(14.50)$$

---

[10] For $m = i$ we have $G_{i,n}(i-1, j) = 0$ and for $n = j$ we have $G_{m,j}(i, j-1) = 0$.

In particular, both of these equations hold at the boundary of the tree. The first equation with $i = 0$ (i.e., no up move) represents the *lower* boundary, the second with $j = 0$ (i.e., no down move) the *upper* boundary of the tree.

With these Arrow-Debreu prices, the value of any arbitrary financial instrument with a payoff profile given by $f(i, j)$ at any *arbitrary* node (m, n) is, analogous to Eq. 14.44, simply

$$V(m, n) = \sum_{(i, j)} G_{m,n}(i, j)\, f(i, j) \; . \qquad (14.51)$$

In particular, the value $B_\tau(m, n)$ of a zero bond at the node (m, n), whose time to maturity at this node is given by τ, is likewise given by Eq. 14.51 where i and j satisfy the following three conditions

$$i + j = m + n + \tau/\delta t$$
$$i \geq m$$
$$j \geq n \; .$$

The first of the three conditions characterizes all nodes corresponding to the time $m + n + \tau/\delta t$, the maturity of the zero bond. The payoff profile $f(i, j)$ of the zero bond equals one at precisely these nodes and zero elsewhere. This condition allows j to be expressed in terms of i. The limits in the sum appearing in Eq. 14.51 can be specified explicitly by the other two conditions which, as a result of Eq. 14.47 must always be satisfied: the lower limit is $i \geq m$, the upper limit follows from $m + n + \tau/\delta t - i = j \geq n$ which can be rewritten as $i \leq m + \tau/\delta t$. The value of the zero bond at node (m, n) with a time to maturity of τ is thus given explicitly by

$$B_\tau(m, n) = \sum_{i=m}^{m+\tau/\delta t} G_{m,n}(i, m + n + \tau/\delta t - i) \; . \qquad (14.52)$$

Thus, at each node of the tree, a complete (future) term structure (i.e., future interest rates for arbitrary times to maturity) can be constructed from Arrow-Debreu prices, since the interest rate at node (m, n) for any arbitrary time to maturity τ is by definition (for continuous compounding) given by

$$r_\tau(m, n) = -\frac{1}{\tau} \ln B_\tau(m, n) \; . \qquad (14.53)$$

With simple compounding, which is the common convention for typical interest rate indexes like 3-months LIBOR, it follows analogously:

$$r_\tau(m, n) = \frac{1}{\tau} \left(B_\tau^{-1}(m, n) - 1 \right) , \qquad (14.54)$$

so that we are now in a position to price any arbitrary derivative on an underlying whose value can be derived from the term structure.

14.7.2 Pricing Caplets Using Arrow-Debreu Prices

In anticipation of Part III, where caps and floors will be defined, we will demonstrate explicitly how the price of a caplet on a 3-month rate with principal N, strike rate K, exercise time (maturity, to be assumed to be equal to forward rate fixing time) T and payment date $T' = T + \tau$ (with $\tau = 3$ months) can be expressed solely in terms of Arrow-Debreu prices. Observe that the stochastic process for the short rate need *not* be specified in order to do so. What will be shown here holds for *any arbitrary* arbitrage-free short rate model.

Consistent with Eq. 18.2 for the payoff profile of a caplet, we will adopt the market convention of using *simple compounding* over a single caplet period (in this case 3 months), which can be calculated by using Eq. 14.54. The 3-month rate at which the payoff profile is to be evaluated at each node (m, n) with $t + (m + n)\delta t = T$, or equivalently, $n = (T - t)/\delta t - m$ is thus:

$$r_\tau(m, a - m) = \frac{1}{\tau} \left[B_\tau(m, a - m)^{-1} - 1 \right] \qquad (14.55)$$

$$= \left(b\delta t \sum_{i=m}^{m+b} G_{m,a-m}(i, a + b - i) \right)^{-1} - \frac{1}{b\delta t}$$

$$\text{for all } 0 \le m \le a ,$$

where we have defined

$$a := (T - t)/\delta t , \quad b := \tau/\delta t .$$

This is the underlying of our caplet at all nodes relevant to the caplet's payoff profile. According to the payoff profile in Eq. 18.2, the values of the caplet at

exercise date T, i.e., at nodes $(m, a - m)$ with $0 \leq m \leq a$, are given by

$$f(m, a - m) = \tau N \, B_\tau(m, a - m) \max\{r_\tau(m, a - m) - K, 0\} .$$

Both the discount factor B_τ appearing in the payoff profile as well as the underlying (the interest rate r_τ) can be expressed in terms of the Arrow-Debreu prices:

$$f(m, a - m) = N \sum_{i=m}^{m+b} G_{m,a-m}(i, a + b - i)$$

$$\times \max\left\{\left(\sum_{i=m}^{m+b} G_{m,a-m}(i, a + b - i)\right)^{-1} - 1 - \tau K, 0\right\} .$$

Equation 14.44 now directly yields the value of this payoff profile, i.e., the caplet value, at time t

$$c^{\mathrm{cap}}(T, T + \tau, K \,|\, t) = \sum_{m=0}^{a} G(m, a - m) \, f(m, a - m)$$

$$= N \sum_{m=0}^{a} G_{0,0}(m, a - m) \sum_{i=m}^{m+b} G_{m,a-m}(i, a + b - i)$$

$$\times \max\left\{\left(\sum_{i=m}^{m+b} G_{m,a-m}(i, a + b - i)\right)^{-1} - 1 - \tau K, 0\right\} .$$

The price of the caplet has now been expressed completely in terms of the Arrow-Debreu prices. Since these prices can be determined using the recursion relations 14.49 and 14.50 with initial values given by Eqs. 14.47 and 14.48, this procedure can be used to price any interest rate instrument.

Practical Implementation of Arrow-Debreu Prices

Arrow-Debreu prices have four indices, two for the position of the cash flow in the tree and two for the position at which the effect of the cash flow is felt. The numerical implementation thus requires in principle the construction of a four-dimensional field, which leads to computer memory problems for

somewhat finer trees as well as performance problems in the computation of all the ADPs. It will prove to be unnecessary in most cases to compute all of the ADPs. In the above case of caplets, for example, it is only necessary to compute the ADPs of the form $G_{m,a-m}(i, a+b-i)$ with the fixed parameters a and b. Thus, only two of the four above mentioned indices are free, namely m and i. Consequently, it is entirely sufficient to determine the two-dimensional field

$$\widetilde{G}_{m,i} := G_{m,a-m}(i, a + b - i)$$

thereby reducing the numerical effort involved considerably. The effort is now only the same as that needed in the determination of the Green's function.

14.8 Explicit Specification of Short Rate Models

Up to this point, the pricing procedure has remained quite general in that no specific term structure model has been used. None of the relations introduced above can as yet be used to calculate an explicit numerical value since, as mentioned above, the arbitrage condition in Eq. 14.46 is by no means sufficient to determine all the (unknown) discount factors appearing in this equation. From now on, we suppose that the term structure model is of the general form 2.19 where we restrict ourselves to functions $b(r, t)$ of the form $b(r, t) = b(t)r^{\beta}$, i.e. models of the form

$$dr(t) = a\,(r, t)\,dt + b(t)r^{\beta}\,dW \ . \tag{14.56}$$

If the exponent $\beta = 0$,

$$dr(t) = a\,(r, t)\,dt + b(t)\,dW \ , \tag{14.57}$$

we end up with a *normal model* or *Gaussian model*. In the special case that parameter a is also independent of r, the model is called *Ho-Lee Model*. Such models with normally distributed short rate have the advantage that they are very easy to implement (for example, as a tree). The Ho-Lee model

$$dr(t) = a(t)dt + b(t)\,dW$$

for constant b is even analytically tractable (for this special case the model was originally invented [97]). It used to be seen as a disadvantage that normal models allow for negative interest rates and that the volatility is absolute rather

than relative with respect to the forward rates. However, with the advent of real negative interest rates for EUR and CHF, these features turned out to be clear advantages, at least for the EUR and CHF currency area. Normal models also tend to be easier calibrated in markets with low, but positive interest rates, avoiding the tendency to push the interest rate level away from zero.

The second class of short rate models does not specify a stochastic process for the short rate itself, but for its logarithm:

$$d \ln r(t) = a(r, t)\, dt + b(r, t)\, dW \qquad (14.58)$$

These models are thus referred to as *lognormal models*. In the special case of the volatility b being independent of r, this process is called the *Black–Derman-Toy model*, which has been especially designed for the binomial tree, as well as the *Black Karasinski model*. Sections 14.11.2 discusses short rate models in more detail. Such lognormal models have the disadvantage that they are much more difficult to implement and cannot be solved analytically. Negative interest rates cannot occur in these models and the parameter b represents the relative volatility, simplifying fitting this model to relative volatilities observed in markets with strictly positive interest rates. Such models can be transformed into models for r via the Ito formula,[11] Eq. 2.21:

$$dr(t) = \left[r(t)a(r, t) + \frac{1}{2}b^2(r, t)r(t) \right] dt + r(t)\, b(r, t)\, dW \ . \qquad (14.59)$$

The numerical evaluation of lognormal models is easier to implement when written in this form.

14.8.1   The Effect of Volatility

According to the Girsanov Theorem (see Sect. 13.4), the drift $a(r, t)$ is uniquely determined by the probability measure used. We fixed this probability measure "by hand" when we required the condition specified in Eq. 14.32 to be satisfied. In doing so, we implicitly fixed the drift as well. The drift

---

[11]For this, we need to consider the following: $\ln r(t)$ in Eq. 14.58 corresponds to S in Eq. 2.21. For the function f we take $f(S) = e^S$ (since this is exactly r). The partial derivatives appearing in Eq. 2.21 are then simply $\partial f / \partial t = 0$ and $\partial f / \partial S = \partial^2 f / \partial S^2 = f = r$.

can thus no longer be given as "input" into the specification of the stochastic process. Since different normal and lognormal models differ only in their drift, this implies that a special choice for Eq. 14.32 effectively belongs to a special choice of the normal resp. lognormal model, namely the Ho Lee and Black Derman Toy model. Only the volatility remains as a parameter through which market information (in addition to the bond prices) may enter into our model for the determination of the discount factors in Eq. 14.46. We will now show that if the volatility is given for the time step n, exactly $n - 1$ conditions are generated on the n instantaneous discount factors at time n. Taken together with the arbitrage condition given by Eq. 14.46, requiring that the observed market price of the zero bond with maturity $T = t + n\delta t$ be reproduced by the model, we have exactly as many conditions as unknowns and the interest rate tree can be uniquely constructed.

In general, from the viewpoint of the node (i, j), the variance of a variable x (we can think of x as representing, for example, the short rate r or its logarithm $\ln r$) is caused by its possible two different values in the next time step, either x_u or x_d. The expectation and variance for random variables of this type are given by[12]

$$E[x] = px_u + (1 - p)x_d \qquad\qquad (14.60)$$

$$\text{Var}[x] = p(1 - p)(x_u - x_d)^2 \, ,$$

In particular, for the case $p = 1/2$ the variance is given by $\text{Var}[x] = (x_u - x_d)^2/4$.

On the other hand, for models of the form given by Eqs. 14.57 and 14.58, the variance of the stochastic variable x (with $x = r$ resp. $x = \ln r$) over an interval of time of length δt is given by $b(r, t)^2 \delta t$.

---

[12]Substituting the expectation into the definition of the variance defined as the expectation of the squared deviation from the expectation gives

$$\text{Var}[x] = E\left[(x - E[x])^2\right]$$

$$= p\,(x_u - E[x])^2 + (1 - p)\,(x_d - E[x])^2$$

$$= p\,(x_u - px_u - (1 - p)x_d)^2 + (1 - p)\,(x_d - px_u - (1 - p)x_d)^2 \, .$$

Multiplying out and collecting terms yields the desired expression.

14.8.2 Normal Models

For models of the form in Eq. 14.57 the variance of the short rate at node (i, j) must satisfy:

$$b(i, j)\sqrt{\delta t} = \sqrt{\mathrm{Var}\,[r(i, j)]} = \frac{1}{2}\,[r(i + 1, j) - r(i, j + 1)]$$

$$\implies \quad r(i + 1, j) = r(i, j + 1) + 2b(i, j)\sqrt{\delta t}\;. \tag{14.61}$$

This enables us to establish a recursion formula (after performing the substitution $i + 1 \to i$) for the instantaneous discount factors at all nodes corresponding to the time slice $n = i + j$

$$
\begin{aligned}
B(i, j) &= \exp\{-r(i, j)\delta t\} \\
&= \exp\left\{-\left[r(i - 1, j + 1) + 2b(i - 1, j)\sqrt{\delta t}\right]\delta t\right\} \\
&= \exp\left\{-2b(i - 1, j)\sqrt{\delta t}\delta t\right\}\exp\{-r(i - 1, j + 1)\delta t\} \\
&= \exp\left\{-2b(i - 1, j)\sqrt{\delta t}\delta t\right\}B(i - 1, j + 1)\;,
\end{aligned}
$$

or

$$B(i, j) = \alpha(i - 1, j)B(i - 1, j + 1) \quad \text{mit } \alpha(i, j) = \exp\left\{-2b(i, j)\delta t^{3/2}\right\}\;. \tag{14.62}$$

Recursive substitution into this equation allows each instantaneous discount factor in the nodes corresponding to this time slice to be expressed in terms of a single discount factor at the "lowest" node $(0, i + j)$:

$$
\begin{aligned}
B(i, j) &= \alpha(i - 1, j)B(i - 1, j + 1) \\
&= \alpha(i - 1, j)\alpha(i - 2, j + 1)B(i - 2, j + 2) \\
&= \alpha(i - 1, j)\alpha(i - 2, j + 1)\alpha(i - 3, j + 2)B(i - 3, j + 3) \\
&= \cdots \\
&= B(0, j + i)\prod_{k=1}^{i}\alpha(i - k, j + k - 1)\;.
\end{aligned}
$$

Observe that only α-parameters (volatility information) in the time slice $i - k + j + k - 1 = i + j - 1$ are required for the determination of this value, i.e., information from the *previous* time step. Volatility information for the value of the interest rate at the node actually being calculated (i.e., for the spacing of the nodes at the time slice $(i + j)$) is *not* required; these values are determined with volatility information from the immediately proceeding time! This (perhaps counter-intuitive) property is of course nothing other than the *previsibility* which we always require for the coefficients of dW and dt in all models of the form 2.19 or 13.17. At this point now, we get an intuitive picture (see also Fig. 14.3) what it means that a and b in Eq. 2.19 are previsible processes. This previsibility is the deeper mathematical reason for why we are able calculate anything at all.

Using this expression for $B(i, j)$ with $j = n - i - 1$ and substituting it into the arbitrage condition, Eq. 14.46, we obtain the arbitrage condition for the discount factor $B(0, n - 1)$ at the lowest node of the time slice $(n - 1)$:

$$B(t, t + n\delta t) = B(0, n - 1) \sum_{i=0}^{n-1} G(i, n - i - 1) \prod_{k=1}^{i} \alpha(i - k, n - i + k - 2) \,,$$
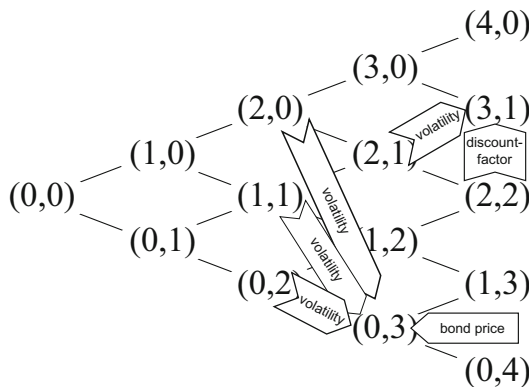
Fig. 14.3 Flow of information when constructing the short rate tree. The discount factor at the lowest node needs the market price of the zero bond maturing one time step *later* and *all* volatility information from one time step *earlier*. For the other discount factors in the time slice it suffices to know the discount factors already calculated at lower nodes in the same time slice and the volatility at the neighboring node one time step earlier

which, after a simple change of index $n \rightarrow (n + 1)$, yields the arbitrage condition for the discount factor $B(0, n)$ at the lowest node of time slice n:

$$B(t, t + (n + 1)\delta t) = B(0, n) \sum_{i=0}^{n} G(i, n - i) \prod_{k=1}^{i} \alpha(i - k, n - i + k - 1) \ . $$

$$(14.63)$$

The left-hand side is the given *market* price of a zero bond with maturity at time slice $(n+1)$, which must be reproduced using zero bonds and the Green's function on time slice n. The α's on the right-hand side are all defined on time slice $(n - 1)$ and are given by the volatility governing the process at this time point. This condition now actually contains only one unknown and can be solved easily for $B(0, n)$:

$$B(0, n) = \frac{B(t, t + (n + 1)\delta t)}{\sum_{i=0}^{n} G(i, n - i) \prod_{k=1}^{i} \alpha(i - k, n - i + k - 1)} \ . $$

From this single value $B(0, n)$ we obtain all further instantaneous discount factors in this time slice by repeated application of Eq. 14.62. Note that we find ourselves at time n at this point of the iteration. The instantaneous discount factors in the time slice n are determined from the price of the bond maturing at time $(n + 1)$ and volatility information from the immediately preceding time step $(n - 1)$. This is illustrated in Fig. 14.3.

Having established the instantaneous discount factors, an expression for the instantaneous short rates results immediately from Eqs. 14.36 and 14.62

$$r(i, j) = -\frac{1}{\delta t} \ln B(i, j) = -\frac{\ln B(i - 1, j + 1)}{\delta t} - \frac{\ln \alpha(i - 1, j)}{\delta t}$$

$$= r(i - 1, j + 1) + 2b(i - 1, j)\sqrt{\delta t}$$

$$= \ldots$$

$$= r(0, j + i) + 2\sqrt{\delta t} \sum_{k=1}^{i} b(i - k, j + k - 1) \ . \qquad (14.64)$$

For volatility structures depending only on time (but not on the interest rate), the volatility values on time slice n are all identical. We can therefore set them

equal to the volatility at the lowest node $(0, n)$. Thus

$$b(i, j) = b(0, i + j) = b(0, n) \equiv \sigma(t + n\delta t)$$
$$\alpha(i, j) = \alpha(0, i + j) = \alpha(0, n) \quad \forall i, j \text{ mit } i + j = n , \quad \quad (14.65)$$

and Eq. 14.62 reduces to

$$B(i, j) = \alpha(0, n - 1)B(i - 1, j + 1)$$
$$= \alpha^i(0, n - 1) \, B(0, n) \quad \text{mit } i + j = n .$$

Equation 14.61 then implies that the short rate in the tree at time slice n changes from node to node by a constant term $2\sigma\sqrt{\delta t}$

$$r(i, j) = r(i - 1, j + 1) + 2\sigma(t + (n - 1)\delta t)\sqrt{\delta t}$$
$$= r(0, n) + 2i\,\sigma(t + (n - 1)\delta t)\sqrt{\delta t} . \quad \quad (14.66)$$

The arbitrage condition for this discount factor at the lowest node of time slice n reduces to

$$B(t, t + (n + 1)\delta t) = B(0, n) \sum_{i=0}^{n} G(i, n - i)\alpha^i(0, n - 1) .$$

14.8.3 Lognormal Models

An analogous procedure for lognormal models can be obtained as follows: in models of the form specified in Eq. 14.58, $b(i, j)^2 \delta t$ is the variance of the *logarithm* of the short rate. Therefore, as seen from the node (i, j), we have

$$b(i, j)\sqrt{\delta t} = \sqrt{\text{Var}\,[\ln r(i, j)]} = \frac{1}{2}\,[\ln r(i + 1, j) - \ln r(i, j + 1)] = \frac{1}{2}\ln(\frac{r(i + 1, j)}{r(i, j + 1)}$$
$$\implies \quad r(i + 1, j) = r(i, j + 1) \exp\left\{2b(i, j)\sqrt{\delta t}\right\} \quad \quad (14.67)$$

This enables us to establish a recursion formula[13] (after performing the substitution $i + 1 \rightarrow i$) for the instantaneous discount factors at all nodes

---

[13] Here we use the property $\exp\{ax\} = (\exp\{x\})^a$ of the exponential function with $a = e^{2b(i-1, j)\sqrt{\delta t}}$.

in time slice $n = i + j$

$$B(i, j) = \exp\{-r(i, j)\delta t\} = \exp\left\{-r(i - 1, j + 1)e^{2b(i-1,j)\sqrt{\delta t}}\delta t\right\}$$

$$= \left[\exp\{-r(i - 1, j + 1)\delta t\}\right]^{e^{2b(i-1,j)\sqrt{\delta t}}} = B(i - 1, j + 1)^{e^{2b(i-1,j)\sqrt{\delta t}}},$$

or

$$B(i, j) = B(i - 1, j + 1)^{\alpha(i-1,j)} \quad \text{mit } \alpha(i, j) = \exp\left\{+2b(i, j)\delta t^{1/2}\right\}.$$
$$(14.68)$$

This is quite similar to the corresponding equation for normal models (Eq. 14.62), but since the volatility information enters the equation as an exponent rather than a factor in the recursion, the structure is somewhat more complicated. Furthermore, a comparison with Eq. 14.62 reveals a sign change in the expression for α. Nevertheless, a simple trick enables us to find a structure which is quite similar to that given in Eq. 14.62. Taking logarithms in Eq. 14.68 yields a recursion relation for the *logarithm* of the discount factors having the same structure as Eq. 14.62:

$$\ln B(i, j) = \alpha(i - 1, j) \ln B(i - 1, j + 1).$$

This recursion can be carried out explicitly giving

$$\ln B(i, j) = \alpha(i - 1, j) \ln B(i - 1, j + 1)$$
$$= \alpha(i - 1, j)\alpha(i - 2, j + 1) \ln B(i - 2, j + 2)$$
$$= \alpha(i - 1, j)\alpha(i - 2, j + 1)\alpha(i - 3, j + 2) \ln B(i - 3, j + 3)$$
$$= \cdots$$
$$= [\ln B(0, j + i)] \prod_{k=1}^{i} \alpha(i - k, j + k - 1),$$

and allowing the instantaneous discount factors in time slice n to be written as a function of the "lowest" node $(0, i + j)$:

$$B(i, j) = \exp\left\{[\ln B(0, j + i)] \prod_{k=1}^{i} \alpha(i - k, j + k - 1)\right\}$$
$$= B(0, j + i)^{\prod_{k=1}^{i} \alpha(i-k,j+k-1)}.$$

Using this $B(i, j)$ for $j = n - i - 1$ in the arbitrage condition Eq. 14.46 we obtain, after performing an index transformation $n \to (n + 1)$, the arbitrage condition for the discount factor $B(0, n)$ at the lowest node in time slice n.

$$B(t, t + (n + 1)\delta t) = \sum_{i=0}^{n} G(i, n - i) \exp \left\{ [\ln B(0, n)] \prod_{k=1}^{i} \alpha(i - k, n - i + k - 1) \right\}$$

$$(14.69)$$

$$= \sum_{i=0}^{n} G(i, n - i) B(0, n)^{\prod_{k=1}^{i} \alpha(i-k,n-i+k-1)} .$$

This can only be solved *numerically* for $B(0, n)$ using, for example, the well-known *Newton-Raphson method.*[14] From this value $B(0, n)$ we obtain all further instantaneous discount factors in this time slice by repeated application of Eq. 14.68. We now have a similar situation as for normal models, see Fig. 14.3: the discount factors in the time slice n are calculated from the price of the zero bond maturing in the *following* time slice $(n + 1)$ and from the volatility information from the immediately *preceding* time slice $(n - 1)$.

Having established the instantaneous discount factors, the short rates follow immediately from Eqs. 14.36 and 14.68.

$$r(i, j) = -\frac{1}{\delta t} \ln B(i, j) = -\alpha(i - 1, j) \frac{\ln B(i - 1, j + 1)}{\delta t}$$

$$= r(i - 1, j + 1) \alpha(i - 1, j) = \ldots$$

$$= r(0, j + i) \prod_{k=1}^{i} \alpha(i - k, j + k - 1) . \qquad (14.70)$$

---

[14]To solve a non-linear equation of the form $f(x) = 0$, the *Newton-Raphson method* uses the following iteration to find the points were the function f equals zero: having an estimate x_i for a zero of f, a better estimate is obtained from the formula

$$x_{i+1} = x_i - f(x_i) \left(\frac{\partial f}{\partial x} \Big|_{x=x_i} \right)^{-1} .$$

We usually start the procedure with a rough estimate x_0 and iterate until the difference between x_{i+1} and x_i is sufficiently small for the required purpose. The iteration sequence converges if

$$\left| f \frac{\partial^2 f / \partial^2 x}{(\partial f / \partial x)^2} \right| < 1$$

holds in a neighborhood of the zero of f. This can always assumed to be the case in our applications.

Exact Reproduction of the Term Structure with the Lognormal Model

Alternatively, we can proceed from the process for the short rate in the log-normal model obtained from Ito's lemma, Eq. 14.59, and derive all equations exactly in the form obtained for the normal model, in particular Eq. 14.63, the only difference being that α is then given by

$$\alpha(i, j) = \exp\left\{-2b(i, j)r(i, j)\delta t^{3/2}\right\} . \tag{14.71}$$

The $r(i, j)$ needed for this equation were calculated in the previous iteration step, so that all terms in Eq. 14.63 are known. Thus, this equation can be solved *analytically* for the lowest bond, even though we are working within the context of the lognormal model.

Both of these two possible methods, i.e., the standard method of solving Eq. 14.69 by means of the Newton-Raphson method, as well as the more elegant method through Eqs. 14.59 and 14.63 with α as in Eq. 14.71, are demonstrated in detail in the Excel workbook TERMSTRUCTUREMODELS.XLSX from the download section [50].

With the interest rate independency of the volatility structure Eq. 14.65 holds again. Thus, Eq. 14.62 reduces to

$$B(i, j) = B(i - 1, j + 1)^{\alpha(0,n-1)}$$
$$= B(0, n)^{\alpha^i(0,n-1)} \text{ with } i + j = n$$

where $\alpha(i, j) = \exp\left\{2\sigma(t + n\delta t)\sqrt{\delta t}\right\}$. Equation 14.67 implies that the short rate is simply to be multiplied by a constant factor $\alpha(0, n - 1)$ when moving from one node to the next in time slice n of the tree, thus

$$r(i, j) = r(i - 1, j + 1)\alpha(0, n - 1)$$
$$= r(0, n)\alpha^i(0, n - 1) , \tag{14.72}$$

and the arbitrage condition for the discount factor at the lowest node of the time slice n reduces to

$$B(t, t + (n + 1)\delta t) = \sum_{i=0}^{n} G(i, n - i)B(0, n)^{\alpha^i(0,n-1)} .$$

14.9 The Example Program TermStructureModels.xlsm

14.9.1 Construction of Interest Rate Trees and Option Pricing

We now demonstrate explicitly how the interest rate tree is constructed for a given term structure, or equivalently, for given market prices of zero bonds,

$$B(t, t + i\delta t) \text{ for } i = 1 \ldots n$$

and given volatilities

$$b(r, t + i\delta t) \text{ for } i = 0 \ldots n - 1 .$$

The calculation of option prices from this interest rate tree—once it has been constructed—has already been discussed in Sect. 14.7.2 in detail for any arbitrary arbitrage-free short rate term structure model using caplets as an example.

- Time step $i = 0$
  According to Eq. 14.47, or in particular according to Eq. 14.41, the Green's function at the node $(0, 0)$ is simply

$$G(0, 0) = 1 .$$

  From Eq. 14.46, it then follows that the discount factor at node $(0, 0)$ can be obtained directly from the market price $B(t, t + 1\delta t)$ of the zero bond:

$$B(t, t + 1\delta t) = G(0, 0)B(0, 0) = B(0, 0) .$$

- Time step $i = 1$
  The discount factors just computed together with the Green's function evaluated at the node $(0, 0)$ are substituted into the recursion relation Eq. 14.43 to determine the values of the Green's function at the next nodes:

$$G(1, 0) = \frac{1}{2}G(0, 0)B(0, 0)$$

$$G(0, 1) = \frac{1}{2}G(0, 0)B(0, 0) .$$

These values, the volatility information from the previous time slice and the price of the zero bond maturing at the next time slice are all used to obtain the discount factor at the lower boundary which here is the node $(0, 1)$. With Eq. 14.63 for the normal model, the discount factor at this node is obtained as

$$B(0, 1) = \frac{B(t, t + 2\delta t)}{G(0, 1) + G(1, 0)\alpha(0, 0)} \;.$$

Application of the recursion relation in Eq. 14.62 yields the other discount factor required for this step

$$B(1, 0) = \alpha(0, 0)B(0, 1) \;.$$

Analogously, the arbitrage condition for the discount factor on the lower boundary in the *log*normal model is, according to Eq. 14.69,

$$B(t, t + 2\delta t) = G(0, 1)B(0, 1) + G(1, 0)B(0, 1)^{\alpha(0,0)} \;,$$

which can only be solved numerically (using the Newton Raphson method, for example) for $B(0, 1)$. Having solved the equation, the next discount factor for this time slice can be obtained immediately using Eq. 14.68:

$$B(1, 0) = B(0, 1)^{\alpha(0,0)} \;.$$

- Time step $i = 2$
  The discount factors just calculated together with the Green's function are substituted into the recursion 14.43 to obtain the values of the Green's function evaluated at the nodes corresponding to this time step

$$G(1, 1) = \frac{1}{2}G(1, 0)B(1, 0) + \frac{1}{2}G(0, 1)G(0, 1)$$

$$G(2, 0) = \frac{1}{2}G(1, 0)B(1, 0)$$

$$G(0, 2) = \frac{1}{2}G(0, 1)B(0, 1) \;.$$

Using these values as well as the volatility information from the previous time slice and the market price of the zero bond maturing at the next time

slice, Eq. 14.63 for the normal model can be applied to obtain the discount factor at the lower boundary, i.e., at node $(0, 2)$

$$B(t, t + 3\delta t) = B(0, 2) \sum_{i=0}^{2} G(i, 2 - i) \prod_{k=1}^{i} \alpha(i - k, 1 - i + k)$$

$$= B(0, 2)G(0, 2)$$
$$+ B(0, 2)G(1, 1)\alpha(0, 1)$$
$$+ B(0, 2)G(2, 0)\alpha(1, 0)\alpha(0, 1) ,$$

which yields for $B(0, 2)$

$$B(0, 2) = \frac{B(t, t + 3\delta t)}{G(0, 2) + G(1, 1)\alpha(0, 1) + G(2, 0)\alpha(1, 0)\alpha(0, 1)} .$$

Applying the recursion formula 14.62 yields the next discount factors belonging to this time step

$$B(1, 1) = \alpha(0, 1)B(0, 2)$$
$$B(2, 0) = \alpha(1, 0)B(1, 1) .$$

Analogously for the *log*normal model, the arbitrage condition Eq. 14.69 for the discount factors on the lower boundary is used to obtain

$$B(t, t + 3\delta t) = G(0, 2)B(0, 2) + G(1, 1)B(0, 2)^{\alpha(0, 1)} + G(2, 0)B(0, 2)^{\alpha(1, 0)\alpha(0, 1)} ,$$

where again we can be solve for $B(0, 2)$ numerically. Once $B(0, 2)$ is known, the additional discount factors for this time step can be computed immediately using Eq. 14.68:

$$B(1, 1) = B(0, 2)^{\alpha(0, 1)}$$
$$B(2, 0) = B(1, 1)^{\alpha(1, 0)} .$$

This procedure is repeated until the entire tree has been constructed up to maturity $T + \tau$ (maturity of the derivative to be priced plus the lifetime of the underlying). After this has been done, all required Arrow-Debreu prices can be determined using Eqs. 14.47, 14.48 and 14.49 and finally, these ADPs are used in pricing the derivative as demonstrated in Sect. 14.7.2 for caplets.

This is all demonstrated explicitly for normal models (referred to there as Ho-Lee models) and lognormal models (referred to as Black-Derman-Toy models) in the Excel workbook TERMSTRUCTUREMODELS.xlsm from the download section [50]. To emphasize that the interest rate tree is independent of the derivative being priced, the structure of the Visual Basic code is extremely modular: first, a short rate tree is generated. This remains in the main memory of the computer until it is used to compute the value of any desired underlying dependent on the term structure (for example a 3-month zero bond yield or a swap rate, etc.) using Arrow-Debreu prices and subsequently pricing any chosen (path-independent) derivative on that underlying with a European payoff mode. Plain vanilla caplets and floorlets on the 3-month rate serve as examples, with an explicit demonstration of their valuation being included in the workbook.

14.9.2 Absolute and Relative Volatilities

As already mentioned, the volatility input for lognormal models has to have the form of a *relative* volatility (for example, 14% of the current underlying value). For normal models, an *absolute* volatility (for example, 0.75 percentage *points*[15]) is required.

If, for comparison, we would like to transform absolute volatilities into relative ones and vice versa, an appropriate reference rate is required in addition. Here, the *forward rate of the underlying at the maturity of the derivative being priced* is a natural choice. This is especially true if we intend to compare the volatilities with the Black-76 volatility, since it is exactly this forward rate which is used as the underlying in the Black-76 model.[16] Therefore the Black-76 volatility quoted in the market belongs to this forward rate. For these reasons, this forward rate has been selected as the multiplicative factor in the Excel workbook TERMSTRUCTUREMODELS.xlsm.

As can be seen from Eq. 14.59, the factor of the stochastic term differs by the additional multiplication with the short rate. The product of relative

---

[15]At an interest rate of 6% this would correspond to a relative volatility of 14%.

[16]As emphasized in Eq. 14.55, the forward rate for the caplet period τ within *linear* compounding serves as the underlying of a cap in the payoff profile. This has already been accounted for in the term structure model through Eq. 14.55. To obtain the correct forward rate as input for the *Black-76* model in the Excel workbook TERMSTRUCTUREMODELS.xlsm from the download section [50] from the current term structure (which holds for continuous compounding), we first determine the forward *discount factors* using Eq. 2.7. From those discount factors the desired forward rates for linear compounding are given by $r = \left(B^{-1} - 1\right)/\tau$.

volatility and short rate is equivalent to the absolute volatility in a Gaussian or normal model. Since the short rate is different at each note for a given time slice, the absolute volatility is different at each node in a log-normal model, while it was the same at each node in a normal model. This is the substantial difference between both models. Indeed, it is enough to change a single line in the implementation of the Gaussian model to transform it into the implementation of a log-normal model according to Eq. 14.59.

14.9.3 Calibration of Volatilities

The current term structure and volatility structure are required as input for the construction of an interest rate tree. In practice, the volatilities $b(r, t + i\delta t)$ themselves are generally not known. Only the prices of options traded on the market (caps, floors, swaptions, etc.) can be observed directly. The tree must then be constructed, leaving the volatilities unspecified as free parameters which are then adjusted until the observed market prices of the options are reproduced by the model. Fitting the parameters $b(r, t + i\delta t)$ to the market prices in this manner is referred to as the *calibration* of the model.

There are many ways of performing such a calibration. For instance we can—exactly as was done when reproducing the market prices of zero bonds—reproduce the market prices of the options stepwise through the tree, beginning with the shortest option lifetimes and proceeding through to the longest. This procedure is demonstrated explicitly in the Excel workbook TermStructureModels.xlsm from the download section [50] where the prices of a strip of caplets at the 3-month rate are available, whose underlyings (the respective 3-month rates) cover the time span under consideration without overlap, i.e.

$$T_{k+1} = T_k + \tau \ .$$

where T_k is the maturity of the kth caplet and τ is the lifetime of the underlying. The calibration starts by assuming that the volatility in the tree between times t and $T_1 + \tau$ is constant,

$$b(r, t + i\delta t) = b(T_1) \quad \text{for all } r \text{ and all } i \text{ with } t \le t + i\delta t \le T_1 + \tau \ .$$

This constant volatility is adjusted (using Newton-Raphson) until the price computed using the tree equals the known market price of the first caplet.

To reproduce the second known caplet price, we assume that the volatility remains equal to the (just calibrated) $b(T_1)$ for the time between t and $T_1 + \tau$, and is constant (although still unknown) at all nodes for times between $T_1 + \tau$ and $T_2 + \tau$:

$$b(r, t + i\delta t) = b(T_2) \quad \text{for all } r \text{ and all } i \text{ with } T_1 + \tau < t + i\delta t \le T_2 + \tau \,.$$

This constant volatility is then adjusted (again using Newton-Raphson) so that the price for the second caplet computed using the tree is equal to the observed market price of this caplet. To reproduce the third caplet price we assume, as above, that the (just calibrated) volatilities computed for the time span from t to $T_2 + \tau$ continue to be valid, and adjust the volatility at all nodes associated to the times between $T_2 + \tau$ and $T_3 + \tau$ until the price of the third caplet computed using the tree matches the observed market price of this caplet, and so on.[17] By this method, we obtain a piecewise constant function for the volatility[18] as a function of time, independent of r.

Such a calibration process yields different volatility values $b(r, t + i\delta t)$ for each different term structure model. We therefore refer to the respective volatilities by the name of the model with we are working, for example *Ho-Lee volatilities*, *Black-Derman-Toy volatilities*, *Hull-White volatilities*, etc. For this reason, it is generally not possible to simply take the Black-76 volatilities as the input values for $b(r, t + i\delta t)$. The tiresome process of calibrating to the observed option *prices* is in most cases unavoidable.

The market quotes option prices often in terms of log-normal volatilities (or, recently, normal model vols), from which prices could be calculated by applying the Black-Scholes resp. Black'76 or Bachelier (for normal vols) formula. Since all other input parameters, e.g. underlying interest rates, are commonly known, the Black-76 model is simply a translation algorithm for moving between the two different ways of quoting the option price, i.e. price vs. volatility.[19] An advantage of quoting volatilities rather than prices is that volatilities change less frequently than other parameters like interest rates. Also, volatilities offer an illustrative, easy to interpret way of comparing options with different strikes and terms to expiry. Therefore, the Black-76 model (or the

---

[17] Alternatively, the volatilities can be (simultaneously) calibrated using a least squares fit. We then minimize the sum of the quadratic differences between the calculated and the traded option prices by varying the volatilities.

[18] Naturally, this procedure can be extended if a "caplet-price-surface", i.e. caplet prices with different times to maturity *and* different strikes, is available to obtain a calibrated volatility surface as a function of time *and moneyness* (the relative or absolute difference between underlying and strike).

[19] Though, in some cases, market makers quote both, prices and volatilities, at the same time.

equivalent normal model) is better understood as a quoting vehicle rather than a pricing model. The procedure behind the quoting is as follows: the market maker calculates the price of an interest rate option which she wishes to quote to the market using the term structure model of her choice (maybe calibrated to other liquid option prices). Before quoting, this price is first translated into a volatility using the Black-76 model. It is this volatility that is quoted on the market. The volatility quoted is simply that which, when used as an input parameter in the Black-76 model, reproduces the price calculated with the bank's (perhaps very complicated and proprietary) term structure model.

The calibration to given (Black-76) caplet (or floorlet) prices are demonstrated in the Excel Workbook TermStructureModels.xlsm from the download section [50]. The calibration to other instruments such as swaption prices is considerably more complicated in its implementation, but is based on the same principles:

1. Calculation of all required Arrow-Debreu prices from the existing tree.
2. Generation of the necessary underlyings, namely the swap rate under consideration, from the Arrow-Debreu prices.
3. Calculation of the payoff profiles of the swaption at the nodes corresponding to the swaption maturity date based on the value of the underlying (swap rate) and with the help of the Arrow-Debreu prices
4. Calculation of the swaption prices at node $(0, 0)$ by discounting the payoff profile with the Green's function
5. Adjusting the volatility in the interest rate tree until the swaption price computed using the tree agrees with the price quoted on the market.

14.10 Monte Carlo on the Tree

In the discussion above, it was emphasized that only path-independent derivatives in the sense of Eq. 14.44 could be priced using the methods introduced in this chapter since the trees recombine and as a result information on the history of the short rate path is lost. However, by combining the presented trees with Monte-Carlo simulations, it becomes possible to price path-dependent derivatives with a pay off profile depending on the past realizations of the interest rate This kind of Monte-Carlo simulation works as follows:

• First observe that the tree needs to be generated (and calibrated!) only once (as described above) and then stays in the memory of the computer.

- The short rate paths are then simulated by jumping randomly from node to node in the tree, always proceeding one time step further with each jump.
- Simulating the jumps from node to node with the appropriate transition probabilities (in our case $p = 1/2$ for up as well as for down moves, see Eq. 14.32) of the chosen martingale measure (in our case the risk-neutral measure, see Eq. 14.31) ensures that the simulated paths already have the correct probability weight needed for pricing financial instruments. This procedure is called *importance sampling*.[20]
- For each simulated path of the short rate the corresponding path of the underlying (for instance 3-month LIBOR) must be calculated using Arrow-Debreu prices.
- At the end of each simulated path the payoff of the derivative resulting from the underlying having taken this path is calculated.
- This payoff is then discounted back to the current time t (i.e., to node $(0, 0)$) at the short rates *along the simulated path*, since after all, we still are within the risk-neutral measure, see Eq. 14.31.
- After many (usually several thousand) paths have been simulated, the (several thousand) generated discounted payoff values can then simply be averaged to yield an estimation for the risk-neutral expectation of the discounted payoff. Here the very simple arithmetic average (with equal weights) of the discounted payoff values can be used without worrying about the correct probability weight of each payoff value since the payoff values (more precisely the paths which generated the payoff values) have already been *simulated* with the correct (risk-neutral) probability. This is (besides the effectiveness in sampling the phase space) another great advantage of importance sampling.
- According to Eq. 14.8, the risk-neutral expectation calculated in this way is directly the desired derivative price.

---

[20]Doing a move with its associated probability only ensures a very effective sampling of the *phase space* (the set of all possible values of the simulated variables): phase space regions (values of the simulated variables) which have low probabilities (and therefore contribute only little to the desired averages of whatever needs to be measured by the simulation) are only visited with low probability (i.e. rarely) while phase space regions with high probabilities (which contribute a lot to the desired averages) are visited with high probability (i.e. often). Because of this feature *importance sampling* is heavily used in thousands of Monte-Carlo applications, especially in physics, meteorology and other sciences which rely on large-scale simulations.

14.11   The Drift in Term Structure Models

At this stage, it makes sense to reconcile the practical computations in Chap. 14 with the profound concepts from the related theory presented in Chap. 13. We first consider the relationship between the underlying instantaneous interest rates and tradable instruments such as zero bonds. We defer proceeding with the short rate models introduced above to take a brief excursion to the models in which the instantaneous *forward* rate plays the role of the underlying.

14.11.1   Heath-Jarrow-Morton Models

We see from the fundamental Eqs. 14.2 or 14.3 and 14.5 that the instantaneous *forward* rates can be used to generate the prices $B(t, T)$ of all zero bonds as well as the entire interest rate term structure $R(t, T)$. The three descriptions of the term structure, the zero bond prices $B(t, T)$, the zero bond yields $R(t, T)$ and the instantaneous forward rates $f(t, T)$ are equivalent. Of these three variables, only a single one needs to be chosen to be modeled by a general stochastic process of the form specified in Eq. 2.19. We take, for example, the forward rates $f(t, T)$ to be modeled by a process of the form:

$$df(t, T) = a(t, T)\, dt + b(t, T)\, dW \quad \text{with} \quad dW = X\sqrt{dt},\ X \sim \text{N}(0, 1)\ .$$

$$(14.73)$$

As mentioned previously, all bond prices and thus the entire term structure can be generated from the solution of this equation. An entire class of term structure models, the *Heath-Jarrow-Morton models* (*HJM models* for short) take this approach of employing the forward rates as the driving factor of the term structure.

   Like all interest rates, the instantaneous forward rates are not tradable (see Sect. 14.4). In Chap. 13, and in particular in Sects. 13.4 and 13.5, a detailed discussion can be found on how to proceed when the underlying is not tradable; we choose a tradable instrument U whose price $U(S, t)$ is a *function* of the underlying. Then all of the results shown in Chap. 13 hold:

- The Harrison-Pliska Theorem establishes the *uniqueness* (in complete markets) of the probability measure with respect to which the prices of tradable financial instruments normalized with an arbitrarily chosen, tradable numeraire instrument Y are martingales.
- According to the Girsanov Theorem, this implies that there is only *one single underlying drift* which may be used for pricing.

- In addition, the drift of the underlying in the real world plays absolutely no role in the world governed by the martingale measure if the numeraire instrument Y satisfies the property 13.34 (which is always the case).

The existence of a function $U(S, t)$ relating the underlying S to a tradable instrument U is the deciding factor for the validity of the above results. If the underlying is the instantaneous forward rate $f(t, T)$, such a functional relation to a tradable instrument exists, namely Eq. 14.3 for the zero bond price $B(t, T)$, and thus the results found in Chap. 13 hold for the Heath-Jarrow-Morton models. In particular, for every complete market, there exists for each numeraire instrument exactly one single underlying drift which may be used for pricing. Therefore, as far as pricing is concerned, the HJM model is uniquely determined by the specification of the volatility term $b(t, T)$ in Eq. 14.73 (which models the process in the real world). The Girsanov Theorem implies that the transition from the real world into the world governed by the martingale measure only effects a (in this case unique) change in the drift; the volatility term $b(t, T)$ is invariant under this transformation. Indeed, the forward rate process in the risk-neutral measure corresponding to the real world process in Eq. 14.73 has the following appearance [92]:

$$df(t, T) = \left[b(t, T) \int_t^T b(t, s) ds \right] dt + b(t, T) d\widetilde{W} ,$$

Here $d\widetilde{W}$ denotes the standard Brownian motion with respect to the martingale measure. The coefficient of dt appearing in square brackets is the drift with respect to the martingale measure. This formula shows *explicitly* that for HJM models, the entire model (including the drift) is uniquely specified through the volatility $b(t, T)$. The drift to be used in the valuation is *unique,* in complete agreement with the general statements made in Chap. 13.

14.11.2 Short Rate Models

As opposed to the HJM models, the term structure models in Sect. 14.6 make use of the instantaneous *spot* rate defined in Eq. 14.1 as the driving factor. However, a one-to-one mapping between the spot rates and the zero bond prices does *not* exist and in consequence, *no* one-to-one mapping between the spot rates and the term structure $R(t, T)$ can exist either. The instantaneous spot rates are not sufficient to generate the term structure. This is indicated by the fact that the instantaneous spot rate $r(t)$ is a function of a single time

variable t in contrast to the processes $B(t, T)$, $R(t, T)$ (and $f(t, T)$ as well!), which are functions of *two* time variables. Taking the limit $dt \to 0$ in the definition of the instantaneous short rate in Eq. 14.1 results in the loss of the second argument (and thus in the loss of the corresponding information). This can be seen explicitly in Eq. 14.6. For this reason, there is *no* analogy to Eq. 14.3 relating the instantaneous spot rates directly to the zero bond prices. The best possible alternative available is to determine the bond prices from the *expectations* of the short rates (see for example Eq. 14.9), but not directly as a *function* of the short rates. This has significant consequences:

The results presented in Chap. 13, in particular those in Sects. 13.4 and 13.5 (regarding the martingale measure, unique drift, etc.), can be shown for non-tradable underlyings only if there exists a tradable instrument whose price process is a *function* of the underlying. This *direct functional relationship* between the underlying (the instantaneous spot rate) and a tradable instrument is *missing* in short rate models. Or from the view point of the Harrison-Pliska Theorem: since $r(t)$ contains less information than $f(t, T)$ or $B(t, T)$, the market is not complete for short rate models. For this reason, the martingale measure in short rate models is *not uniquely determined* by fixing the numeraire instrument. Thus, the Girsanov theorem asserts that we retain the freedom of choosing from various drift terms in the model. The information lost in the transition shown in Eq. 14.6, for example, must be reinserted into the model "by hand". This is accomplished by directly specifying a drift in the world governed by the *martingale measure.* This is the essential difference in the models here compared to those encountered in the previous chapters, where the drift was always specified in the *real* world. The situation for short rate models is different; the drift is specifically chosen for the world governed by the *martingale measure* rather than for the real world. Only through this drift specification is the martingale measure uniquely determined in short rate models.

Frequently, it is seen as a requirement that the has drift shows an effect called *mean reversion* which is observed in the evolution of interest rates but not seen in e.g., stock prices. Interest rates do not rise or fall to arbitrarily high or low levels but tend to oscillate back and forth about a long-term mean. This can be modeled with a drift in the functional form $\mu - vr$ for some $v > 0$: for values of r small enough so that $vr < \mu$ holds, the drift is positive and, consequently, r tends on average toward larger values. Conversely, for values of r large enough so that $vr > \mu$ holds, the drift is negative and r tends to drift toward smaller values on average. The interest rate thus tends to drift toward a mean value μ at a rate v. Naturally, a stochastic component driven by $\sim dW$ is superimposed onto this deterministic movement. An example of

a *mean reversion model* is the *Hull-White model* [104]. This model specifies the following stochastic process for the short rate with respect to the *risk-neutral martingale measure:*

$$dr(t) = [\mu(t) - v(t)r] \, dt + \sigma(t) \, dW \tag{14.74}$$

This is only one of many examples. Several well-known models, each named after their respective "inventors" can be distinguished from one another, after having separated them into categories of normal and lognormal models, essentially through the form of their drift. The best known representatives of these models are summarized in the following list:

- Normal models $dr(t) = a(r, t)dt + b(r, t) \, dW$

  - Stationary models $b(r, t) = \sigma$

    * Arbitrage-free models

      · Hull-White $a(r, t) = \mu(t) - vr$ (mean reverting)
      · Ho-Lee $a(r, t) = \mu(t)$

    * Equilibrium models (not arbitrage free because of too few degrees of freedom)

      · Vasiceck $a(r, t) = \mu - vr$ (mean reverting)
      · Rendleman-Barter $a(r, t) = \mu$

  - Non-stationary models $b(r, t) = \sigma(t)$

- Lognormal models $d \ln r(t) = a(r, t)dt + b(r, t) \, dW$

  - Stationary models $b(r, t) = \sigma$
  - Non-stationary models $b(r, t) = \sigma(t)$

    * Arbitrage-free models

      · Black-Karasinski $a(r, t) = \mu(t) - v(t) \ln r$ (mean reverting)
      · Black-Derman-Toy $a(r, t) = \mu(t) - \frac{\partial \sigma(t)/\partial t}{\sigma(t)} \ln r$

The models are called either *stationary*[21] or *non-stationary* depending on whether or not the volatility is assumed to be a function of time.[22] The models

---

[21]This is not to be confused with the definition of stationary time series in Chap. 32.

[22]Ho-Lee and Hull-White are often applied for time-dependent volatilities. Their inventors, however, originally assumed constant volatilities.

allowing for neither a time dependence nor an r dependence in the drift or volatility terms (Vasiceck, Rendleman-Barter) obviously cannot reproduce the current term structure arbitrage-free. For these models, only a "best fit" can be obtained (for example by minimizing the *root mean square error*). These models are referred to as *equilibrium models*. Since equilibrium models are not arbitrage free, they will receive no further attention in this book. Nowadays they are no longer very important in practice.

The volatility of dr is independent of r in normal models and proportional to r in lognormal models (see, Eq. 14.59). An intermediate scheme between these two possibilities is the *Cox-Ingersoll-Ross* model [42] for which the volatility is assumed to be proportional to $\sqrt{r}$.

$$dr = (\mu - vr)dt + \sigma\sqrt{r}dW \ .$$

Note that all these processes are modeled in the *risk-neutral martingale measure* directly, i.e., should be used directly in the valuation of the financial instrument under consideration without first performing a drift transformation through an application of the Girsanov Theorem. The volatility term is invariant under the Girsanov transformation. This implies that the volatility taken for the valuation is the same as that observed in the real world. The form taken on by the drift, however, is a result of the particular choice of the measure (coordinate system) established for pricing through an application of the Girsanov Theorem and as such, a rather artificial construct. It is thus not readily apparent why a specific form of drift (for example mean reversion) should be modeled in a specific (for example risk-neutral) artificial world (dependent on the selection of a particular numeraire) when our intuitive conception of the drift actually pertains to the *real* world.

Or more precisely: according to the Girsanov Theorem, the process modeled with respect to the martingale measure differs from the real world process by a previsible process $\gamma(r, t)$. This previsible process is *arbitrary* (with the restriction that it must satisfy the boundedness condition $E\left[\exp\left(\frac{1}{2}\int_0^T \gamma(r, t)dt\right)\right] < \infty$). Therefore the choice of model with respect to a martingale measure provides as good as no information about the drift of the short rate in the *real* world. For example, the model given by Eq. 14.74 has a mean reversion in the world governed by the martingale measure, but has the form

$$dr(t) = [\mu(t) - v(t)r(t) + \gamma(r, t)] \, dt + \sigma(t) \, dW$$

in the real world with a (practically) *arbitrary* previsible process $\gamma(r, t)$. It is therefore by no means clear that this process shows any mean reversion in the real world.

After having placed such emphasis on the necessity of specifying a drift, i.e., a martingale measure, explicitly in short rate models (despite the results in Chap. 13 the martingale measure is not unique here, even when the numeraire has been fixed), the attentive reader will surely have asked why only the volatility but no drift information has entered as input into our explicit computations in Sect. 14.8. For both the normal and the lognormal models, the interest rate trees were constructed in their entirety and no drift information from the specific stochastic processes was needed at any point.

This stems from the condition in Eq. 14.32, which we introduced "by hand" for the sake of simplicity; the probability of an up move was simply *set* to $p = 1/2$. Through this choice, we have explicitly selected one particular measure from the family of arbitrage-free martingale measures belonging to the risk-neutral numeraire instrument (the bank account). This fixed the drift in accordance with the Girsanov Theorem. We can also see this fact explicitly since the drift can be quite simply determined from the generated tree. As was shown in Sect. 14.8 for the variance, it follows from the general equation 14.60 for a binomial tree with $p = 1/2$ that the expectations of the short rate as seen from the node (i, j) are

$$
\begin{aligned}
a(i, j)\delta t &= \mathrm{E}\left[r(i, j)\right] = \tfrac{1}{2}\left[r(i+1, j) + r(i, j+1)\right] && \text{normal model} \\
a(i, j)\delta t &= \mathrm{E}\left[\ln r(i, j)\right] = \tfrac{1}{2}\left[\ln r(i+1, j) + \ln r(i, j+1)\right] && \text{log-normal model,}
\end{aligned}
$$

from which the drift at each node in the tree can be immediately determined since all the short rates have already been established (the tree has already been built).

For example, Eq. 14.66 can be used to compute the drift in a normal model with r-independent volatility as

$$
a(i, j)_{\text{normal}} = \frac{r(0, i + j + 1)}{\delta t} + (2i + 1)\frac{\sigma(t + (i + j)\delta t)}{\sqrt{\delta t}} \ .
$$

Analogously, the drift for *log*-normal models with r-independent volatility $b(0, n) = \sigma(t + n\delta t)$ can be computed explicitly, employing Eq. 14.72 and α as in Eq. 14.68:

$$
\begin{aligned}
a(i, j)_{\text{log-normal}} &= \frac{1}{2\delta t} \left[\ln r(i + 1, j) + \ln r(i, j + 1) \right] \\
&= \frac{1}{\delta t} \left[\ln r(0, i + j + 1) + (i + \frac{1}{2}) \ln \alpha(0, i + j) \right] \\
&= \frac{\ln r(0, i + j + 1)}{\delta t} + (2i + 1) \frac{\sigma(t + (i + j)\delta t)}{\sqrt{\delta t}} \; .
\end{aligned}
$$

Had the probability of an up move not been fixed at $1/2$, a free parameter p would remain unspecified in Eq. 14.60 which (if dependent on time and perhaps on r as well) would allow for many different drift functions.

14.12 Short Rate Models with Discrete Compounding

The discount factor over a single time period used in this chapter was always of the form $B(t, t + \delta t) = e^{-r(t)\delta t}$ (see for example Eq. 14.36). Intuitively, in view of Eq. 2.3, this means that interest has been paid infinitely often in the reference period δt, and that these payments were then immediately reinvested at the same rate. Strictly speaking, this contradicts the concept of a tree model, for which time has been *discretized* into intervals of positive length δt, implying by definition that nothing can happen in between these times. To be consistent, we should have therefore used *discrete* compounding, allowing the payment and immediate reinvestment of interest solely after each δt. Then, only in the limiting case $\delta t \rightarrow 0$ will the discount factor for continuous compounding be obtained. If we wish to be consistent, we would therefore have to write

$$
B(t, t + \delta t) = \frac{1}{1 + r(t)\delta t} \xrightarrow[\delta t \to 0]{} e^{-r(t)\delta t} \; .
$$

Despite the inconsistency, the discount factor for continuous compounding is commonly used in the literature. In *this* section, we will collect and present the differences caused by using discrete rather than continuous compounding and, in doing so, show how short rate models with discrete compounding

can be treated. The discount factor in Eq. 14.36 has the following form when adopting the convention of discrete compounding

$$B(i, j) = \frac{1}{1 + r(i, j)\delta t} \ . \tag{14.75}$$

Formulating recursion equations as in Eqs. 14.62 or 14.68 for these discount factors $B(i, j)$ in a time slice or for the lowest zero bond in a time slice (see Eqs. 14.63 and 14.69) is quite awkward. Such conditions are more easily formulated for the *interest rate* $r(i, j)$. All recursion relations for the short rate follow from the arbitrage condition for the Green's function, Eq. 14.46. This expression for the instantaneous discount factor in the form of Eq. 14.75 is

$$B(t, t + n\delta t) = \sum_{i=0}^{n-1} \frac{G(i, n - i - 1)}{1 + r(i, n - i - 1)\delta t} \ . \tag{14.76}$$

14.12.1   Normal Models

If the short rate $r(i, j)$ in Eq. 14.75 is governed by a stochastic process of the form given in Eq. 14.57, then Eq. 14.61 holds. The recursion relation for the short rate in the normal model is given by Eq. 14.64:

$$r(i, j) = r(i - 1, j + 1) + 2b(i - 1, j)\sqrt{\delta t}$$

$$= r(0, j + i) + 2\sqrt{\delta t} \sum_{k=1}^{i} b(i - k, j + k - 1) \ . \tag{14.77}$$

Substituting this into Eq. 14.76 with $j = n - i - 1$ and performing the transformation $n \rightarrow n + 1$, we obtain a condition analogous to Eq. 14.63 for the interest rate $r(0, n)$ at the lowest node in the time slice n

$$B(t, t + (n + 1)\delta t) = \sum_{i=0}^{n} \frac{G(i, n - i)}{1 + r(0, n)\delta t + 2\delta t^{3/2} \sum_{k=1}^{i} b(i - k, n - i + k - 1)} \ .$$

This equation can only be solved numerically for $r(0, n)$. Once this value is known, Eq. 14.77 provides all other $r(i, j)$ on the time slice n. From the $r(i, j)$, the discount factors can then be calculated immediately using Eq. 14.75. Note that in the case of discrete compounding, the arbitrage

condition, Eq. 14.76, can no longer be solved analytically, not even in the context of the normal model.

14.12.2 Lognormal Models

If the short rate $r(i, j)$ in Eq. 14.75 is governed by a stochastic process of the form 14.58, then Eq. 14.67 holds. The recursion relation for the short rate in the lognormal model is given by Eq. 14.70:

$$r(i + 1, j) = r(i, j + 1) \exp\left\{2b(i, j)\sqrt{\delta t}\right\}$$

$$= r(0, j + i) \prod_{k=1}^{i} \alpha(i - k, j + k - 1) \qquad (14.78)$$

with $\alpha(i, j)$ as defined in Eq. 14.68. Substituting this into Eq. 14.76 with $j = n - i - 1$, and performing the transformation $n \rightarrow n + 1$, we obtain the condition on the interest rate $r(0, n)$ at the lowest node in the time slice n analogous to Eq. 14.69

$$B(t, t + (n + 1)\delta t) = \sum_{i=0}^{n} \frac{G(i, n - i)}{1 + r(0, n)\delta t \prod_{k=1}^{i} \alpha(i - k, n - i + k - 1)} .$$

Again, this can only be solved for $r(0, n)$ numerically. Once this has been done, the other values $r(i, j)$ in the time slice n can be calculated immediately using Eq. 14.78. The discount factors are finally determined from the $r(i, j)$ using Eq. 14.75.

14.13 Other Interest Rate Models

The world of interest rate models is not limited to the models presented so far. Instead, there are a couple of other models used in practice. For example, a further, quite important class of models are market models, which in contrast to Heath Jarrow Morton models use forward rates for finite periods as driving factors. The *LIBOR market model* plays here an especially dominant role. Such a model may use, e.g., 6M LIBOR rates as driving factors. For a time horizon of 30 years, a model with 60 non-overlapping forward rates as stochastic factors

could be constructed. Each forward rate follows a log-normal (or, alternatively, normal) process, which are not independent, but correlated with each other. In general, the number of independent stochastic drivers is reduced by means of a principal component analysis (see Chap. 34) to a few factors only (typically, 3 or 4). Each forward rate is a martingale in its own T measure, where T is the end date of the period the interest rate belongs to. By application of a unique numeraire to all 60 forward rates, a drift term is implied by the change of numeraire. For an extensive discussion of the LIBOR market model see for example [4, 21].

Part III

Instruments

15

Simple Interest Rate Products

In Part II, a whole array of very workable methods for the valuation and hedging of financial instruments was introduced. We now continue in Part III with the explicit valuation of the most important and common financial instruments. We restrict our considerations to simple (for the most part, plain vanilla) instruments which still represent the largest proportion of all trades in financial markets today. The methods presented in Part II do in fact allow much more complicated instruments than those introduced here to be effectively priced. Seen from this point of view, the application of the material introduced in Part II is much more extensive than its restriction to the instruments defined in Part III might suggest. More involved applications (such as term structure models for Bermudan swaptions or multi-dimensional finite difference schemes for convertible bonds, etc.) often contain so many details specific to the individual implementation, that it is easy to become distracted from the essential ideas. Therefore such complicated examples are not particularly appropriate for discussion in an introductory text. However, Part II enables the reader to develop pricing techniques for quite complex products even if they don't receive specific treatment in Part III.

In the following sections, the valuation of different plain vanilla instruments will often be explicitly demonstrated on the basis of tables. All of these tables have been extracted from the Excel workbook PLAINVANILLA.xls on the website [50] accompanying this text. The computations contained in this workbook are carried out solely using cell calculations. VBA-programs will not be needed.

15.1 Zero Bonds

Most actively traded zero bonds have terms to maturity of up to one year, but zero bonds do trade occasionally for longer periods, too. Compared to instruments with regular interest payments, zero bonds have a significantly higher exposure against default risk, since the exposure is increasing over time. This is also reflected in a higher duration (Table 15.1). The invested capital is bound for a very long time, making long term zero bonds less attractive as an investment. Traded zero bonds can be separated into *compounding papers* and *discount papers*. In case of a compounding paper, the full notional is paid at the trade date, while at maturity the buyer receives a payment of notional plus interest accrued over the full life time. Therefore, it is just identical to an instrument paying regular interest with only one interest period. A discount paper pays back at maturity the notional only, but is traded at a discount to the full notional (the discounted notional) at trade date. From a valuation perspective, it is only relevant that the buyer receives only a single, deterministic cash flow at maturity. Typical compounding papers are *deposits*, while *commercial papers* are typical representatives of discount papers.

15.1.1 Cash Flows and Present Value

A *zero bond* is a security with a single cash flow: the *redemption* at maturity, i.e., payment of the nominal N at maturity. The cash flow C_T resulting from this payment is:

$$C_T(T) = N .$$

The present value, defined as the sum of all discounted future cash flows, is accordingly

$$V(t) = N B(t, T) . \tag{15.1}$$

15.1.2 Yield to Maturity and Par Rate

A zero bond does not pay coupons at a fixed rate K. At best we could say the coupon rate is $K = 0$ (hence the name). The yield of a zero bond results from the difference between its present value, Eq. 15.1, and its nominal amount N and is thus equal to the current spot rate $R(t, T)$ over the lifetime of the zero

Table 15.1 Present value and sensitivities of a zero bond in the three most commonly used compounding methods. For the sake of simplicity, the abbreviation $\tau = T - t$ is used to represent the time from value date t until maturity T

	Present value	Mod. duration = key rate dur.	Convexity
General	$B(t,T)$	$\frac{-1}{B(t,T)}\frac{\partial B(t,T)}{\partial R}$	$\frac{1}{B(t,T)}\frac{\partial^2 B(t,T)}{\partial R^2}$
Continuous	$e^{-R\tau}$	τ	τ^2
Discrete	$\frac{1}{(1+R)^\tau}$	$\frac{\tau}{1+R}$	$\frac{\tau(\tau+1)}{(1+R)^2}$
Simple	$\frac{1}{1+R\tau}$	$\frac{\tau}{1+R\tau}$	$\frac{\tau^2}{(1+R\tau)^2}$

bond. As explained is Sect. 5.3, a "par rate" only makes sense in the context of fixed-rate instruments and thus will not be defined for zero bonds. Since there is only one cash flow, the yield to maturity (YTM) is equal to the current spot rate until maturity T, which in turn is the only relevant key rate.

15.1.3 Sensitivities

Zero bonds are the basic elements which can be used to replicate all securities whose value derives from future cash flows. This is because the present value of any arbitrary future cash flow is nothing other than the value of a zero bond with a principal equal to that cash flow and maturity equal to the payment date of that cash flow. Because of the immense importance of the zero bond, its present value, modified duration and convexity for all compounding conventions are listed explicitly in Table 15.1. As can be seen from the table, modified duration and convexity depend on the choice of the compounding method. Without explicit specification of the compounding method used, any mentioning of the duration value is incomplete, though most often (discrete) annual compounding can be assumed.

Since the YTM of a zero bond is equal to the spot rate for its maturity, the modified duration and the key rate duration agree. For continuous compounding, the sensitivities have a particularly simple and intuitive form: the modified duration is exactly equal to the time to maturity and the convexity is the square of the time to maturity.

15.2 Forward Rate Agreements

A *forward rate agreement*, abbreviated as *FRA*, is a contract in which both contract parties agree to a fixed rate K on a principal N to be paid for some future interest period between T and T'. At the start T of this period, the

difference to the then valid interest rate for this period from T to T' (e.g., 6M-EURIBOR for a 6 months period; which interest rate index is relevant has to be fixed contractually) will be paid, discounted from T' to T. Therefore, the pay off profile of a FRA is

$$N \frac{(L(T) - K)\tau(T, T')}{1 + L(T)\tau(T, T')} \ ,$$

where $L(T)$ is the interest rate index starting at T and ending at T'. Because of default and liquidity risks, $L(T)$ differs in general from the risk free forward rate, which could be calculated by means of Eq. 2.8, by a small spread, which is called the basis $s(T)$ (see Sect. 15.6.1):

$$L(T) = F(t, t_{\text{fix}}, T, T') + s(T) = \frac{1}{\tau(T, T')} \left(B^{-1}(T, T'|t_{\text{fix}}) - 1 \right) + s(T) \ .$$
$$(15.2)$$

For the EUR-zone, the fixing date t_{fix} lies typically two bank working days before T. Alternatively, $L(t)$ can be calculated by means of a *tenor* or *forward curve*, which takes the basis spread implicitly into account. This yields the expression

$$L(T) = \frac{1}{\tau(T, T')} \left(B_{\text{6M}}^{-1}(T, T'|t_{\text{fix}}) - 1 \right) \ , \tag{15.3}$$

where $B_{\text{6M}}(t, T)$ represents the 6M forward curve (see Sect. 29.7).

An FRA can be interpreted as an agreement loan to be made in the future with an interest rate already fixed today. The party receiving the loan makes the fixed interest payments. In contrast to bonds, we will refer to this party's position as a *long* position in the FRA, whereas the counterparty receiving the interest payments is *short* in the FRA. A (long) FRA can thus be interpreted as an agreement on two future cash flows: a receipt of the principal N at time T (the loan is made) and a payment at maturity T' of the FRA in the amount of the principal N compounded at the agreed rate K over the period $T' - T$ (the loan plus interest is paid back).

The present value $F(T, T', K|t)$ of the FRA at time t is given by discounting the cash flow at time T to $t < T$:

$$
\begin{aligned}
F(T, T', K|t) &= B(t, T)N\frac{(L(T) - K)\,\tau(T, T')}{1 + L(T)\tau(T, T')} \\
&= B(t, T)N\frac{B^{-1}(T, T'|t_{\text{fix}}) - 1 + (s(T) - K)\,\tau(T, T')}{1 + B^{-1}(T, T'|t_{\text{fix}}) - 1 + s(T)\tau(T, T')} \\
&= B(t, T)N\left(1 - \frac{1 + K\tau(T, T')}{B^{-1}(T, T'|t_{\text{fix}}) + s(T)\tau(T, T')}\right),
\end{aligned}
$$
(15.4)

where the definition of the forward rate, Eq. 2.6, was used in the last step.

It is common practice in the market for forward transactions, the interest rate K is chosen so that the contract is worthless at the time it is concluded. As can be deduced from Eq. 15.4, this condition is satisfied if the term in brackets in the last line vanishes:

$$
B^{-1}(T, T'|t_{\text{fix}}) + s(T)\tau(T, T') = 1 + K\tau(T, T')
$$

and therefore

$$
K = \frac{1}{\tau(T, T')}\left(B^{-1}(T, T'|t_{\text{fix}}) - 1\right) + s(T)\,.
$$

This is exactly the case if the fixed rate K equals the *forward rate* at time t corresponding to the FRA period from T until T' plus the basis spread, which is identical to the expectation of the forward rate index $L(t)$.

Though the FRA's present value could be represented as the sum of two discounted cash flows, there is actually only one cash flow at the start of the interest rate period T equal to the difference of the present values of the two other cash flows (cash settlement). For this, the virtual cash flow at time T' needs to be discounted to time T by means of the reference rate index $L(t)$.

Since $L(t)$ differs by the basis spread $s(t)$ from the risk free interest rate, we have a slight inconsistency here, because on the other hand the risk free discount factor $B(t, T)$ is used for discounting from T to t to calculate the present value. This has a historical as well as a practical reason. Historically, before 2007, basis spreads have been so small that basis spread effects haven't had a big impact and have often been neglected. The practical reason is that the risk free interest rate is no directly observable mark rate index and could

not be determined exactly without any doubt, other than the reference rate index, which is daily quoted publicly. The impact on the fair strike price K of the FRA is small, though, and will be dominated by other effects, e.g. counterparty credit risk and collateral management.

15.3 Coupon Bonds

15.3.1 Cash Flows and Present Value

Coupon bond, often simply referred to as a *bond*, is a security with a *coupon*, which specifies a fixed interest rate K. In general, the coupon is given by an annualized interest rate (i.e. the interest rate for an annual term, even if the actual coupon payment frequency is, for example, quarterly). The cash flows arising from the interest payments C_Z of a coupon bond equals the interest the principal N of the bond, compounded at the rate K, earns over the interest period under consideration (for example, 6 months). In addition, a cash flow C_T for the redemption of the principal occurs at maturity.

$$C_Z(t_i) = K\tau(t_{i-1}, t_i)N \quad \forall i = 1, \ldots, n$$
$$C_T(t_n) = N \ . \tag{15.5}$$

Here, $\tau(t_1, t_2)$ is the time difference in years between t_1 and t_2 calculated with a day count convention consistent with the fixed interest rate K. The present value given as the sum of all discounted future cash flows is

$$V(t) = N \sum_{i=m+1}^{n} B(t, t_i)K\tau(t_{i-1}, t_i) + NB(t, t_n)$$
$$= NKA(t, t_0, t_n) + NB(t, t_n) \tag{15.6}$$

with

$$A(t, t_0, t_n) = \sum_{i=1}^{n} B(t, t_i)\tau(t_{i-1}, t_i)\mathbf{1}_{t<t_i} \ . \tag{15.7}$$

Here, t_0 is the start date of the bond, which could be in the past, i.e. $t_0 < t$. The factor $\mathbf{1}_{t<t_i}$, which equals 1 if $t < t_i$ and 0 otherwise, ensures that only future cash flows will be taken into account of the present value. The term

$A(t, t_0, t_n)$ is also called *Annuity*. Sometimes, the same expression is used for the bond itself or more generally for all products with regular (often annually) fixed cash flow payments. Usually, it is clear from the context what is meant by annuity.

Actually, Eq. 15.6 is only a simplification in the sense that there is no differentiation between coupon dates and payment dates. According to the business day rules applied, both types of dates may differ. We neglected this difference to simplify the notation. Most often, it is clear which dates are actually coupon dates (e.g., the arguments of $\tau(\cdot, \cdot)$) and where payment dates are needed (e.g., as second argument of $B(t, \cdot)$).

It is a frequent requirement that all coupon periods (measured in the specified day count rule) shall have exactly the same length, such that all coupon amounts are equal. The day count rule ICMA was especially designed to guarantee this feature for "natural" interest periods. For example, consider a bond with coupon dates on the 20th of March, June, September and December, as long as it does not fall on a bank holiday, otherwise on the next following business day. Using the day count rule ICMA will ensure that the length of such a coupon period will always be exactly quarter of a year, i.e. 0.25 years, long, independent of the actual number of days or if start and/or end date is a bank holiday. The ICMA rule is also called EUR-bond-method, since it is often applied to EUR denominated bonds. Fix a fixed interest rate K for each period, the Coupon $K \tau N$ is for all coupons equal to a constant c. In this case, the present value simplifies to

$$V(t) = c \sum_{i=m+1}^{n} B(t, t_i) + N B(t, t_n) \quad \text{with} \quad c = K \tau N \quad \forall i . \qquad (15.8)$$

15.3.2 Yield to Maturity

By definition of the yield to maturity, discounting in Eq. 15.6 at the *yield to maturity* $\overline{R}$ (which is constant until maturity) instead of the current zero bond rates must result in the same present value. Assuming that the coupon periods are equally long, results in YTM discount factors which are all identical over entire interest periods (t_i, t_{i+1}). We can therefore set $B_A(\overline{R}, t_i, t_{i+1}) = b \; \forall i$.[1]

---

[1]Here, $B_A(R, t_i, t_{i+1})$ is the discount factor based on a discrete, annual compounding interest rate R.

We then re-write the YTM discount factors in the following way

$$B_A(\overline{R}, t, t_i) = \underbrace{B_A(\overline{R}, t, t_{m+1})}_{\substack{\text{partial} \\ \text{coupon period}}} \underbrace{B_A(\overline{R}, t_{m+1}, t_{m+2})}_{b} B_A(\overline{R}, t_{m+2}, t_i)$$

$$= B_A(\overline{R}, t, t_{m+1}) b^2 B_A(\overline{R}, t_{m+3}, t_i)$$

$$= B_A(\overline{R}, t, t_{m+1}) b^{i-m-1} \underbrace{B_A(\overline{R}, t_{m+(i-m)}, t_i)}_{=1} .$$

Replacing the spot rate discount factors in Eq. 15.8 with these YTM discount factors gives

$$B_A(\overline{R}, t, t_n) = B_A(\overline{R}, t, t_{m+1}) b^{n-(m+1)}$$

$$\sum_{i=m+1}^{n} B_A(\overline{R}, t, t_i) = \underbrace{B_A(\overline{R}, t, t_{m+1})}_{\substack{\text{partial} \\ \text{coupon period}}} \sum_{i=m+1}^{n} \underbrace{b^{i-(m+1)}}_{\substack{\text{entire} \\ \text{coupon period}}}$$

$$= B_A(\overline{R}, t, t_{m+1}) b^{n-(m+1)} \frac{1 - b^{m-n}}{1 - b^{-1}} , \qquad (15.9)$$

where the formula for the sum of a *geometric series* has been used in the last step. Geometric series are introduced in all elementary mathematics texts and collections of mathematical formulas.[2] We will often use such series in what follows. The quintessential property of geometric series is that their finite sums are given by the following equation:

$$\sum_{k=1}^{n} c^{k-1} = \frac{c^n - 1}{c - 1} \quad \text{for any constant } c . \qquad (15.10)$$

---

[2]A sequence x_k, having the property that the quotient of two elements, one immediately following the other, is constant, $x_{k+1}/x_k = c$ for all k, is called a *geometric sequence*. Thus, for each of the elements of a geometric sequence, $x_k = x_1 c^{k-1}$. The sum of the elements of a geometric sequence is called a *geometric series*. The following formula holds for series of this type:

$$\sum_{k=1}^{n} x_k = x_1 \sum_{k=1}^{n} c^{k-1} = x_1 \frac{c^n-1}{c-1}.$$

Application of this equation in the above sum over the powers of the constant b leads[3] to Eq. 15.9. The present value of a bond with coupon periods of equal length expressed in terms of its yield to maturity is thus

$$V(t) = N B_A(\overline{R}, t, t_{m+1}) b^{n-(m+1)} \left[1 + c \, \frac{1 - b^{m-n}}{1 - b^{-1}} \right]$$

$$\text{with} \quad b = B_A(\overline{R}, t_m, t_{m+1}) \,, \quad c = K\tau \,. \tag{15.11}$$

This is a closed form expression for the present value of a bond as a function of its yield to maturity. It contains the discount factor for the partial interest period from t (today) until the next coupon payment date t_{m+1} and two parameters b and c which are determined through the discount factors with respect to the coupon and the yield to maturity for the next complete coupon period (from t_m until t_{m+1}).

In Table 15.2, fifteen different bonds with a 6% coupon, annual coupon payments and remaining lifetimes between 1 and 15 years were priced using Eq. 15.8 with the spot rates. In addition, the YTM for each bond and the par rate for each term to maturity is given. The par rates have been calculated in advance with Eq. 15.12 (see next section). In the case of the bond with time to maturity of only one year, zero rate, YTM and par rate are identical. For all other terms to maturity, these three values differ by up to 30 basis points.

15.3.3 Par Rates

Bonds are fixed-rate instruments. Therefore, the at-par condition 5.5 can be used in calculating the par rate $\bar{K}$. Since plain vanilla bonds before maturity pay only interest (amortization payments being made only upon maturity) the "residual debt" during the entire lifetime is equal to the principal N plus the accrued interest since the last coupon payment. Inserting the cash flows in

---

[3] Set $k = i - m$. The geometric series is then:

$$\sum_{i=m+1}^{i=n} b^{i-(m+1)} = \sum_{k+m=m+1}^{k+m=n} b^{k+m-(m+1)} = \underbrace{\sum_{k=1}^{n-m} b^{k-1}}_{\text{Geom. Series}} = \frac{b^{n-m} - 1}{b - 1}$$

$$= b^{n-m-1} \frac{b(1 - b^{m-n})}{b - 1} = b^{n-m-1} \frac{b(1 - b^{m-n})}{b(1 - b^{-1})} \,.$$

Table 15.2 Par rate, YTM and present value for 15 coupon bonds with a 6% annual coupon and time to maturity from 1 to 15 years. For comparison, the zero rate is given in discrete annual compounding

Term [years]	Zero rate	Zero bond	Par rate	YTM	Present value
1	2.20%	0.97847	2.200%	2.200%	103.72%
2	2.50%	0.9518	2.496%	2.492%	106.76%
3	2.90%	0.91781	2.886%	2.872%	108.87%
4	3.30%	0.87821	2.268%	3.245%	110.18%
5	3.70%	0.83389	3.643%	3.612%	110.75%
6	4.00%	0.79031	3.919%	3.884%	111.13%
7	4.30%	0.74475	3.188%	4.149%	111.05%
8	4.50%	0.70319	4.366%	4.328%	111.11%
9	4.70%	0.66142	4.539%	4.500%	110.90%
10	4.90%	0.61979	4.706%	4.668%	110.46%
11	5.00%	0.58468	4.793%	4.757%	110.45%
12	5.10%	0.55051	4.878%	4.844%	110.34%
13	5.20%	0.51736	4.959%	4.926%	110.13%
14	5.25%	0.48853	5.004%	4.974%	110.18%
15	5.30%	0.46087	5.047%	5.018%	110.18%

Eq. 15.5 into the general at-par condition 5.5, the principals N cancel and the at-par condition for bonds becomes

$$1 + \bar{K}\tau(t_m, t) = \sum_{i=m+1}^{n} B(t, t_i)\bar{K}\tau(t_{i-1}, t_i) + B(t, T)$$

$$= \bar{K}A(t, t_0, t_n) + B(t, T) \tag{15.12}$$

with $t_m \leq t < t_{m+1} < t_{m+2} < \cdots < t_n = T$, $n \geq m + 1$.

The interest accrued until time t for the current interest rate period normalized to one unit of currency appears on the left hand side. On the right, we have the present value at time t of a bond with coupon $\bar{K}$ and a principal of 1 monetary unit. Solving for the par rate yields

$$\bar{K} = \frac{1 - B(t, T)}{A(t, t_0, t_n) - \tau(t_m, t)} \ . \tag{15.13}$$

15.3.4 Sensitivities

The sensitivities with respect to the yield to maturity, namely the modified duration and the convexity, are calculated using Eq. 15.6:

$$
D_{\text{mod}} = -\frac{1}{V}\frac{\partial V}{\partial \overline{R}} = -\frac{NK}{V}\sum_{i=m+1}^{n}\tau(t_{i-1}, t_i)\frac{\partial B_A(\overline{R}, t, t_i)}{\partial \overline{R}} - \frac{N}{V}\frac{\partial B_A(\overline{R}, t, t_n)}{\partial \overline{R}}
$$

$$(15.14)$$

$$
\text{Convexity} = \frac{1}{V}\frac{\partial^2 V}{\partial \overline{R}^2} = \frac{NK}{V}\sum_{i=m+1}^{n}\tau(t_{i-1}, t_i)\frac{\partial^2 B_A(\overline{R}, t, t_i)}{\partial \overline{R}^2} + \frac{N}{V}\frac{\partial^2 B_A(\overline{R}, t, t_n)}{\partial \overline{R}^2} \ .
$$

If all coupon periods are of equal length, the closed form expression in Eq. 15.11 can be differentiated explicitly with respect to the yield to maturity to obtain closed form expressions for the modified duration and the convexity of the bond. These expressions are not listed here as they are quite lengthy and illustrate no new concepts.

The sensitivities with respect to the spot rate, i.e., the key rate durations are

$$
D_{\text{key}}(t, t_i) = \frac{NK}{V}\tau(t_{i-1}, t_i)\frac{\partial B_R(t, t_i)}{\partial R} \quad \forall \ i = m+1, \ \dots \ , n-1
$$

$$
D_{\text{key}}(t, t_n) = \frac{NK}{V}\tau(t_{n-1}, t_n)\frac{\partial B_R(t, t_n)}{\partial R} \ .
$$

$$(15.15)$$

The expressions for the sensitivities in Eqs. 15.14 and 15.15 look more complex than they actually are. The derivatives can be interpreted as the modified duration, key rate duration, and convexity of zero bonds, each multiplied by the present value of the respective zero bond (see Eq. 5.15). For each zero bond, we must use the interest rate R with respect to which the derivative was taken, explicitly: the YTM must be used for the modified duration and convexity, the spot rate of the appropriate interest period for the key rate duration. The present value and sensitivities of the zero bond for all of the usual compounding conventions can be taken from Table 15.1. The calculation of the sensitivities of the fifteen coupon bonds in Table 15.2 is demonstrated explicitly in Table 15.3.

Table 15.3 The sensitivities of the coupon bonds from Table 15.2. Annualized compounding was used throughout

Term [years]	Modified duration [years]	Key rate duration [years]	Convexity [years]	Modified duration × PV	Convexity × PV
1	0.978	0.978	1.915	0.957	1.874
2	1.898	1.844	5.503	1.857	5.436
3	2.761	2.605	10.533	2.676	10.405
4	3.572	3.272	16.810	3.402	16.470
5	4.332	3.848	24.142	4.023	23.289
6	5.050	4.349	32.420	4.563	30.737
7	5.721	4.771	41.441	5.004	38.421
8	6.356	5.136	51.183	5.390	46.482
9	6.951	5.434	61.430	5.694	54.471
10	7.504	5.670	72.040	5.919	62.187
11	8.035	5.878	83.185	6.137	70.279
12	8.532	6.038	94.584	6.299	78.077
13	8.996	6.154	106.137	6.408	85.472
14	9.443	6.252	118.032	6.513	93.046
15	9.863	6.316	130.016	6.581	100.234

15.4 Forward Bonds

15.4.1 Present Value

Equation 15.6 for the present value of a bond also holds for a future time T. For the discount factors with respect to the future time T, the current forward rates must be used in place of the spot rates to ensure arbitrage freedom. Thus, the present value of forward bond differs from the present value of a regular bond only by replacing the start time t_0 by the future time T. With maturity $T' = t_n$ we have

$$V(T) = NKA(t, T, T') + B(t, T')N . \qquad (15.16)$$

The value of a forward bond with interest rate K differs only by the additional nominal payments from the fixed leg of a forward swap (Eq. 15.24).

15.4.2 Forward Par Rate and Forward Yield to Maturity

The calculation of forward yield to maturity and forward par rate is analog to the calculation of these figures for a regular bond with Eqs. 15.9 and 15.13, but without the terms for the current interest period. Thus, we get the equation

for the yield to maturity (assuming periods of equal length) by setting $m = 0$ in Eq. 15.9:

$$\sum_{i=1}^{n} B_A(\overline{R}, t, t_i) = B_A(\overline{R}, t, t_1) \sum_{i=1}^{n} b^{i-(1)}$$

$$= B_A(\overline{R}, t, t_1) b^{n-1)} \frac{1 - b^{-n}}{1 - b^{-1}} \, .$$

The forward par rate $\bar{K}$ follows in a similar way by replacing the zero bond and annuity by their forward zero bond and forward annuity equivalents in Eq. 15.13:

$$\bar{K} = \frac{1 - B(T, T'|t)}{A(T, T, T'|t)} = \frac{B(t, T) - B(t, T')}{A(t, T, T')} \, .$$

In the last expression, both nominator and denominator have been multiplied by $B(t, T)$ to emphasize the similarity with the forward par swap rate, which differs only by an additional basis spread term.

15.5 Interest Rate Futures

15.5.1 Futures on Zero Bonds

The holder of a zero bond receives no interest payments before maturity. This corresponds to an underlying "without dividends". The forward price as seen from today's date t of a zero bond at time T having a face value of one monetary unit and maturing at $T' > T$ is given by Eq. 6.1 as

$$S(t, T) = \underbrace{B^{-1}(t, T)}_{\text{compounding until } T} \underbrace{B(t, T')}_{\text{present value at time } t} = \frac{B(t, T) B(T, T'|t)}{B(t, T)} = B(T, T'|t) \, .$$

$$(15.17)$$

The forward price as seen from time t of a zero bond at T is thus the price which the zero bond would have at T if the forward rate as seen at t were the actual spot rate at time T. With Eq. 6.6 we see that the present value of a future with delivery price K on such a zero bond is this forward price less the delivery price.

$$F_S^{\text{zero bond}}(t, T, T', K) = B(T, T'|t) - K \ .$$

15.5.2 Futures on Coupon Bonds

We will determine the forward price of a coupon bond from the cash & carry arbitrage which led to Eq. 6.1. Regardless of the underlying instrument, only two questions were of importance for the cash & carry arbitrage:

- What amounts must actually be paid in order to purchase the underlying today and at an agreed upon future date?
- Which cash flows arise for the holder of the underlying up to maturity which do *not* arise for the holder of the future?

For futures on coupon bonds, often just called *bond futures* for short, we therefore have to apply the cash & carry arbitrage for forward contracts on underlyings with dividend payments, where the dividend payments of course correspond to the coupon payments of the underlying bond. But *only* those coupon payments which arise during the lifetime of the forward contract are to be taken into consideration.[4]

Denoting the coupon payments at t_i by $C(t_i)$, the following holds for the dirty forward price of the bond

$$S(t, T) = \frac{S(t) - D(t, T)}{B(t, T)} \quad \text{with} \quad D(t, T) = \sum_{t < t_i \leq T} C(t_i) B(t, t_i) \ .$$
$$(15.18)$$

In this equation

- $D(t, T)$ is the present value of all coupon payments to be made up to the maturity of the future;

---

[4]Bonds *cannot* be interpreted as an underlying with a dividend *yield q* when pricing options and forward contracts on bonds. The assumption of a dividend yield assumes payments relative to the underlying price, which is not a useful approximation for bonds. In addition, the accrued interest on the underlying is exchanged directly between the counterparties through the dirty price. In cash & carry arbitrage, only the "extra" cash flows which are not exchanged by the counterparties of the forward contract must be taken into consideration. These are just the coupon *payments* during the lifetime of the forward contract. These are received by the holder of the underlying, but not by the holder of the forward contract. Hence, the only consistent treatment of a bond is to interpret it as an underlying with dividend payments during the lifetime of the forward contract and to perform each calculation (for the spot price, the forward price and option strikes) with the dirty price (= clean price plus accrued interest).

- $S(t, T)$ is the dirty price (including accrued interest) to be paid for the bond at time T as seen from time t;
- $S(t)$ is the dirty price (including accrued interest) at which the bond is available for purchase at time t.

According to Eq. 6.6, the value of a future with a delivery price K on such a bond is this dirty forward price less the dirty delivery price. In the market, however, *clean* prices are quoted. Therefore the interest accrued up to time T must be added to the quoted (clean) delivery price. For forward and option contracts, all calculations must generally be made on the basis of dirty prices, i.e., the accrued interest should always be taken into consideration.

The bond futures traded on exchanges do not normally refer to real bonds as their underlying but rather to *synthetic bonds* which are defined by their coupon and time to maturity. Usually there are several real bonds which are quite similar to such a synthetic bond and which are therefore admissible as an underlying for the future under consideration. A future very popular in Germany, the *Euro Bund Future* (previously referred to as *Bund Future*) has as its underlying a synthetic German government bond with 10 years to maturity and a coupon of 6%, the underlying for the *Euro Bobl Future* (previously referred to as *Bobl Future*) is a synthetic bond with a 6% coupon and 5 years to maturity and the underlying of the *Euro Schatz Future* (previously referred to as *Schatz Future*) is a synthetic bond with a 6% coupon and 2 years to maturity. The face value of the synthetic underlying, called the *contract size*, is 100,000 euros for all these futures. The admissible real underlyings that the holder of a short future position is allowed to deliver at the maturity of such futures are

- for *Euro Bund Futures:* German federal bonds (*Bundesanleihen*) with a remaining term upon delivery of $8\frac{1}{2}$ to $10\frac{1}{2}$ years.
- for *Euro Bobl Futures:* German federal bonds (*Bundesanleihen*) and German federal debt obligations (*Bundesobligationen*) with a remaining term upon delivery of $4\frac{1}{2}$ to $5\frac{1}{2}$ years.
- for *Euro Schatz Futures:* German federal treasury notes (*Bundesschatzanweisungen*), German federal debt obligations (*Bundesobligationen*) and German federal bonds (*Bundesanleihen*), all with a remaining term upon delivery of $1\frac{3}{4}$ to $2\frac{1}{4}$ years.

In order to enable the investor to compare the prices of all the different instruments allowed to be delivered at maturity of a future, the futures exchange *EUREX* calculates a *conversion factor* for each real deliverable bond which converts the clean price of the synthetic bond into the clean price of

the deliverable bond. The EUREX calculates this conversion factor P and the accrued interest Z as follows:

$$P = \frac{1}{(1 + K)^{m/12}} \left[\frac{K_L}{K} (1 + K) + \frac{1}{(1 + K)^n} \left(1 - \frac{K_L}{K} \right) \right] - K_L (1 - m/12)$$

$$Z = \frac{i}{360} N K_L$$

where

> K_L = coupon of the deliverable bond
> K = coupon of the synthetic bond = $0, 06$
> n = full years until the maturity of the deliverable bond
> m = full months until the next coupon payment of the deliverable bond
> i = interest days since the last coupon payment of the deliverable bond
> N = contract size = 100,000 euros

The same conversion factor is used for Bund, Bobl and Schatz futures (as well as for options on these instruments). These futures differ only in the lifetime of their underlying synthetic bond.

At the maturity of the future, the holder of the long position in this future must purchase the bond delivered by the holder of the short position at the future price $S(T, T) = S(T)$, multiplied by a conversion factor P, in addition to the accrued interest Z which is always paid upon purchase of a bond.[5] For a contract size N the total amount (denoted by g) payable by the holder of the long future is

$$g = N \, S(T) \, P + Z$$

The holder of the short position delivers a deliverable bond in exchange for this amount g. This bond costs him or her the face value N multiplied by the quoted (clean) price $S_L(T)$ of the deliverable bond at time T plus the accrued interest. Subtracting from these costs the amount received from the holder of

---

[5]The actual price $S(T)$ and not the originally agreed upon delivery price $S(t_0, T)$—where t_0 denotes the date the future contract was entered into—must be paid since the differences between the actual forward price of the underlying and the delivery price $S(t_0, T)$ have already been settled on a daily basis because of the variation margin system.

the long future, the net costs are

$$N\,S_L(T) + Z - g = N\,[S_L(T) - P\,S(T)]$$

These costs vary from bond to bond. The holder of the short position will therefore choose from all the deliverable bonds exactly that bond, which minimizes the holder's costs. In other words, the bond which minimizes the value of $S_L(T) - P\,S(T)$ will be chosen. This bond is called the *cheapest to deliver* or *CTD* for short. In fact, futures (and options as well) on synthetic bonds are priced as if they were written directly on the cheapest to deliver bond instead of on a synthetic bond. This means that the future's price is that given by Eq. 15.18 using the forward price of the CTD bond (quoted as the clean price, i.e., less the interest accrued up to T and divided by the conversion factor). Without this procedure, Eq. 15.18 could not be applied in practice since the coupon dates of the synthetic bond are not defined and thus the present value of the coupon payments of synthetic bonds cannot be determined.

The CTD bond is generally known on the market. From this bond, the forward price of the synthetic underlying (which is the one and only thing quoted in such futures markets) can be calculated in five steps

- Determine the dirty price at time t of the CTD from its quoted (clean) price plus the accrued interest.
- Determine the present value of the CTD's coupon payments due between times t and T.
- Calculate the dirty price of the CTD for the date T using Eq. 15.18.
- Determine the (clean) price of the CTD for the date T from the dirty price less the interest accrued up to T.
- Divide this price by the conversion factor for the CTD to obtain the (clean) forward price of the synthetic bond.

15.6 Floaters

A *floater* is a security that pays periodically interest based on a variable interest rate index and the nominal at maturity. The variable interest payments refer to a *reference rate* observable on the market (for example, the 6-month *EURIBOR*). The reference rate determines the interest rate that is applied to the nominal N of the floater to calculate the interest compounded over the interest rate period (e.g. 6 months). The reference rate is determined at the

fixing date or brief *fixing*, which is typically at or before the beginning of each interest rate period. The reference rate fixed at time t relates to an interest rate period starting at $t + v$, where v is the number of spot days (for 6-months EURIBOR, this is 2 business days). Therefore, the fixing date is usually two business days before the interest period start date. The interest coupon is generally paid at the end of the interest rate period at the next possible business day. In the following, we assume that the interest payment day falls together with the interest period end date, if not explicitly stated otherwise. $t_m \leq t$ denotes the last interest payment date (i.e. end date of the previous interest rate period), where t is the current date. Then, t_{m+1} is the payment date of the current interest rate period. The maturity $T = t_n$ equals the last interest payment date of the floater. t_0 is the start date of the first interest rate period. Therefore, we have:

$$\underbrace{t_1 < t_2 < \cdots t_m}_{\text{past}} \leq t < \underbrace{t_{m+1} < t_{m+2} < \cdots t_n = T}_{\text{future}} .$$

The fixing date at which the interest rate for the payment date t_m is fixed will be denoted by $F(t_m)$. At time t, only the next cash flow (if $F(t_{m+1}) \leq t_m \leq t < t_{m+1}$) and the second next cash flow (if $F(t_{m+2}) \leq t \leq t_{m+1} < t_{m+2}$) may be deterministic already. All further cash flows will have to be determined in the future.

15.6.1   Cash Flows and Present Value

(Uncollateralized) forward rate indexes are traded with an add-on on top of the (default) risk free interest rate, the so-called *basis spread* or just *basis*,[6] taking credit and liquidity risks into account. The basis depends on the start and length of the forward rate reference's respective interest rate period. For a regular floater, all interest rate periods have the same length. Therefore, it is here sufficient to consider only the start date dependency of the basis. The basis of the forward rate fixed at $F(t_m)$ is denoted in the following as s_m. Using arbitrage free arguments, the (default) risk free forward rates can be identified with those calculated by means of Eq. 2.7. We can now write down the cash

---

[6]Small differences between similar, but not identical entities are frequently called basis.

flows from interest payments C_Z and nominal repayment C_T at maturity

$$C_Z^{\text{fix}}(t_{m+1}) = R^{\text{fix}}\tau(t_m, t_{m+1})N$$

$$C_Z(t_i) = \left[B^{-1}(t_{i-1}, t_i|t) - 1 + s_i\tau(t_{i-1}, t_i)\right]N$$

$$= \left[\frac{B(t, t_{i-1})}{B(t, t_i)} - 1 + s_i\tau(t_{i-1}, t_i)\right]N \quad \forall\, i = m+2, \ldots, n$$

$$C_T(t_n) = N \,. \tag{15.19}$$

Here, R^{fix} is the fixed rate for the current interest rate period (for simplicity, we have assumed that $t < F(t_{m+1})$, otherwise, the interest rate for the next period would also be fixed already and we would have two terms with a deterministic interest cash flow). $\tau(t_1, t_2)$ is the year fraction between start t_1 and end t_2 of an interest period, calculated by application of a proper day count convention. The present value of the sum of all future cash flows is then

$$V(t) = C_Z^{\text{fix}}(t_{m+1})B(t, t_{m+1}) \tag{15.20}$$

$$+ N \sum_{i=m+2}^{n} B(t, t_i)\left[\frac{B(t, t_{i-1})}{B(t, t_i)} - 1 + s_i\tau(t_{i-1}, t_i)\right] + NB(t, t_n)$$

$$= C_Z^{\text{fix}}(t_{m+1})B(t, t_{m+1})$$

$$+ N \sum_{i=m+2}^{n} \left[B(t, t_{i-1}) - B(t, t_i) + B(t, t_i)s_i\tau(t_{i-1}, t_i)\right] + NB(t, t_n) \,.$$

Since the discount factor $B(t, t_i)$ appearing with a negative sign in the above formula is equal to the discount factor with a positive sign $B(t, t_{i-1},)$ in the next term in the sum, i.e., for $i \rightarrow i+1$, the series is telescoping, with all terms canceling except the discount factors with the smallest and the largest index. Thus, the above equation reduces to

$$V(t) = C_Z^{\text{fix}}(t_{m+1})\,B(t, t_{m+1}) + N\left[B(t, t_{m+1}) - B(t, t_n)\right]$$

$$+ N \sum_{i=m+2}^{n} B(t, t_i)s_i\tau(t_{i-1}, t_i) + NB(t, t_n) \,.$$

Together with Eq. 15.19 for the present value of the fixed cash flow this can be further reduced to

$$V(t) = N B(t, t_{m+1}) \left[1 + R^{\text{fix}} \tau(t_m, t_{m+1}) \right] + N \sum_{i=m+2}^{n} B(t, t_i) s_i \tau(t_{i-1}, t_i) .$$

(15.21)

This is quite a remarkable result! Neglecting for a moment the basis spread term, the present value of a floater is given by the sum of the principal N and the (already) fixed interest cash flow $C_Z^{\text{fix}}(t_{m+1})$, discounted from the next interest payment date t_{m+1} back to t. This corresponds to a case in which the fixed interest cash flow together with the entire principal is due on the next payment date $t_m + 1$, after which no further payments occur. However, since the financial crisis 2008, basis spreads have widened and can no longer be neglected. From this point of view, floater bear primarily a basis spread risk, i.e. the risk that the basis spread may change.

> The interest rate risk of a floater is for pricing and risk management equivalent with a zero bond with maturity at the next interest payment date and a principal equal to the sum of the floater's nominal and the already fixed interest payment. The basis risk, however, could be significant.

The present value of a floater is especially simple, if the basis risk is neglected and the fixing dates is identical with the period start date $F(t_{m+1}) = t_m$. Evaluating the floater at the cash flow date t_m yields the floater's nominal as present value, since the floater equals according to Eq. 15.21 exactly its nominal

$$R^{\text{fix}} = \frac{1}{\tau(t_m, t_{m+1})} \left(B^{-1}(t_m, t_{m+1}) - 1 \right) \quad \Rightarrow \quad V(t = t_m) = N .$$

This result is independent on compounding method, length of interest period, or anything else.

Floaters are typically traded with a *spread*. This means that the cash flows are not fixed precisely at a reference rate rather, but a spread K is added to or subtracted from to the rate (for example, 3-month LIBOR plus 10 basis points). With such a spread, the risk of an issuer default could be compensated in such a way, that the floater could be quoted at 100 at issue date. Here, we will for the moment continue to neglect the default risk effects on the valuation.

Such a (long) floater generates the same cash flows as a portfolio consisting of

- a floater without a spread, sometimes called *LIBOR flat* (long),
- a coupon bond (long) with the basis spread as a coupon and having the same principal, coupon dates and maturity as the floater,[7]
- a coupon bond (long) with the spread K as a coupon and having the same principal, coupon dates and maturity as the floaters,
- and a zero bond (short) whose principal and maturity is the same as that of the floater.

Including the short zero bond is necessary to compensate for the final payment at maturity of the coupon bond. A floater with a spread can thus be priced by means of the *stripping* procedure described above.

15.6.2   Yield to Maturity, Par Rate and Sensitivities

A floater does not earn at a fixed rate K. As indicated by its name, it is a floating rate instrument. A "par rate" as described in Sect. 5.3 therefore makes no sense in this context. The YTM, however, can be defined for floating rate instruments. Since a floater is equivalent to a zero bond in terms of its interest rate dependence (present value and risk with neglecting the basis spread), the YTM and sensitivities, i.e., the modified duration, convexity and the key rate duration are calculated exactly as for the associated zero bond. Taking the basis spreads and spreads into account, these figures could be calculated similarly to coupon bonds. Alternatively, the forward rates could be frozen at levels calculated at valuation date, making all cash flows effectively deterministic. This allows to perform the same calculations as for coupon bonds. However, this would change the meaning of this sensitivities, which could no longer be interpreted in the usual way.

15.7 Swaps

Swaps are contracts used to exchange interest payments. In their simplest form, a plain vanilla swap, fixed cash flows are exchanged for floating cash flows.

---

[7] Here, the basis spread was assumed to be deterministic and time dependent. Though in reality, the basis will change over time, which could be simulated in a more complex model.
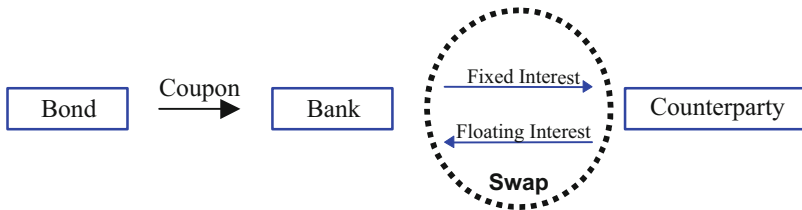
Fig. 15.1 Reducing the interest rate risk of a bond by using a plain vanilla interest rate swap

A bank, which holds a coupon bond, for example, but which wishes to reduce the interest rate risk of this fixed rate instrument can initiate a Swap according to which the counterparty of the bank receives fixed interest payments from the bank while the bank in return receives floating interest payments from the counterparty (see Fig. 15.1). If the swap's negotiated fixed rate is equal to the coupon of the bond then the fixed payments from the bond are transferred entirely to the counterparty and the bank's net return is the floating interest received from the counterparty.

The meaning of the terms "long" and "short" are a bit ambiguous for swaps. We will therefore refer to swaps more precisely as payer or receiver swaps. In this book, the usual market convention will be adopted: "long" means "long payer", i.e., one *pays* the fixed rates. This is the exact opposite as was the case for bonds: one is long in a bond when one *receives* its fixed rate (i.e., coupon) cash flows. The long/short conventions for swaps extend to options on swaps, called *swaptions* (see Sect. 18.6.4). All conventions for swaps and swaptions are summarized in Table 15.4.

15.7.1 Cash Flows and Present Value

The cash flow stream of a *plain vanilla interest rate swap* (long payer) is equivalent (except for default risks) to a portfolio consisting of a long floater and a short bond paying a coupon at the same rate as the fixed side of the swap. If the floater has a spread over the reference rate, it can be added to the coupon of the fixed side for the purpose of pricing the swap if the payment dates of the fixed and floating sides of the swap coincide. Otherwise, the floating side can, as mentioned in Sect. 15.6, be interpreted as a portfolio consisting of a floater, a bond and a zero bond. The present value of a swap can be computed directly from the present values of a floater and a bond. For example, for the

Table 15.4 Long/short conventions for swaps and swaptions. The parts in brackets are often left out. Expressions used most often are printed in bold

Short name	Full name	Corresponds to	Meaning
Payer swap	Long (payer) swap	Short receiver swap	Pay fixed and receive variable interest
Receiver swap	Short (payer) swap	Long receiver swap	Receive fixed and pay variable interest
Long call	(Long) **call on payer swap**	Long put on receiver swap	Right to pay fixed and receive variable interest
Long put	(Long) **put on payer swap**	(Long) call on receiver swap	Right to receive fixed and pay variable interest
Short call	**Short call on payer swap**	Short put on receiver swap	Obligation to receive fixed and pay variable interest
Short put	**Short put on payer swap**	Short call on receiver swap	Obligation to pay fixed and receive variable interest

case of a long payer swap

$$V(t) = V^{\text{Floater}} - V^{\text{Bond}}$$

$$= B(t, t_{m_f+1}) \left[N + C_Z^{\text{fix}}(t_{m_f+1}) \right] + N \sum_{i=m_f+2}^{n_f} B(t, t_i) s_i \tau(t_{i-1}, t_i)$$

$$- N \sum_{i=m_b+1}^{n_b} B(t, t_i) K \tau(t_{i-1}, t_i) - N B(t, t_{n_b})$$

$$= N \left[B(t, t_{m_f+1}) - B(t, t_n) \right] - N K A(t, t_0, t_n) \qquad (15.22)$$

$$+ B(t, t_{m_f+1}) C_Z^{\text{fix}}(t_{m_f+1}) + N \sum_{i=m_f+2}^{n_f} B(t, t_i) s_i \tau(t_{i-1}, t_i) .$$

Here, we have introduced the sub-indexes f (for floater) and b (for bond) to differentiate between coupon period indexes related to the variable and fixed leg of the swap. This is required, since interest rate periods for both legs are generally different. Often, the interest periods for the fixed leg are twice as long as for the variable leg, such that only every second coupon payment date of the variable leg falls together with one of the fixed rate leg. Start date t_0 and maturity $t_n = t_{n_f} = t_{n_b}$ are identical for both legs of a regular swap,

though. The annuity $A(t, t_0, t_n)$ always relates to the fixed side of the swap, the additional sub-index b may therefore be dropped. In the case of identical period lengths for both legs (which is rarely the case), the expression further simplifies to:

$$V(t) = N\left[B(t, t_{m+1}) - B(t, t_n)\right] - N \sum_{i=m+2}^{n_f} B(t, t_i)(K - s_i)\tau(t_{i-1}, t_i)$$
$$+ B(t, t_{m+1})\left(C_Z^{\text{fix}}(t_{m+1}) - K\right).$$

In Table 15.5, we consider a swap with a notional of 250,000 euros in which the 1-year floating rate is exchanged for a fixed interest rate of 6% over a period of 15 years. We do so by constructing a cash flow table of the swap, i.e. the list of all expected cash flows as seen from today are calculated for both, the fixed and floating sides (called *legs*) and the resulting net cash flow is determined. The present value is the sum of all net cash flows discounted at the current spot rates corresponding to the respective payment dates. The cash flows on the fixed side have already been established through the fixed interest rate and the principal agreed upon in the swap contract. The expected cash flows on the floating side are obtained from the forward rates for 1-year deposits incl. the basis spread. Calculation of the forward rates is demonstrated in Fig. 2.2. In addition, both swap legs are calculated according to Eq. 15.22 as bond resp. floater. The difference also equals the swap present value, since the additional nominal repayments cancel each other out.

15.7.2   Par Swap Rate and Yield to Maturity

The *par swap rate* is the analogue to the bond par rate. The definition of par rates can be transferred to swaps by defining that a swap quotes at par, if its present value, Eq. 15.22, is equal to zero. Then, the interest rate of the fixed leg is equal to the par swap rate, if the swap priced with given rate curves for calculation of forward rates and discount factors is at par. With this definition,

Table 15.5 Cash flow table of a swap. The forward rates for 1-year deposits in 1 to 14 years as calculated in Fig. 2.2, using discrete annual compounding. The swap present value is −2,441 EUR and its YTM is 1.48% per year

Term [years]	Spot rate	Start in years	Forward rate	Cash flow	PV based on spot rates	Cash flow	Present value	Cash flow	PV based on spot rates	PV based on YTM
1	2.20%	0	2.20%	15,000	14,677	5,500	5,382	−9,500	−9,295	−9,361
2	2.50%	1	2.80%	15,000	14,277	7,002	6,665	−7,998	−7,612	−7,766
3	2.90%	2	3.70%	15,000	13,767	9,262	8,501	−5,738	−5,267	−5,490
4	3.30%	3	4.51%	15,000	13,173	11,273	9,900	−3,727	−3,273	−3,513
5	3.70%	4	5.32%	15,000	12,508	13,289	11,081	−1,711	−1,427	−1,590
6	4.00%	5	5.51%	15,000	11,855	13,783	10,893	−1,217	−962	−1,114
7	4.30%	6	6.12%	15,000	11,171	15,296	11,391	296	220	267
8	4.50%	7	5.91%	15,000	11,548	14,777	10,391	−223	−157	−198
9	4.70%	8	6.31%	15,000	9,921	15,785	10,440	785	519	687
10	4.90%	9	6.72%	15,000	9,297	16,793	10,408	1,793	1,111	1,548
11	5.00%	10	6.01%	15,000	8,770	15,013	8,778	13	8	11
12	5.10%	11	6.21%	15,000	8,258	15,516	8,542	516	284	432
13	5.20%	12	6.41%	15,000	7,760	16,019	8,287	1,019	527	841
14	5.25%	13	5.90%	15,000	7,328	14,755	7,208	−245	−119	−199
15	5.30%	14	6.00%	15,000	6,913	15,006	6,916	6	3	5
Sum of present values					160,224		134,784		−25,441	−25,441
Amortization bond and floater				250,000	115,216	250,000	115,216	250,000		
					PV bond	275,441	PV floater		PV swap	

we get for the par swap rate K_S the expression:

$$K_S = \frac{1}{A(t, t_0, t_n)} \left[B(t, t_{m_f+1}) - B(t, t_n) + B(t, t_{m_f+1}) C_Z^{\text{fix}}(t_{m_f+1}) \right.$$

$$\left. + \sum_{i=m_f+2}^{n_f} B(t, t_i) s_i \tau(t_{i-1}, t_i) \right].$$

Setting the basis spread to zero, $s_i = 0$, the par swap rate is identical to the par rate (Eq. 15.12) of a bond equivalent to the fixed leg at the beginning of an interest period, i.e. for $t = t_m$, since the float leg equals the nominal:

> At the start of an interest period, the swap rate is equal to the par rate of the bond on the fixed side of the swap, if basis spreads are neglected.

If a new swap is traded, the fixed rate is set equal to the par swap rate, such that the swap present value equals zero, if we consider a default risk free swap (e.g. a swap that is collateralized or traded via a *Central Counter Party* or *CCP*).

In principal, the *yield to maturity* of a swap could be calculated in the same way as the YTM of a bond via its present value Eq. 15.22. According to the definition of YTM, that interest rate needs to be determined (e.g. by numerical iteration), which, entered into Eq. 15.22 for all spot rates, gives the current swap present value. This would be the YTM of the swap. Note that we have to deal with the complication that the interest rate curve is also needed to calculate the forward rate for the variable leg. For a meaningful definition, these forward rates would have to be calculated with the YTM, too, interpreting the basis spread s_i as a spread over YTM. Another possibility would be to freeze the forward rates in a first step and keep them constant later on.

Unfortunately, a YTM calculated in this way would not have the same meaning as the bond YTM. In case of a bond, the YTM is equal to the yield, the investor who holds the bond until maturity would earn (if the issuer won't default). In contrast, the yield of a swap cannot be known in advance because of its dependency on the stochastic forward rates. The yield could even turn negative, if the interest rate level changes. Therefore, YTM is of only limited help for analysis and valuation of swaps and somewhat misleading. Instead, other sensitivities should be used.

15.7.3 Sensitivities

Classic fixed income sensitivities like modified duration, convexity and key rate duration could formally be calculated for swaps, too, but not without losing most of their significance for risk management. E.g., the duration of a swap could become negative, which can be seen easily if the swap is decomposed into a floater and a bond. The duration of the floater equals more or less (depending on the spread) the end date of the current interest period and is in general much smaller than the duration of the bond. Therefore, the duration of a payer swap would become negative. This is not consistent with the interpretation of the duration as a weighted average lifetime. For the same reason, it is problematic to calculate the duration for a portfolio with long and short positions in fixed income securities, since the duration could also become negative. Often, this problem is circumvented by calculating the duration of the swap separately for each leg, i.e. the equivalent floater and bond, and averaging over both values. However, it is preferable to use other sensitivities described below, which could be calculated and interpreted for any financial product regardless of their structure. They are more useful do address the requirement to measure sensitivities with respect to a changing interest rate curve.

The most important figure or sensitivity for measuring interest rate sensitivities is the PV01 (*Present Value of a Basis Point*), which is the change of the present value of a financial instrument, if the interest rate curve is shifted up by 1 basis point. Alternatively the DV01 (*decrease of a basis point*) could be used, which is calculated by decreasing the rate curve by 1 basis point.

The PV01 could be calculated for the whole interest rate curve (parallel shift, i.e. shift of all zero rates at the same time) or for special time nodes only (*time-dependent PV01* or *bucketed PV01*). In the latter case, the resulting figures include detailed information on the exposure to certain spot interest rate terms, which is very important for steering the interest rate risk. Let $\tilde{B}(t, t_i)$ be the discount factor analog to $B(t, t_i)$, but based on a spot rate shifted by one basis point. To make this definition precise, we need to define the compounding method to be used to calculate the 1 basis point add-on. E.g., in continuous compounding we get the simple result

$$\tilde{B}(t, t_i) = e^{(t_i - t)/10000} B(t, t_i) .$$

The expressions for simple compounding

$$\tilde{B}(t, t_i) = \frac{1}{B^{-1}(t, t_i) + (t_i - t)/10000}$$

and discrete, annual compounding

$$\tilde{B}(t, t_i) = \left(\frac{1}{[B(t, t_i)]^{\frac{1}{t_i - t}} + (t_i - t)/10000} \right)^{t_i - t}$$

are somewhat more complex. To determine time-dependent PV01s, we also need to define the time points, at which we want to shift the interest rate curve. In addition, we need to define how to interpolate the discount factors in between the selected time points.

The discount factor curve can be defined as a set RC of the discount factors the swap depends upon, i.e.

$$RC = \left\{ B(t, t_i) | i \in m_f, \ldots, n \vee i \in m_b, \ldots, n \right\}$$

and for the parallel shifted curve

$$\widetilde{RC} = \left\{ \tilde{B}(t, t_i) | i \in m_f, \ldots, n \vee i \in m_b, \ldots, n \right\} .$$

Then, we get for the PV01

$$PV01 = V(t, \widetilde{RC}) - V(t, RC) .$$

Here, we have interpreted the present value of the swap $V(t) = V(t, RC)$ explicitly as a function of the current interest rate curve.

15.8 Forward Swaps

A *forward swap*, also called a *deferred swap* is a swap with a lifetime over a *future* time span from T to T', for example in three years ($T = 3$) over two years ($T' = 5$). The cash flow dates are denoted by $T_0, T_1, ..., T_n = T'$ where $T_0 = T$ equals the start of the first interest period. The times $T_1, ..., T_n$ are thus, as seen from T, still future cash flow dates. The relation between these

times can be summarized by

$$t < T_0 = T < T_1 < T_2 < \cdots < T_n = T' \tag{15.23}$$

15.8.1   Present Value

Equation 15.22 for the present value of a swap holds for the future time T as well. As far as the first fixing date T_{fix} of the variable leg lies in the future (which we will assume in the following, since otherwise, there would be no remarkable difference to a regular swap), we only have to replace the term for the already fixed rate by an equivalent term including the forward rate.

$$V(t, T) = N\left(B(t, T) - B(t, T')\right) - NKA(t, T, T') + N\sum_{i=1}^{n_f} B(t, T_i)s_i\tau(T_{i-1}, T_i) . \tag{15.24}$$

Table 15.6 shows an example for the valuation of a forward swap.

15.8.2   Forward Par Swap Rates

In analogue to the definition of the par swap rate for regular swaps in Sect. 15.7.2, the forward par swap rate $K_S(t, T, T')$ is determined as the interest rate for the fixed leg that sets the present value of the forward swap in Eq. 15.24 equal to zero:

$$0 = N\left(B(t, T) - B(t, T')\right) - NK_S(t, T, T')A(t, T, T') + N\sum_{i=1}^{n_f} B(t, T_i)s_i\tau(T_{i-1}, T_i) ,$$

i.e.

$$K_S(t, T, T') = \frac{1}{A(t, T, T')}\left(\left(B(t, T) - B(t, T')\right) + \sum_{i=1}^{n_f} B(t, T_i)s_i\tau(t_{i-1}, T_i)\right) . \tag{15.25}$$

Table 15.6 Forward swap fix vs. 12-months variable reference index starting in 3 years and terminating 10 years later

| | Interest rates | | | | | Swap fix vs. 1Y float over 10Y in 3Y | | | | | |
| | Spot | | Annual period | In 3 years | | Fixed leg 6% | | Float leg Nominal | | Net cash flow 250,000 | |
Term	Spot rate	In years	Forward rate	Term in years	Forward rate	Cash flow	Present value at T	Cash flow	Present value at T	Cash flow	Present value at T
1	2.20%	0	2.20%								
2	2.50%	1	2.80%								
3	2.90%	2	3.70%								
4	3.30%	3	4.51%	1	4.51%	15,000	14,353	11,273	10,787	−3,727	−3,566
5	3.70%	4	5.32%	2	4.91%	15,000	13,628	13,289	12,074	−1,711	−1,555
6	4.00%	5	5.51%	3	5.11%	15,000	12,916	13,783	11,868	−1,217	−1,048
7	4.30%	6	6.12%	4	5.36%	15,000	12,172	15,296	12,411	296	240
8	4.50%	7	5.91%	5	5.47%	15,000	11,492	14,777	11,321	−223	−171
9	4.70%	8	6.31%	6	5.61%	15,000	10,810	15,785	11,375	785	565
10	4.90%	9	6.72%	7	5.77%	15,000	10,129	16,793	11,340	1,793	1,211
11	5.00%	10	6.01%	8	5.80%	15,000	9,556	15,013	9,564	13	8
12	5.10%	11	6.21%	9	5.84%	15,000	8,997	15,516	9,307	516	309
13	5.20%	12	6.41%	10	5.90%	15,000	8,455	16,019	9,030	1,019	574
14	5.25%	13	5.90%	11	5.90%						
15	5.30%	14	6.00%	12	5.91%						
Sum of present values at start date T							112,509		109,077		−3 431
Sum of present values at valuation date t (today)							103,262		100,112		−3,149

This expression for $K_S(t, T, T')$ becomes identical to the par swap rate $K_S(T, T, T')$ once the first period of the float leg is fixed. As an example, Table 15.7 demonstrates the calculation of the par swap rate of the forward swap in Table 15.6.

15.8.3   Annuity as Numeraire

In Eq. 15.24, we calculate the present value of the forward swap by discounting the individual cash flows. Alternatively, the present value can be determined by comparing the difference between the forward par swap and the forward swap with fixed strike K. The forward par swap is the forward swap with fixed strike equal to the future par swap rate $K_S(T, T, T')$. The float leg of both swaps is identical and vanishes while calculating the difference of both swap values, with only the difference of the fixed legs remaining.

Today, the future value of the par swap rate is not known yet, therefore we have to calculate the expected value of this difference. In the terminal measure with maturity T we have, according to Eq. 14.13, the following result for the present value $V(t)$ of a forward payer swap:

$$
\begin{aligned}
\frac{V(t)}{B(t, T)} &= \mathrm{E}_t^T \left[(K_S(T, T, T') - K) A(T, T, T') \right] \\
&= \mathrm{E}_t^T \left[(K_S(T, T, T') A(T, T, T') \right] - K A(T, T, T') .
\end{aligned}
$$

The calculation of the expected value requires to choose a specific interest model, since it depends on the distribution of the par swap rate at the future time T in the terminal measure.

On the other hand, we are not bound to use the terminal measure, but are free to choose a more convenient measure and respective numeraire. An especially appealing choice of numeraire is the annuity: as a linear combination of zero bonds, the annuity is a tradable asset and fulfills all requirements of a numeraire. Also, the term in brackets in Eq. 15.25 is another linear combination of zero bonds and a tradable asset, too. In fact, it's just the present value of the float leg of the forward swap. As a consequence, $K_S(t, T, T')$ can be written as the quotient of the present value of the float leg and the numeraire

Table 15.7 The forward par swap rate equivalent to the forward swap in Table 15.6. Only the fixed rate has changed, i.e. the cash flows of the fixed leg. With a forward swap rate of 5.82% the present value of the swap is zero

	Interest rates					Forward-par-swap					
	Spot (Annual period)			In 3 years		Nominal 250,000				Forward swap rate 5.82%	
						Fixed leg		Float leg		Net cash flow	
Term	In years	Spot rate	Forward rate	Term in years	Forward rate	Cash flow	Present value at T	Cash flow	Present value at T	Cash flow	Present value at T
1	0	2.20%	2.20%	1	4.51%	14,543	13,915	11,273	10,787	−3,269	−3,128
2	1	2.50%	2.80%	2	4.91%	14,543	13,213	13,289	12,074	−1,254	−1,139
3	2	2.90%	3.70%	3	5.11%	14,543	12,522	13,783	11,868	−760	−654
4	3	3.30%	4.51%	4	5.36%	14,543	11,800	15,296	12,411	753	611
5	4	3.70%	5.32%	5	5.47%	14,543	11,142	14,777	11,321	234	180
6	5	4.00%	5.51%	6	5.61%	14,543	10,480	15,785	11,375	1,242	895
7	6	4.30%	6.12%	7	5.77%	14,543	9,820	16,793	11,340	2,251	1 520
8	7	4.50%	5.91%	8	5.80%	14,543	9,264	15,013	9,564	471	300
9	8	4.70%	6.31%	9	5.84%	14,543	8,723	15,516	9,307	973	584
10	9	4.90%	6.72%	10	5.90%	14,543	8,197	16,019	9,030	1,476	832
11	10	5.00%	6.01%	11	5.90%						
12	11	5.10%	6.21%	12	5.91%						
13	12	5.20%	6.41%								
14	13	5.25%	5.90%								
15	14	5.30%	6.00%								
Sum of present values at start date T						109,077			109,077		0
Sum of present values at valuation date t (today)						100,112			100,112		0

(i.e., the annuity). Together with Eq. 13.1 we get:

$$
\begin{aligned}
K_S(t, T, T') &= \frac{1}{A(t, T, T')} \left(\left(B(t, T) - B(t, T') \right) + \sum_{i=1}^{n_f} B(t, T_i) s_i \tau(t_{i-1}, T_i) \right) \\
&= \mathrm{E}_t^{A(t,T,T')} \left[\frac{1}{A(T, T, T')} \left(\left(B(t, T) - B(t, T') \right) \right. \right. \\
&\qquad\qquad \left. \left. + \sum_{i=1}^{n_f} B(t, T_i) s_i \tau(t_{i-1}, T_i) \right) \right] \\
&= \mathrm{E}_t^{A(t,T,T')} \left[K_S(T, T, T') \right] .
\end{aligned}
$$

Here, $\mathrm{E}_t^{A(t,T,T')}$ denotes the expected value in the annuity measure, where the arguments of A are often suppressed. This expression is analog to Eq. 14.18 for the forward rate $r_f(T, T + \tau | t)$, with the difference that we have chosen the annuity instead of the zero bond as numeraire. Thus, the forward par swap rate is a martingale in the annuity measure.

16

FX Derivatives

16.1 FX Forward Rate and Cross Currency Basis

The underlying of a *FX forward contract* (*FX* or *Forex* are abbreviations for foreign currency exchange) is a forward FX rate. In the following, we define one of the two currencies as the domestic currency. All figures are given in terms of the domestic currency if not stated otherwise. The other involved currency is denoted as foreign currency. Figures in foreign currency will be marked with the index f.

For valuation purposes, a FX rate corresponds to an underlying with a dividend yield q, where q is equal the risk-free interest rate r_f in foreign currency. Then, according to Eq. 6.1 the *forward FX rate*, i.e., the forward price of the foreign currency, is

$$S(t, T) = \frac{B^f(t, T)}{B(t, T)} S(t) . \tag{16.1}$$

Here, the discount factor in foreign currency is $B^f(t, T)$.

The above relation follows from analyzing the arbitrage free replication of the forward FX contract: An amount of $N\,€$ is used to purchase the amount of $N/S(t)$ in \$ for the spot exchange rate given by $S(t)$ €/\$. This amount is invested in the US money market over the time period $(T - t)$ until maturity at a rate r_f. Afterwards, the entire sum is transfered to € at the exchange rate $S(T)$ €/\$ valid at the future time T. The value of this amount in euros at the

end of these transactions is

$$\underbrace{S(T)}_{\text{Exchange into Euros}} \quad \underbrace{1/B^f(t,T)}_{\text{Compounding in USD}} \quad \underbrace{N/S(t)}_{\text{Amount in USD}} \quad .$$

On the other hand, if the capital N is invested directly in the EUR money market for the time period $(T-t)$ at the rate r, it will have earned $N/B(t,T)$.

If no arbitrage is possible between the spot and forward market with respect to the foreign currency exchange rate, both strategies must yield the same amount *if* the exchange rate at the future date T is already fixed at time t in a foreign exchange forward contract. The forward exchange rate $S(t,T)$ for this transaction is the value for $S(T)$ established by an arbitrage-free market as seen from today. Equating both strategies immediately yields Eq. 16.1.

So far the theory. Comparison with reality reveals, that actually traded foreign FX rate differ slightly from those calculated by Eq. 16.1. The tiny difference is called *cross currency basis* or *CCY basis*. For the major currency pairs, the CCY basis amounts to a few basis points. The origin of this CCY basis is not entirely clear. Frequently discussed arguments involve the different levels of average default probabilities in both currency areas or an unbalance of bid and offer for both currencies. Fact is that it is not possible to avoid this CCY basis, since all forward FX contracts, including the most liquid instruments like FX swaps and cross currency swaps, observe this small correction of implied forward FX rates. Anybody interested in hedging future exposure to foreign currency has to take the CCY basis into account, though it is in general not explicitly quoted (except for CCY basis swaps). For the valuation of forward FX contracts, one ore both discount factors involved will be modified such that Eq. 16.1 is exactly true for those modified interest rate curves (see Sect. 29.8). Therefore, in the following, we still could use Eq. 16.1 for pricing forward FX contracts without specifying explicitly the CCY basis.

16.2 FX Swaps

FX swaps are the most liquidly traded forward FX contracts at all. Therefore, FX swap quotes are important benchmark instruments for calibration of all valuation models which take FX risk into account, as well as for the construction of foreign interest rate curves, as we will see later. A FX swap combines a FX spot trade with a reverse forward FX contract at some future time (see Fig. 16.1) and could be stripped in these two trades. More often,
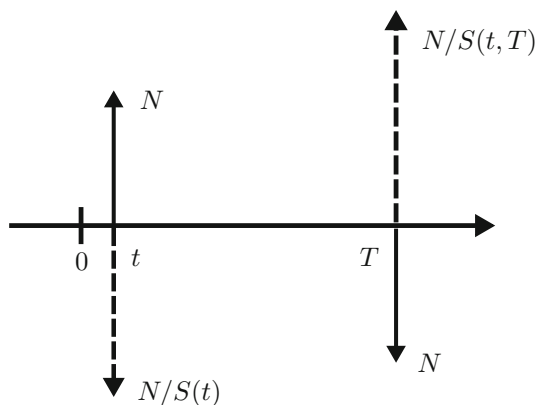
Fig. 16.1 Sale of $N€$ at a price of $N/S(t)\$$ at time t (*spot day*) and buy back of $N€$ for $N/S(t, T)\$$ at time T

however, FX forwards (or *FX outrights*) are synthesized by means of a FX swap and a FX spot trade.[1]

The difference between the forward price and the spot price is sometimes denoted as the *basis* of the FX swap, but the term *swap points* or brief *pips* is more frequently used, also avoiding confusion with the CCY basis. The swap points $b(t, T)$ depend only on the differences in risk free interest rates for both involved currencies. In simple compounding, $b(t, T)$ can be given explicitly by (see Table 2.5):

$$
\begin{aligned}
b(t, T) &\equiv S(t, T) - S(t) \\
&= \left[\frac{B^f(t, T)}{B(t, T)} - 1 \right] S(t) && \text{general} \\
&= \left[\frac{1 + r(T - t)}{1 + r_f(T - t)} - 1 \right] S(t) && \text{simple compounding} \\
&\approx (r - r_f)(T - t)S(t) && \text{linear approximation.}
\end{aligned}
\tag{16.2}
$$

In linear approximation (e.g. for short terms), the swap rate is simply the spot exchange rate multiplied by the interest spread of the two currencies involved and by the length of the time period under consideration.

---

[1]Though in general, single FX outrights are not hedged by micro hedges, but at the portfolio level.

16.2.1 Short Term Financing of Positions in Foreign Currency

With only a few exceptions, spot transactions are executed with two spot or value days, i.e. the settlement of the transaction actually takes place two business days after trade execution. Therefore, the spot price quoted today is strictly speaking the price of the asset in two business days.

For the short term financing of positions in foreign currency, it is sometimes required to trade maturities shorter than that. Here, FX swaps play an important role. At the end of each trade day, the accumulated cash positions in foreign currencies could be rolled forward by one business day (i.e. moved to the next business day) by means of FX swaps. Hereto, *overnight* (o/n) FX swaps are traded. o/n FX swaps start today (i.e. $t = 0$) and end one business day later $t + 1\text{bd}$ (here, bd stands for business day) . Equivalently, it is possible to trade *tomorrow next* (t/n, or *tom-next*) FX swaps with start at $t = 1\text{bd}$ from today and end at maturity $t = 2\text{bd}$ (i.e. spot day) to roll forward tomorrow's currency exposure. Usually, the FX swaps are structured such that the amount in foreign currency is the same for the start and the end date, accumulating the interest rate difference (pips) in domestic currency. These roll-over trades help to avoid foreign currency positions during the night.

Pips for o/n or t/n FX swaps can also be calculated as in Eq. 16.2, i.e. in relation to the spot rate. Since now $t > T$, cash flows at time T needs to be compounded rather than discounted to t, though. E.g., we get for o/n pips with spot day= 2bd:

$$b(2\text{bd}, 0) = \left[\frac{B(0, 2\text{bd})}{B^f(0, 2\text{bd})} - 1 \right] S(2\text{bd}) \ .$$

16.3 FX Forwards and FX Futures

FX forwards or *FX outrights* differ from simple FX spot trades only by the term to maturity, i.e. the contract is settled after the spot day. A FX swap could be stripped into a FX spot trade and a FX forward. Since FX swaps are more liquidly traded than FX forwards, the reverse case is more common, i.e. that a FX forward is synthesized by a portfolio consisting of a FX spot and a FX swap, where the FX spot component of the FX swap neutralizes the FX spot contract. Corporates frequently use FX forwards to hedge themselves against FX risk of future payments in foreign currency. At some exchanges, it is possible to trade *FX futures*, i.e. the exchange traded counterpart of FX

forwards, which are OTC contracts. They differ from FX forwards by the requirement to maintain a margin account (similar to other futures contracts) and standardized conventions defined by the exchange.

16.4 Cross Currency Swaps

16.4.1 Plain Vaniall CCY Swap

FX swaps with more than two years term to maturity are usually not traded without additional exchange of interest rate payments in both currencies. FX swaps with interest rate payments are called *cross currency swaps* or brief *CCY swaps*. Interest payments for both legs could be based on a variable rate index that is related to the currency area of each leg's currency. Consider, for example, a CCY swap with semi-annual payments based on the 6M EURIBOR in EUR against quarterly payments in USD based on the 3M USD LIBOR rates. Other combinations, i.e. fixed vs. float, float vs. fixed, or even fixed vs. fixed are also possible. Since the coupon payments are in different currencies, the latter form fixed-fixed does make sense here, in contrast to an interest rate swap with both legs in the same currency. Similar to an FX swap, a CCY swap may include the exchange of notional at start and at maturity in both currencies. At trade date, nominal and interest rate (if at least one of both legs pays or receives coupons based on a fixed interest rate) are determined such that the present value vanishes. In general, for one of the two currencies, the notional is some constant value over the whole term to maturity. A typical use case for such a CCY swap is the hedge against the FX risk initiated by a position in a bond or loan in foreign currency. In this case, the CCY swap would be structured such that the leg in foreign currency matches exactly all payments of the original position with opposite sign. If the payments in home currency are variable, the whole package equals a *synthetic floater*. The risk profile of such a synthetic floater differs from a "real" floater in respect of default risk, CCY basis spread risk (see below), and liquidity risk.

16.4.2 Mark to Market CCY Swaps

At trade date, the present value of a CCY swap is zero and this holds also for the nominal exchange at maturity in particular. Over the full life time of the swap the FX rate will vary and the present value of the final nominal exchange may fluctuate considerably, especially with for the commonly long

terms to maturity of up to 20 or 30 years. If the CCY swap trade is not subject to a collateral management agreement, this exposure is fully exposed to credit default risk and the financial institution is required to reserve a proper amount of capital for this trade.

To avoid this credit risk effect, it is possible to enter instead into a *mark to market CCY swap* or brief MTM CCY swap. Here, the nominal for one leg of the CCY swap is fixed for the full life time, while the nominal for the other leg of the swap will be adjusted at the beginning of each interest rate period by applying the FX spot rate at that future date, such that both nominals will have the same present value again. The difference between the nominal of the previous period and the new nominal will compensated by a payment equal the nominal difference (in nominal currency). Therefore, the present value of the nominal exchange caused by FX rate changes will be reset to zero after each period.

16.4.3 Cash Flows and Present Value

For the present value calculation, we will restrict ourselves to fix-float MTM CCY swaps, since the consideration could easily be extended to other variants. We assume that the MTM CCY swap pays fixed rate coupons in foreign currency and receives floating rate coupons in domestic currency, and the MTM adjustments will be made on the leg in domestic currency. Then, the nominal in foreign currency N_f is constant over the full lifetime of the swap. The FX spot rate $S(t)$ is given as price of one unit in foreign currency, expressed in domestic currency. $S(t, T)$ is the appropriate forward FX rate. In practice, the nominal exchange at the start of the swap is often excluded from the trade, since it could be replaced by a spot trade and is anyway worthless once the swap has started already. For simplifying the notation, we will therefore neglect this first notional exchange.

With the above preliminaries, the present value of the fixed leg $V_{\text{MTM CCY}}^{\text{fix}}(t)$ equals the present value of a bond in foreign currency (Eq. 15.6):

$$V_{\text{MTM CCY}}^{\text{fix}}(t) = \frac{1}{S(t)} \left(N_f K A_f(t, t_0, t_n) + N_f B_f(t, t_n) \right) ,$$

where

$$A_f(t, t_0, t_n) = \sum_{i=1}^{n} B_f(t, t_i) \tau(t_{i-1}, t_i) \mathbf{1}_{t < t_i} .$$

is in analogy to Eq. 15.7 the annuity in foreign currency and K is the fixed interest rate. The factor $1/S(t)$ transfers the result from foreign to domestic currency.

For valuation of the variable leg, we require to change the expression for the present value of a floater (Eq. 15.20): The nominal hast to be adjusted according to the MTM rule and the additional payments for the nominal differences need to be added. Thus, the present value for the float leg is

$$
\begin{aligned}
V_{\text{MTM CCY}}^{\text{float}}(t) =\, & S(t_m)N_f \left(1 + R^{\text{fix}}\tau(t_m, t_{m+1})\right) \\
& + N_f \sum_{i=m+2}^{n} S(t, t_{i-1})B(t, t_i) \left[\frac{B(t, t_{i-1})}{B(t, t_i)} - 1 + s_i \tau(t_{i-1}, t_i)\right] \\
& + N_f S(t, t_{n-1})B(t, t_n) \\
& + N_f \sum_{i=m+1}^{n-1} B(t, t_i)\left(S(t, t_i) - S(t, t_{i-1})\right) \ .
\end{aligned}
$$

Since $t > t_m$, $S(t_m)$ lies in the past, and therefore denotes a FX rate which has already been fixed at t_m. In contrast to the floater Eq. 15.20, this expression could not further be simplified because of the additional factors $S(t, t_{i-1})$ in the second line. Instead of giving the basis spreads explicitly, we could again make use of a forward curve in accordance with Eq. 15.3, which yields

$$
\begin{aligned}
V_{\text{MTM CCY}}^{\text{float}}(t) =\, & S(t_m)N_f \left(1 + R^{\text{fix}}\tau(t_m, t_{m+1})\right) \\
& + N_f \sum_{i=m+2}^{n} S(t, t_{i-1})B(t, t_i) \left[B_\tau^{-1}(t_{i-1}, t_i|t) - 1\right] \\
& + N_f S(t, t_{n-1})B(t, t_n) \\
& + N_f \sum_{i=m+1}^{n-1} B(t, t_i)\left(S(t, t_i) - S(t, t_{i-1})\right) \ .
\end{aligned}
$$

Here, $B_\tau(t, T)$ denotes the forward curve for the interest rate index of the float leg. It is worth mentioning that the leg's present value could become negative due to the MTM payments related to the nominal, if the FX rate falls significantly over time.

Now, we can simply write for the present value of the MTM CCY swap

$$V_{\text{MTM CCY}}(t) = V_{\text{MTM CCY}}^{\text{float}}(t) - V_{\text{MTM CCY}}^{\text{fix}}(t) \; .$$

16.5 CCY Basis

So far, we have just used the discount factor curves in domestic and foreign currency without consideration how these curves have been constructed. The first approach would be to use the risk neutral curves in each currency, which can be constructed as described in detail in Sect. 29.6. For example, one would use the USD OIS curve for the USD market and the EONIA curve for the EUR market. However, if one would calculate the PVs of actually quoted FX and CCY swaps, the resulting PVs would be systematically slightly different from zero in contradiction to the assumptions. Indeed the previously described procedure is not fully consistent with market practice. Instead, for the valuation of forward FX derivatives, for at least one of the two involved currency a discount factor curve that has been adjusted for the so called *cross currency basis* or brief *CCY basis* needs to be applied.

For collateralized forward FX trades, the currency of the cash collateral determines, which curve needs to be adjusted. For most currency pairs, USD is the collateral currency. In this case the unchanged USD OIS curve would serve as discount curve for cash flows in USD, regardless whether USD is the domestic or foreign currency. Then, the discount curve for the other, non-USD currency, needs to be adjusted for the CCY basis spread. In the case of a CCY swap exchanging payments in GBP vs. EUR with collateral in USD, this would imply that the cash flows for both legs need to be discounted by a CCY discount curve (i.e. a discount curve adjusted for the CCY basis). Liquid quotations of FX and CCY swaps are used to calculate the implied CCY basis spread curve (see Sect. 29.8). This way, it is guaranteed that all forward FX derivatives can be valuated consistently.

The existence of the CCY basis causes many problems. An especially nasty problem is related to Hedge Accounting according to IFRS. The forward FX rate $S(t, T)$ is implicitly dependent on the CCY basis, too, i.e. one or both discount curves included in Eq. 16.1 are CCY discount curves, in order to calculate the market conform forward FX rate. Although the CCY basis could just be added to or subtracted from the risk neutral rate curve to form a CCY discount curve, the CCY basis is no real interest rate, similar to the tenor basis spread of forward rates. The question arises, which effect causes the CCY basis. An often heard explanation argues that the CCY basis reflects the difference

of the average default risk levels of market participants in both currency areas. A more pragmatic view is this: Since it is not possible to avoid the CCY basis while trading and financing foreign currency derivatives, market maker could just factor them in to reflected differences in offer and demand for certain currency pairs.

17

Variants of Fixed Income Instruments

Interest rate derivatives trade with a very broad variety of different features. In the following, a selection of these various features is presented. Though each of the following examples will focus on one particular feature, it should be noted that different features are frequently combined in a single trade.

17.1 Basis Swaps

Basis swaps have two legs like Plain Vanilla swaps, but none of these are based on a fixed rate. Instead, the interest coupons on both legs are based on variable interest rate index (i.e. floating legs), but for different interest rate period lengths. For example, a typical basis swap exchanges pays quarterly a coupon based on the 3M EURIBOR reference index and receives semi-annually a coupon based on the 6M EURIBOR. The tenor basis spread over the risk free interest rate will be different for both rate indexes. Therefore, a small spread adjustment is added to the rate index for one leg in order to make the basis swap worthless at trade day. In general, the spread is added to the leg with the shorter interest rate period (or tenor), yielding in a positive spread. The interest rate periods are in general chosen to be consistent with the periods belonging to the rate reference index.

17.1.1   Present Value and Cash Flows of Basis Swaps

The present value of a floating swap leg equals the present value of a swap minus the fixed coupon payments. With Eq. 15.22, we get for both legs 1 and 2 of the basis swap:

$$V_1(t) = N \left[B(t, t_{1,m_1+1}) - B(t, t_{1,n_1}) \right] + B(t, t_{1,m_i+1}) C_Z^{\text{fix}}(t_{1,m_i+1})$$

$$+ N \sum_{i=m_1+2}^{n_1} B(t, t_{1,i}) \tau(t_{1,i-1}, t_{1,i}) s_{1,i}$$

$$V_2(t) = N \left[B(t, t_{2,m_2+1}) - B(t, t_{2,n_2}) \right] + B(t, t_{2,m_2+1}) C_Z^{\text{fix}}(t_{2,m_2+1})$$

$$+ N \sum_{i=m_2+2}^{n_2} B(t, t_{2,i}) \tau(t_{2,i-1}, t_{2,i}) (s_{2,i} + s_{\text{b}}) . \qquad (17.1)$$

Here, index 1 denotes the leg based on the reference index with the longer period. The present value of the basis swap is simply the difference of both leg's present values, i.e. $V(t) = V_1(t) - V_2(t)$ or $V(t) = V_2(t) - V_1(t)$, depending on which leg pays/receives interest coupons. This both legs terminate at the same maturity date, we have $t_{1,n_1} = t_{2,n_2}$, though $n_1 \neq n_2$ because of the different periods lengths For the same reason the coupon dates $t_{1,i}$ and $t_{2,i}$ differ. In general, all coupon payment dates of leg 1 (with the longer interest rate period) are also payment dates for leg 2, but not vice versa. The time-dependent basis spreads $s_{1,i}$ and $s_{2,i}$ also differ for the different rate reference indexes. s_b is the flat basis spread to be added to the rate index with the shorter period such that the present value vanishes at trade date.

Alternatively, the present value could be expressed in terms of special tenor interest rate curves, as has been used in Eq. 15.3. Naming both tenor curves $B_1(t, T)$ and $B_2(t, T)$ respectively, we get

$$V_1(t) = N \left[B(t, t_{1,m_1+1}) - B(t, t_{1,n_1}) \right] + B(t, t_{1,m_i+1}) C_Z^{\text{fix}}(t_{1,m_i+1})$$

$$+ N \sum_{i=m_1+2}^{n_1} B(t, t_{1,i}) \left(B_1^{-1}(t_{1,i-1}, t_{1,i}|t) - 1 \right)$$

$$V_2(t) = N \left[B(t, t_{2,m_2+1}) - B(t, t_{2,n_2}) \right] + B(t, t_{2,m_2+1}) C_Z^{\text{fix}}(t_{2,m_2+1})$$

$$+ N \sum_{i=m_2+2}^{n_2} B(t, t_{2,i}) \left(B_2^{-1}(t_{2,i-1}, t_{2,i}|t) - 1 + \tau(t_{2,i-1}, t_{2,i}) s_{\text{b}} \right) .$$

17.1.2   Par Basis Swap Spread

The par basis swap spread s_b is determined at trade date, such that the present value vanishes, i.e. that both legs in Eq. 17.1 have the same value $V_1(t) = V_2(t)$. If the swap starts at value date and the rate references for both first coupon periods have not been fixed yet, the terms with the fixed coupons can be dropped. In addition, the term $B(t, t_0) - B(t, t_n)$ with $t_n = t_{1,n_1} = t_{2,n_2}$ is identical for both legs and vanishes, too. It remains

$$
s_b = \frac{1}{\sum\limits_{i=1}^{n_2} B(t, t_{2,i}) \tau(t_{2,i-1}, t_{2,i})} \left(\sum\limits_{i=1}^{n_1} B(t, t_{1,i}) \tau(t_{1,i-1}, t_{1,i}) s_{1,i} \right.
$$

$$
\left. - \sum\limits_{i=1}^{n_2} B(t, t_{2,i}) \tau(t_{2,i-1}, t_{2,i}) s_{2,i} \right) .
$$

If we approximate the par basis swap spread s_b applying the (false) assumption the coupon periods have equal lengths and both basis spreads are constant over time, i.e. $s_{1,i} = s_1$ and $s_{2,i} = s_2$, then s_b is just given by the difference of both spreads, $s_b \approx s_1 - s_2$.

17.2   Annuity Loans

In light of the increasing demand for improved credit risk management it is required to evaluate a bank's loan portfolio according to standard arbitrage free pricing methods on a regular basis. This is especially necessary, if the interest rate and credit risks embedded in the loan portfolio should be hedged means of interest rate swaps, credit default swaps, and other derivatives.

As an example for the mark-to-market of loans, we consider in the following the calculations for a typical annuity loan.

17.2.1   Cash Flows and Residual Debt

Like a bond, an *annuity loan* is determined by its cash flows. These cash flows are the constant installments C paid during the lifetime of the loan and the payment of the residual debt at maturity. The fundamental difference between an annuity loan and a coupon bond is that payments on the principal are made

during the lifetime of the loan, i.e., before maturity. Thus, the *residual debt* $N(t)$ by reference to which the interest payments are determined, decreases with time. The amount C paid at the end of each interest period covers the interest accrued in that period. The amount of C exceeding the interest payment due goes toward repaying the principal on the loan. The contractually agreed upon interest rate is denoted by K. The payment C can be decomposed into an interest cash flow C_Z and an amortization cash flow C_T as follows:

$$C = C_Z(t_i) + C_T(t_i) \quad \forall \ i = 1, \ \dots \ , \ n \tag{17.2}$$

with

$$C_Z(t_i) = N(t_{i-1})K\tau(t_{i-1}, t_i)$$
$$C_T(t_i) = C - N(t_{i-1})K\tau(t_{i-1}, t_i) \ .$$

The residual debt $N(t_i)$ at time t_i is the amount on which interest must still be paid after the time t_i. An expression for this residual debt can be derived recursively. For simplification of notation, we set $N_i = N(t_i)$ and $z_i = 1 + K\tau(t_{i-1}, t_i)$. For the first interest period, the nominal equals $N_0 = N$, i.e. the initial loan amount. For the following interest rate periods, the remaining debt could be calculated iteratively as follows:

$$N_1 = Nz_1 - C$$
$$N_2 = N_1z_2 - C = (Nz_1 - C)z_2 - C = Nz_1z_2 - C(1 + z_2)$$
$$N_3 = N_2z_3 - C = (Nz_1z_2 - C(1 + z_2))z_3 - C = Nz_1z_2z_3 + C(1 + z_2z_3 + z_3)$$

$$N_i = N\prod_{j=1}^{i} z_j - C\left[1 + \sum_{k=2}^{i}\prod_{j=k}^{i} z_j\right] \ . \tag{17.3}$$

In the case that all interest periods have exactly the same period lengths, i.e. $z_i = z = $ const., the expressions simplifies to

$$N_i = Nz^i - C\sum_{j=0}^{i-1} z^j = Nz^i - C\frac{1 - z^i}{1 - z} \ . \tag{17.4}$$

In the last step of the last transformation, we have again applied the formula for the sum of the Geometric series. At some time t between to coupon dates,

$t_m < t < t_{m+1}$, the residual debt increases by the accrued interest until t. This yields for the residual debt $N(t)$ at time t

$$N(t) = \left[N z^m - C \frac{1 - z^m}{1 - z} \right] (1 + K \tau(t_m, t)) \qquad (17.5)$$

17.2.2 Present Value

In addition to the installment payments, the final amortization of the residual debt $N(t_n)$ occurs at maturity, $T = t_n$. This is treated as a cash flow for the purpose of calculating the present value. The amount of this cash flow is given by Eq. 17.3 with $i = n$. The present value as the sum of all discounted future cash flows thus becomes:

$$V(t, T) = \sum_{i=m+1}^{n} B(t, t_i) \underbrace{C}_{\text{installments}} + B(t, t_n) \underbrace{N(t_n)}_{\text{residual debt}} \quad \text{for } t_m \leq t < t_{m+1}$$

$$(17.6)$$

$$= C \sum_{i=m+1}^{n} B(t, t_i) + B(t, t_n) \left[N \prod_{j=1}^{i} z_j - C \left(1 + \sum_{k=2}^{i} \prod_{j=k}^{i} z_j \right) \right].$$

Note that the residual debt and the installment payments are fully determined by the terms of the loan contract. The only variable parameters in the present value are the discount factors obtained from the current interest rate curve. In the corresponding equation for a bond (Eq. 15.6), the spot rate discount factors also appear as the only parameters not fixed and are thus the only variables which are influenced by the market. Thus, by equating the coefficient of $B(t, t_i)$ for $i = m + 1, \ldots n$ in Eqs. 15.6 and 17.6, we obtain a bond whose behavior with respect to its present value and market risk is exactly the same as the loan, if the same spot rate curve is used as was for the loan.

$$N_{\text{bond}} K_{\text{bond}} \tau(t_{i-1}, t_i) = C_{\text{loan}} \qquad (17.7)$$

$$N_{\text{bond}} = N(t_n)_{\text{loan}}$$

$$= \left[N \prod_{j=1}^{i} z_j - C \left(1 + \sum_{k=2}^{i} \prod_{j=k}^{i} z_j \right) \right]_{\text{loan}}.$$

An annuity loan with a principal of N, a fixed rate K and constant installments C can, in accordance with Eq. 17.7, be interpreted as a coupon bond whose principal is equal to the residual debt of the loan at maturity and whose coupon is chosen such that the cash flows resulting from the coupon payments are exactly the same as the cash flows from the interest *and amortization* of the loan. All cash flows of such a bond are thus exactly those of the corresponding loan. In terms of valuation, cash management, and risk management, there are therefore no fundamental differences between bonds and annuity loans.

In general, the residual debt of the loan at maturity is not equal to zero, even when the term of the loan runs until complete amortization. At time t_n of the last installment payment, the residual debt is not exactly equal to the amortization component of the installment C, but is usually somewhat larger. The last payment date t_n of such a loan is the date after which the residual debt is smaller than C. As a rule, this residual debt is paid at t_n as well. For this type of loan, the residual debt and thus the principal of the associated bond is indeed quite small but Eq. 17.7 can be applied nevertheless. If a loan happens to be completely amortized upon payment of the final installment, Eq. 17.6 gives the present value as simply $\sum B(t, t_i) C$, i.e., equal to the present value of a series of zero bonds, each with face value C and whose respective maturities correspond to the installment payment dates of the loan. An interpretation as a coupon bond is thus no longer necessary.

An explicit example of these concepts is provided in Tables 17.1 and 17.2 for an $8\frac{1}{2}$ year loan of one million euros at a fixed rate of 8.4%, with semi-annual installment payments of exactly 50,000 euros (this corresponds to an

Table 17.1 The equivalent coupon bond for an annuity loan with fixed rate 8.40% per year and semi-annual payments

	Bond	Loan
Nominal	788,264	1,000,000
Value date	21. Feb. '01	21. Feb. '01
Original term		8.5
Time to maturity	7.358	7.358
Coupon period	0.50	0.50
Installment C	50,000	50,000
Initial yearly amortization		1.81%
Fixed rate	13.09%	8.40%
YTM	9.35%	9.35%
Par rate	9.45%	9.39%

Table 17.2 The equivalent coupon bond's cash flows from Table 17.1. All present values are as of February, 21st, 2001

| | Rate curve | | | Equivalent bond | | | |
| | | | | coupon = fixed rate | | coupon = par rate | |
	Term [years]	Spot rate	Cash Flow	PV based on spot rate	PV based on YTM	PV based on spot rate	Cash Flow
30. Jan. '01	0.358	2.20%	50,000	49,612	48,424	36,124	36,407
31. Dez. '01	0.861	3.20%	50,000	48,662	46,297	35,432	36,407
30. Jun. '02	1.358	4.50%	50,000	47,098	44,285	34,294	36,407
31. Dez. '02	1.861	5.50%	50,000	45,258	42,339	32,954	36,407
30. Jun. '03	2.358	6.30%	50,000	43,291	40,499	31,521	36,407
31. Dez. '03	2.861	7.10%	50,000	41,090	38,720	29,919	36,407
30. Jun. '04	3.358	7.90%	50,000	38,732	37,037	28,202	36,407
31. Dez. '04	3.861	8.50%	50,000	36,490	35,410	26,569	36,407
30. Jun. '05	4.358	9.00%	50,000	34,344	33,871	25,007	36,407
31. Dez. '05	4.861	9.30%	50,000	32,451	32,383	23,629	36,407
30. Jun. '06	5.358	9.50%	50,000	30,745	30,975	22,386	36,407
31. Dez. '06	5.861	9.60%	50,000	29,217	29,614	21,274	36,407
30. Jun. '07	6.358	9.70%	50,000	27,754	28,327	20,208	36,407
31. Dez. '07	6.861	9.80%	50,000	26,326	27,083	19,169	36,407
30. Jun. '08	7.358	9.90%	838,264	418,510	434,318	411,724	824,670
			Present value	949,581	949,581	798,413	
			Residual debt	802,120	802,120	798,413	
			Accrued interest	13,856	13,856	10,149	
			Market price	118.71%	118.71%	100.00%	

initial annual amortization of 1.81%). The associated bond has a face value of
746,470 euros and a coupon of 13.85%. Observe that for the calculation of
the bond, all relevant information can be obtained from the terms of the loan
contract. The current term structure and even the current present value is not
required.

The YTMs and par rates for the loan and its equivalent bond are included
in Fig. 17.1. A further discussion of the YTM and par rate for the loan can
be found in Sect. 17.2.3. The interest rates for the bond are determined as
follows: in the column "PV Using YTM" the *future* cash flows are discounted
at the YTM. In accordance with Table 5.2, the value of the YTM was adjusted
(using the Excel function "Goal Seek") until the sum of the thus discounted
cash flows equaled the present value obtained by using the current spot rates.
In the column "PV Using Spot Rate" under "Coupon = Par Rate", the present
value of a bond with the par rate as the coupon was calculated varying the par
rate until the present value thus calculated equaled the residual debt.

Note that both for the bond and the loan, the par rate is not the same as
the YTM. Furthermore, the par rate of the loan and the equivalent bond are
differ significantly although the future cash flows are exactly the same! This is
because the bond and the loan have different *current* residual debts. The par
rate, an extremely instrument-sensitive parameter, reveals this difference. An
instrument is thus not fully characterized by its future cash flows. The open
claims associated with it are needed as well. The instrument is only completely
specified when both the future cash flows and the current residual debt are
known. The YTM depends only on the future cash flows and is therefore
identical for both, loan and equivalent bond.

In the above example, the accrued interest, defined as the difference between
the residual debt and the face value, is also calculated, as well as the price of
the bond, quoted as the clean price (present value less the accrued interest)
divided by the face value. As expected, the bond price with the par rate as the
coupon is exactly 100%.

In the special case of constant interest rate period lengths, Eq. 17.4 can be
used for the residual debt and the present value Eq. 17.6 of the loan becomes

$$
V(t, T) = \left[Nz^n - C\frac{1 - z^n}{1 - z} \right] B(t, t_n) + C \sum_{i=m+1}^{n} B(t, t_i) . \qquad (17.8)
$$

17.2.3   Yield to Maturity and Par Rates

Substituting the cash flows 17.2 into the at-par condition, Eq. 5.5, this condition reads for annuity loans

$$N(t) = C \sum_{i=m+1}^{n} B(t, t_i) + B(t, t_n) N(t_n) \text{ mit}$$

$$t_m \leq t < t_{m+1} < t_{m+2} < \cdots < t_n = T \ , \ n \geq m+1 \ . \tag{17.9}$$

Finally, with Eq. 17.3 for the residual debt, we have

$$\left[N \prod_{j=1}^{m} z_j - C \left(1 + \sum_{k=2}^{m} \prod_{j=k}^{m} z_j \right) \right] (1 + K\tau(t_m, t))$$

$$= C \sum_{i=m+1}^{n} B(t, t_i) + B(t, t_n) \left[N \prod_{j=1}^{n} z_j - C \left(1 + \sum_{k=2}^{n} \prod_{j=k}^{n} z_j \right) \right] . \tag{17.10}$$

The fixed rate K, on which the z_i depend, satisfying this condition is the par rate. Comparison with Eq. 15.12 reveals that, because of the amortization, the par rates of an annuity loan are quite different from those for bonds. This even holds for the equivalent bond associated with the loan as is quite evident from Fig. 17.1. Thus, the yield curve (par rates of coupon bonds) cannot be utilized for the valuation of loans!

To calculate the YTM, the discount factors $B(t, t_i)$ on the right hand side of Eq. 17.6 need to be replaced by discount factors $B_A(\bar{R}, t, t_i)$ based on a time independent YTM $\bar{R}$. Then, $\bar{R}$ has to be varied until both sides of the equation are equal, i.e. the right hand side equals the present value $V(t, T)$.

Here again, it can be shown that in the special case of constant interest rate period lengths of exactly one year the YTM equals at the start of an interest rate period exactly the par rate. We abstain here to write down the explicit calculation, since it does not provide any new enlightenment (see Sect. 5.3).

In Table 17.3, all cash flows of the loan from Fig. 17.2 are listed and priced with respect to the spot rates belonging to the debtors credit worthiness. The two columns "interest" and "amortization" show explicitly how the constant installment (cash flow) can be decomposed into interest (at the fixed rate of 8.40% specified in the contract) and amortization payments. The residual debt

Table 17.3 Cash flow table of the annuity loan from Table 17.2. The next payment date is June 30th, 2001. Thus, in our notation we have $m = 2$ and $n = 17$. Par Rate and YTM were determined via numerical iteration. The results for those yields are displayed in Table 17.1. All present values are given as of February 21st, 2001

	IR Term Structure						Annuity Loan			
i	Payment Date	Term [years] 30E/360	Spot Rate	Cash Flow	Interest	Amortization	Residual Debt	Discounted at Spot Rates	Discounted at YTM	Residual Debt using par rate as fixed rate
0	31. Dez. '99	−1.142					1,000,000			1,000,000
1	30. Jun. '00	−0.642		50,000	41,153	8,847	991,153			995,905
2	31. Dez. '00	−0.142		50,000	40,789	9,211	981,942			991,622
3	30. Jun. '01	0.358	2.20%	50,000	40,410	9,590	972,352	49,612	48,424	987,143
4	31. Dez. '01	0.861	3.20%	50,000	40,015	9,985	962,368	48,662	46,297	982,458
5	30. Jun. '02	1.358	4.50%	50,000	39,605	10,395	951,972	47,098	44,285	977,558
6	31. Dez. '02	1.861	5.50%	50,000	39,177	10,823	941,149	45,258	42,339	972,433
7	30. Jun. '03	2.358	6.30%	50,000	38,731	11,269	929,880	43,291	40,499	967,073
8	31. Dez. '03	2.861	7.10%	50,000	38,268	11,732	918,148	41,090	38,720	961,467
9	30. Jun. '04	3.358	7.90%	50,000	37,785	12,215	905,933	38,732	37,037	955,603
10	31. Dez. '04	3.861	8.50%	50,000	37,282	12,718	893,215	36,490	35,410	949,470
11	30. Jun. '05	4.358	9.00%	50,000	36,759	13,241	879,973	34,344	33,871	943,056
12	31. Dez. '05	4.861	9.30%	50,000	36,214	13,786	866,187	32,451	32,383	936,347
13	30. Jun. '06	5.358	9.50%	50,000	35,646	14,354	851,833	30,745	30,975	929,330
14	31. Dez. '06	5.861	9.60%	50,000	35,056	14,944	836,889	29,217	29,614	921,991
15	30. Jun. '07	6.358	9.70%	50,000	34,441	15,559	821,330	27,754	28,327	914,316
16	31. Dez. '07	6.861	9.80%	50,000	33,800	16,200	805,130	26,326	27,083	906,288
17	30. Jun. '08	7.358	9.90%	50,000	33,134	16,866	788,264	24,963	25,906	897,891
					PV final residual debt			393,547	408,412	448,279
					Present value loan			949,581	949,581	1,004,313
					Present residual debt			993,227	993,227	1,004,313

remaining after each payment is shown in the column "Residual Debt". All of the values for the residual debt listed in the example were computed using Eq. 17.3 or Eq. 17.5. At the valuation date Feb-21-01, the present value of the loan is 949,581.00 EUR, the current residual debt (i.e., the residual debt as of the last payment date compounded up to February 21st), however, is equal to 993,227.00 euros. The present value is thus less than the remaining debt! This means that the interest paid by the debtor on the loan is too low considering his or her current credit worthiness. This example demonstrates the mark to market of a loan.

In the column "Residual Debt Using Par Rate as Fixed Rate", the residual debt was calculated as if the par rate were the fixed interest rate instead of the rate agreed to in the contract. Using Eq. 17.9 and Eq. 17.10, the par rate was adjusted (using the Excel function "Goal Seek") until the present value of the loan (calculated as before with the spot rates) was equal to the current residual debt. The par rate is thus the rate which the debtor should pay if the transaction is to be "fair". At 9.49%, it is substantially higher than the agreed upon 8.40% rate. In the column "Present Value with YTM" the cash flows and the final residual debt (calculated at the contractually agreed upon rate) were discounted at the YTM rather than with the spot rates and the YTM was adjusted until the sum of these discounted cash flows equaled the actual present value. All of these yields are listed in Fig. 17.1.

17.2.4 Sensitivities

The sensitivities of an annuity loan with respect to the yield to maturity, i.e., the modified duration and the convexity are given in their full generality (i.e., without the assumption that the interest periods are of equal length) by Eq. 17.3 as

$$D_{\mathrm{mod}} = -\frac{1}{V}\frac{\partial V}{\partial \overline{R}} = -\frac{C}{V}\sum_{i=m+1}^{n}\frac{\partial B_A(\overline{R}, t, t_i)}{\partial \overline{R}} - \frac{N(t_n)}{V}\frac{\partial B_A(\overline{R}, t, t_n)}{\partial \overline{R}}$$

$$(17.11)$$

$$\mathrm{Convexity} = \frac{1}{V}\frac{\partial^2 V}{\partial \overline{R}^2} = \frac{C}{V}\sum_{i=m+1}^{n}\frac{\partial^2 B_A(\overline{R}, t, t_i)}{\partial \overline{R}^2} + \frac{N(t_n)}{V}\frac{\partial^2 B_A(\overline{R}, t, t_n)}{\partial \overline{R}^2} \ .$$

The sensitivities with respect to the spot rates, namely the key rate durations are

$$D_{\text{key}}(t, t_i) = -\frac{C}{V} \frac{\partial B(t, t_i)}{\partial R_i} \quad \forall \; i = m+1, \ldots, n-1$$

$$D_{\text{key}}(t, t_n) = -\frac{C + N(t_n)}{V} \frac{\partial B(t, t_n)}{\partial R_n} \;. \tag{17.12}$$

The derivatives can, again using Eq. 5.15, be interpreted as the sensitivities of zero bonds, each multiplied by the present value of the corresponding zero bond.

Alternatively, the YTM and sensitivities can be calculated for the equivalent bond using Eq. 17.7. Since the sensitivities depend exclusively on the future cash flows and not on the residual debt, the sensitivities for the bond listed in Fig. 17.1 are equal to those of the associated annuity loan as well. In this way, mark-to-market and risk management of a credit portfolio can be performed using the traditional sensitivities.

17.3 Fixing-in-Arrears

A floater or a swap's floating leg with *fixing in arrears* is a variant of where the variable reference rate index is not fixed at the start of the interest rate period, but at the end. Typically, the fixing day lies as many business days before the end of the coupon period as is given by the spot day convention, such that the actual start date the rate index is related to equals the end date T of the current coupon period. This forward interest rate is no longer a martingale in the T-measure. Instead, a change of measure from the T_f-measure (T_f being the actual maturity of the interest rate period the rate index refers to) to the T-measure is required. This yields a drift correction, also called *convexity adjustment* (see Sect. 14.5).

17.4 Float Rate with Cap/Floor

Another variant of a floater or swap is a limitation of the variable rate index by an upper limit (cap) or lower limit (floor). Such a Cap or Floor is effectively an option that can be stripped from the original product. This type of option will be discussed in more detail in Sect. 18.6.2.

17.5 Call Rights and Break Clause

Bonds and Swaps have often embedded rights to cancel (call) the product that can be exercised at some fixed date (European call right) or at multiple dates (Bermudan call right). In a view cases, the call right can be exercised at any time before maturity (American call right). From a mathematical perspective, a bond a swap with embedded call rights can always be stripped into a bond resp. swap without call rights and a separate option to sell the bond or swap, a so called swaption. Therefore, we refer to Sect. 18.6.4 for a detailed discussion.

A *break clause* also allows to terminate the contract prior to maturity, but at the then current market value. Without considering credit default risk, such a feature does not have an impact on the valuation. Taking credit default risk into account, the impact on valuation maybe significant, though, since the exposure is limited.

17.6 Reverse Floater

The reverse floater's coupon is based on a fixed interest rate, the *offset*, subtracted by a variable rate reference index (i.e. 6M-LIBOR), sometimes multiplicated by some factor. In addition, the coupon is often limited by an upper (cap) or lower (flow) bound. E.g., the interest rate R for an interest rate period is calculated as follows:

$$R = \min\left(\max\left(O - \alpha L(t_{\text{fix}}), F\right), C\right) ,$$

where

$$O \text{ offset}$$
$$\alpha \text{ multiplicative factor}$$
$$L \text{ LIBOR interest rate}$$
$$t_{\text{fix}} \text{ fixing day of LIBOR Rate}$$
$$F \text{ lower bound of coupon rate}$$
$$C \text{ upper bound of coupon rate}$$

For valuation purposes, the reverse floater could be stripped into a fixed rate bond, a Plain Vanilla floater, a Floor, and a Cap.

17.7 Constant Maturity Swaps

Constant maturity swaps or CMS are special variants of Plain Vanilla swaps where the variable rate reference index is itself a par swap rate. For example, one leg of the swap pays an annually coupon with an amount equal to the quoted swap rate for a fixed maturity (= "Constant Maturity") of, e.g., 10 years multiplied by the nominal and the lengths of the coupon period in years. The other leg could be based on the 6M-EURIBOR plus a constant spread.[1] The constant spread is chosen at trade date such that the present value of the swap is zero. Alternatively, the other leg may pay a fixed rate coupon.

Since the present value of a CMS is just the sum of the present values of the single cash flows (the single cash flows are independent from each other), we consider for simplicity only the present value of a single CMS cash flow for the coupon period with start date T_1 and end date T_2. Then, in the annuity measure, we get the following expression:

$$
\begin{aligned}
V_{\text{CMS Kupon}} &= A(t, T, M)\mathrm{E}_t^A \left[N \frac{B(T, T_p)K_S(T, T, M)}{A(T, T, M)} \tau(T_1, T_2) \right] \\
&= N\tau(T_1, T_2)A(t, T, M)\mathrm{E}_t^A \left[\frac{B(T, T_p)K_S(T, T, M)}{A(T, T, M)} \right].
\end{aligned}
\tag{17.13}
$$

T is the start date of the underlying par swap rate index and M its maturity. T_p is the coupon payment day and must be greater than or equal to the fixing date T, since the coupon amount is otherwise unknown, therefore $T_p \geq T$ (for simplicity, we assume that T is equal to the fixing date; often, the fixing date is one ore two banking days prior to the start date).

Further model assumptions are required for the calculation oft the expected value. In general, T equals T_1 (fixing in advance) or T_2 (fixing in arrears). In the special case $T = T_2 = T_p$, the expression for the expectation value simplifies a bit.

---

[1]For CMS and other structured swaps the definition of receiver and payer swaps is ambiguous, since both legs may pay variable coupons. In such a case, a swap is a receiver or payer, when the CMS-leg payments are received or payed, i.e., payer and receiver always belong to the more "interesting" leg.

Approximately, the expectation value could be calculated as a convexity adjustment. The general procedure follows the approach in Sect. 14.5. Formally, the convexity adjustment $K_S^{\text{adjustment}}(t, T, M, T_p)$ is given by

$$
\begin{aligned}
K_S^{\text{adjustment}}(t, T, M, T_p) &\equiv \mathrm{E}_t^T\left[K_S(T, T, M)\right] - K_S(t, T, M) \\
&= \frac{A(t, T, M)}{B(t, T_p)} \mathrm{E}_t^A\left[\frac{B(T, T_p)K_S(T, T, M)}{A(T, T, M)}\right] - K_S(t, T, M)\,.
\end{aligned}
$$
$$(17.14)$$

18

Plain Vanilla Options

18.1 Traditional and General Definition of an Option

In the traditional sense, an *option* gives its purchaser the right to buy (*call*) or sell (*put*) a specified underlying at a fixed price (*strike*) at (*European*) or up to (*American*) a fixed date (*maturity*).

Nowadays, this definition is no longer sufficient. In order to include the "exotics", this definition requires generalization: an option on one or several underlyings with prices given by $S_1, S_2, \ldots S_m$ is characterized by its payoff profile. The *payoff* of the option is a function $F(S_1, S_2, \ldots S_m)$ of the underlying prices and one or more points in time and indicates the cash flows arising for the option's holder upon exercise.[1] American options can be exercised at any time during the lifetime of the option in contrast to the European option, which can only be exercised at maturity. There are also options which can be exercised at several specific times during their lifetime. These are called *Bermudan options* or *Atlantic options* since, in a sense, they are an intermediate form between American (exercise is possible at any time during the option's lifetime) and European (exercise is possible at a *single* time, namely at maturity) options.

---

[1]For example, *plain vanilla* European calls and puts with maturity T and strike K on an underlying with a price S are uniquely defined by the payoff

$$c_S^{\text{plain}}(T, T, K) = \max\{0, S(T) - K\}, \quad p_S^{\text{plain}}(T, T, K) = \max\{0, K - S(T)\}.$$

Furthermore, options can be separated in *strategy dependent options* and *strategy independent options*. A strategy independent option does not require an active decision whether the option should be exercised or not, since the optimal exercise time could be easily determined. At expiry, it is always optimal to exercise an European Plain Vanilla option if the option is in the money, i.e. its present value at exercise is positive. E.g., options embedded in exchange traded certificates will usually be exercised automatically, if it is to the advantage of the option holder, without any interference of the holder. For some options, it is less simple to decide whether exercise is optimal, e.g. if the underlying is more complex. A simple example for a strategy dependent option is a Bermudan swaption. Here, the option holder needs to decide at which (if any) of the agreed exercise dates the option should be exercised. In case of exercise, the option becomes worthless. If the option is not exercised, the option continuous to exist for the remaining term to maturity, which need to be taken into account in pricing the contract.

18.2 Conventions

In this chapter, the pricing of options on different underlying types using the Black-Scholes model and the Black-76 model is presented. The values listed in Table 18.1 will repeatedly be used in the following examples.

For simplicity's sake, we adopt the *Act/365* day count convention, for which $T - t$ is equal to the number of calendar days between T and t divided by 365. Furthermore, the examples are calculated using the continuous compounding convention and constant interest rate, with the following implication holding for the discount factor $B(t, T)$:

$$B(t, T) = e^{-r(T-t)}$$

$$\Rightarrow \quad \frac{\partial B(t, T)}{\partial t} = r B(t, T) \,, \quad \frac{\partial B(t, T)}{\partial r} = -(T - t) B(t, T) \,.$$

Table 18.1 The input parameters for all of the following examples

Risk neutral refinancing rate	r	3.20%
Continuous dividend yield	q	4.80%
Volatility	σ	12.50%
Spot price	$S(t)$	$80.00
Valuation date	t	7. Jan. '01
Maturity of Forward/Futures	T'	7. Apr. '01
Expiration date of option	T	8. Mrz. '01
Strike price of option	K	$79.00

All of the following examples have been extracted from the Excel workbook PLAINVANILLA.xlsx of the download section [50] accompanying this book.

18.3 Options on Spot and Forward Prices

18.3.1 European Options

Liquidly traded European plain vanilla options are usually priced on the market with the Black-Scholes model or more precise: quoted. In fact, there is no pricing model required for liquidly traded options: the price is just the result of offer and demand. However, it turned out to be convenient to quote options in terms of its Black-Scholes volatility, rather than quoting the price directly. The volatility, which hast to be input into the Black-Scholes formula to get back a given price is called the *implicit (Black-Scholes) volatility*. Usually, the volatility changes less frequently than other input parameters of the Black-Scholes formula, e.g. the underlying price or the interest rate curve. Quoting the volatility instead of the price, the quotes need to be updated less frequently. In general, in can be observed that quoted implicit volatilities depend significantly on the strike price of the option (that is the *volatility smile* or *volatility skew*, depending on the shape of the dependence). Since the strike is a property of the option, but the volatility is a property of the stochastic process, the existence of a smile contradicts the model assumptions. However, this strike dependence can be interpreted as a correction to the Black-Scholes model, which describes the underlying process only approximately correct, taking into account that the underlying actually does not follow a strict log-normal process as is assumed by the Black-Scholes model (see also Sect. 30.1).

To use the Black-Scholes model, Eqs. 8.7 for spot options and 8.8 for options on futures are applied. This is demonstrated in Table 18.2. The forward prices themselves are determined using the arbitrage Eq. 6.1. They are also presented in Table 18.2 among the "Interim Results". The values of the futures and forward positions, each with a delivery price K were calculated from the forward price $S(t, T')$ using Eqs. 6.5 and 6.6. The sensitivities for options on the spots and futures as given in Table 12.2 are included in Table 18.2 as well.

Sensitivities like vega, rho, and rho$_q$ are commonly given in %, i.e. as derivative of the price with respect to the risk factors volatility σ, interest rate r, or dividend yield q divided by hundred. The indicate the price change

Table 18.2 Valuation of European plain vanilla spot and futures options using the Black-Scholes model

Interim results				Spot option		Futures option	
				Call	Put	Call	Put
$T - t$	0.1644	$T' - t$	0.2466				
$B(t,T)$	0.9948	$B(t,T')$	0.9921				
$B_q(t,T)$	0.9921	$B_q(t,T')$	0.9882				
$S(t,T)$	$79.790	$S(t,T')$	$79.685				
$\sigma(T-t)$	0.0507	$\sigma(T'-t)$	0.0621				
x	0.2216	x'	0.1701				
$N(x)$	0.5877	$N(x')$	0.5675				
		Price		$2.020	$1.234	$1.959	$1.277
		Δ		0.5831	−0.4091	0.5745	−0.4202
		Γ		0.0953	0.0953	0.0964	0.0964
		Ω		23.0905	−26.5093	23.3657	−26.2009
		Vega		0.1253	0.1253	0.1258	0.1258
		Θ		−0.0108	−0.0144	−0.0129	−0.0130
		ρ		0.0734	−0.0558	0.0720	−0.0571
		ρ_q		−0.0767	0.0538	−0.0753	0.0550
Futures' PV			$0.685				
Forward's PV			$0.680				

if the according risk factor changes by 1%. Theta as derivative with respect to the term to expiry is divided by 365 and equals the price change, if the option is priced one day (not one year) later (if all the other parameters remain unchanged).

The price of the futures option was calculated using Eq. 8.8. In order to calculate the sensitivities, an additional procedure was implemented: as mentioned in the vicinity of Eqs. 8.8 and 12.11, an option on a futures contract is the same as an option on the spot price of an underlying whose dividend yield is exactly the risk-free rate ($q = r$), and whose spot price corresponds to the forward price $S(t, T')$. Therefore, the same procedure as used for the spot option was applied again for the futures option but with the appropriately modified input parameters.

Gamma and delta of a spot option refer to the spot price of the underlying. In the case of the futures option, on the other hand, the forward price is the reference price. This means, for example, that the delta of the futures option indicates how many *futures* are required to (delta) hedge the option, whereas the delta of the spot option gives the number of *underlyings* required for the hedge. The required number of instruments for other combinations can be calculated using Eq. 12.15 and Table 12.1 with either of these deltas. For example, to hedge the spot option requires either 0.5831 underlyings or $0.5831 B_r(t, T')/B_q(t, T') = 0.5854$ futures.

The price of options on the corresponding forward contract can be calculated with Eq. 8.9. It yields a price of \$1.954 for the call and \$1.274 for the put. Options on forward contracts are not traded nearly as liquidly as options on futures.

Of course, the binomial model can be applied for the valuation as well. The appropriate payoff profile is to be implemented at the end of the tree but otherwise the procedure for options on a spot price, a future or a forward is always the same. For options on the spot price, the price can be determined directly from Eq. 9.19 using the binomial distributions. The parameters u, d are calculated using Eq. 9.31. The values for the probabilities p and $\hat{p}$ follow directly from Eq. 9.9. The results of the individual steps are displayed in Table 18.3. The price of the put was determined directly from the price of the call with the help of the put-call parity, Eq. 6.8.

Table 18.3 Valuation of European plain vanilla options using the binomial distribution

Interim results				Option price	
$dt = (T - t)/n$	0.001644	u	1.005080895	Call	$2.02
B	0.999947	d	0.994945	Put	$1.24
B_q	0.999921	p	0.496138209		
$T - t$	0.164384	p'	0.49867215		
$B(t, T)$	0.994753537	y	49		
$B_q(t, T)$	0.992140637	$Bp(j >= y)$	0.587977672		
		$Bp'(j >= y)$	0.607630103		

18.3.2 American Options

The binomial model is very effective in calculating American options. The following example will serve to demonstrate the valuation of both options and futures. The values for the futures and forward positions, each with a delivery price K listed among the final results, are computed from the forward price $S(t, T')$ making use of Eqs. 6.5 and 6.6.

First, a tree for the underlying price must be constructed. In Fig. 18.1, we use the method prescribed by Eq. 9.31 for the calculation of the up and down parameters, in the same way as was already done in Table 18.3. The calculation is subsequently performed backwards through the entire tree:

- Time $t + 3dt = T'$

  Once the tree for the underlying has been constructed, the values of the derivative on the boundary of the tree should be inserted if these are known. In our example, these are the forward prices at the maturity of the future T'. These are just the spot prices effective at time T'.

- Time $t + 2dt = T$

  This time is a maturity date as well, but for the options. Thus, at this point the prices of the *spot option* at maturity are set equal to its payment profile. The *future prices* at each node are determined using Eq. 9.13 from the prices in the next time step. Once the future prices at the maturity of the option on the future are known, the price of the *futures option* can be set equal to its payoff profile.

- Time $t + dt$

  The *futures prices* at each node are computed using Eq. 9.13 from the prices holding for the next time step. The prices of both the *spot* as well as the *futures* option at each node are, in accordance with Eq. 9.8, computed from the prices in the next time step. The right to early exercise is taken into
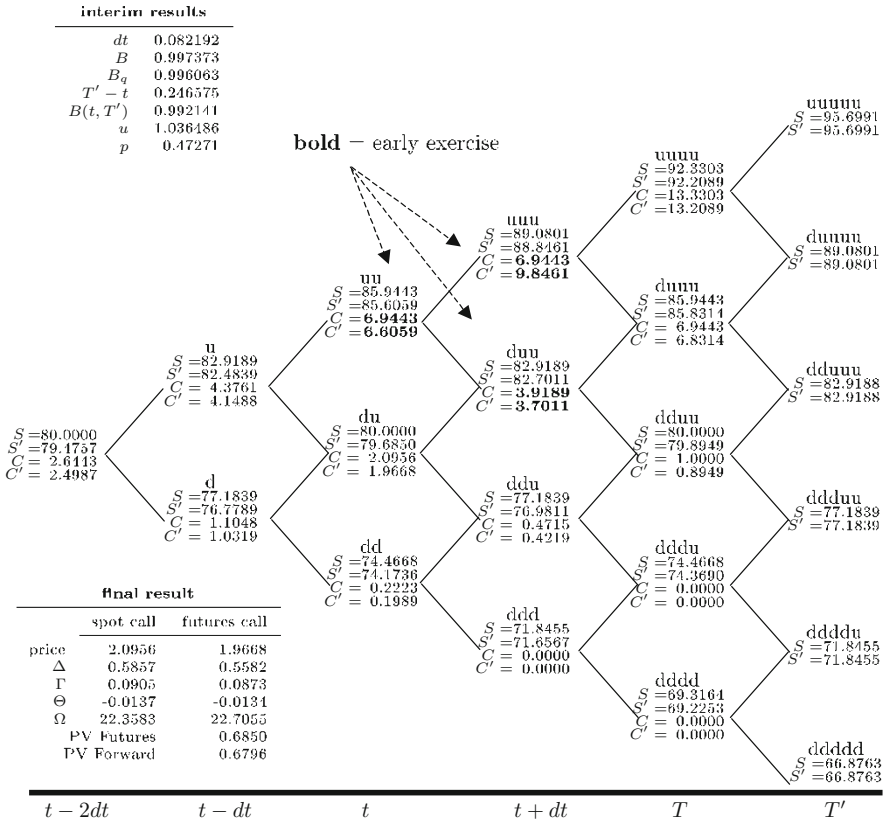
Fig. 18.1   Valuation of an American call c on the spot price S, and of an American call c' on the forward price S' using a binomial tree. The tree for the options comprises two time steps into the future. The tree for the forward price consists of three time steps into the future

account by comparing the results just obtained to the intrinsic values $S - K$ and $S' - K$ (S' denotes the forward price in this example) taking in each case the greater of the two values as the corresponding option price.

- Times $t, t - dt, t - 2dt$
  The same procedure is carried out as that described for time $t + dt$.

Once the entire tree is constructed, prices at three different nodes are available for use at time t, from which the middle node (the **du** node) contains the desired prices for the futures contract, the option on the spot price and the option on the futures contract.

Sensitivities are calculated as difference quotients. For example, the tree could be calculated by constructing a new tree with a slightly changed interest rate $r + h$ (with $h > 0$). Then, we get for the interest rate sensitivity rho:

$$\text{rho} = \frac{V(r + h, t) - V(r, t)}{(r + h) - r} = \frac{V(r + h, t) - V(r, t)}{h} .$$

This procedure could be applied for all sensitivities, though it is numerically inefficient since it is required to construct a new tree. It is also unstable since the parameter shift could cause jumps in the price of $V(t)$. Therefore, delta and gamma are calculated on the existing nodes of the tree according to Eq. 12.27, using the nodes **uu** and **dd** at time t. This requires to let the tree start at time $t - 2dt$ rather then t. This is not required for the calculation of other sensitivities. The values for delta, gamma and omega given here for both options are related to the underlying spot price (not the futures price). For Vega, using a constant volatility, it is also possible to construct a tree which uses the same nodes, but which as an extra time step (i.e. the tree is compressed, which changes the volatility slightly.)

18.4 Index Options and Index Futures

Index futures and options on an index or on an index future correspond to the general case treated in Sect. 8.3, where the spot price S represents the state of the index and q the average dividend yield of the index's component stocks (*as long as these are not directly reinvested in the index!*). Examples are the (European) options on the DAX index traded on the EUREX, known as *ODAX* and the (American) Options on DAX-futures, known as *OFDX*. Table 18.3 and Fig. 18.1 again serve as examples of the valuation of index futures, index spot options and index futures options, where the dividend yield q is the average dividend yield of the index and the underlying price S is the state of the index.

The German *DAX* is a *performance index*, which means that the dividends of the component stocks are immediately reinvested in the DAX, so that in principle, no decline in price is observed when a dividend payment is made and consequently, q should be set equal to zero. The reason why q is not exactly zero in practice relates to the effects of corporate taxation. If the risk-free rate r is known, the DAX together with the DAX future can be used in Eq. 6.2 to find the *implied dividend yield* which can then be used as q when calculating option prices.

18.5 Foreign Exchange Options

Trading of *Foreign exchange* or brief *FX Options* belongs to the most liquid option markets at all. Partly, this is due to the omnipresence of FX risks in an environment of global trades. In addition, the majority of traded FX options is limited to the most important major currencies, i.e. the number of different underlyings is quite limited, e.g., compared to the equity market. Certain kinds of exotic FX options , for example many variants featuring barrier components (see for example [36, 195]) are also traded quite frequently.

FX options belong also to the general case treated in Sect. 8.3, where the spot price S is replaced by the FX (spot) rate and q is replaced by the risk-free interest rate r_f in foreign currency. The logic behind this choice is the following: the holder of the underlying (i.e., the foreign currency) receives, in contrast to the holder of the option or the future, interest payments at the interest rate valid for the foreign currency much the same as the holder of a stock receives a dividend yield q of the stock. Again Table 18.3 and Fig. 18.1 serve as examples for the valuation of foreign exchange futures, spot options and futures options, where here the dividend yield q is equal to the risk-free rate of the foreign currency and the underlying price S is merely the foreign exchange rate expressed in terms of the domestic currency. At an exchange rate of 1.15 euros to the US$, for example, the underlying price would be $S(t) = 1.15$.

18.5.1 Put-Call-Equivalence for FX Options

An interesting relation exists between foreign exchange puts and calls, which follows immediately from the nature of these transactions: The right to *buy* 1 USD for K EUR is obviously the same as the right to *sell* K EUR for 1 USD. The EUR/USD exchange rate indicates how many euros have to be paid per dollar. The USD/EUR exchange rate, indicating how many dollars are to be paid per euro, is naturally $1/S(t)$. A *call* on the EUR/USD exchange rate with a strike price K is thus the same as K *puts* on the USD/EUR exchange rate with a strike price $1/K$. This can also be seen by considering the payoff profile. The payoff in dollars of K puts on the euro is

$$K \max\left\{0, \frac{1}{K} - \frac{1}{S(T)}\right\} = \max\left\{0, 1 - \frac{K}{S(T)}\right\} = \frac{1}{S(T)} \max\{0, S(T) - K\} \ .$$

Here, all terms are expressed in USD units of currency since dollars are received for selling euros. The payoff of a *call* on USD, however, is more naturally expressed in terms of EUR since euros are paid when buying dollars. Before the payoffs can be compared, they both need to be expressed in terms of the same currency. The conversion of the above payoff into EUR can be performed by simply multiplying with the EUR/USD exchange rate $S(T)$. This simply yields the payoff profile of the call max $\{0, S(T) - K\}$ on the right-hand side. If the payoff profiles agree, the options must have the same value for all earlier times t. We thus obtain the *put-call equivalence for foreign currency options*

$$c_S(t, T, K) = K \ p_{1/S}(t, T, 1/K) \ . \tag{18.1}$$

18.6 Interest Rate Options

In Part II the valuation of interest rate options by means of term structure models was demonstrated. In the presentation of the following examples, however, plain vanilla interest rate options are priced using the Black-Scholes and Black-76 models still commonly used in the market for the quotation of these types of options.[2]

18.6.1 Options on Bonds

European bond options are usually priced resp. quoted using the Black-76 model, Eq. 8.10. If the Black-76 is used (forward price equals expectation of future price), the model's parameter σ can be set equal to the volatility of the respective underlying's future price.[3]

For options on zero bonds, $S(t, T)$ is the forward price as given by Eq. 15.17, for options on coupon bonds, it is the (dirty) forward price according to Eq. 15.18. As presented in Sect. 15.5.2, calculations made in pricing derivatives on coupon bonds should always be performed with the amounts of money which must actually be paid, i.e., accrued interest should always be included in the computations. Thus, for options on coupon bonds, the *dirty strike* of

---

[2]In markets with negative interest rates like EUR or CHF, these models are increasingly replaced by shifted-lognormal or normal models (i.e. the Bachelier model).

[3]This volatility is also called *price volatility*. Alternatively, bond options could be prices similar to swaptions, such that the interest rate or the interest rate curve is modeled as stochastic underlying rather than the bond price. Then, the respective volatility is called *interest rate volatility*.

the option should be used. This is the amount of money paid by the holder of a call upon exercise in order to purchase the underlying. In general, bond options are quoted for a *clean strike*. The interest accrued up to time T must then be added to this quoted strike. Often though, exercise dates and coupon period start resp. end dates fall together, at which dates accrued interest is zero.

American options on bonds—as all options on underlyings with dividend payments during the lifetime of the option—generate binomial trees which *no longer* recombine after the *ex-dividend day t_D* (=coupon payment date):

$$u\, d S(t) \neq d\, u\, S(t) \ \text{ for } \ t > t_D \ .$$

Therefore, other methods like finite differences (see Chap. 10) are better suited for pricing American options. In practice, Bermudan bond options are traded more frequently traded than American options. Bermudan options could be exercised only on a given set of dates, which typically fall on coupon period start dates. Pricing Bermudan options is as elaborate as pricing American options, except for the fact that accrued interest does not matter, if the exercise dates fall together with coupon start dates.

Bond options are always traded on real rather that synthetic underlyings as is the case for futures contracts. Therefore, bond options are traded less liquidly than corresponding options on bond futures contracts.

This is the reason why options on bonds are a good deal rarer in electronic exchanges than futures options. For example, on the EUREX, options are available on the Bund Future, the Bobl future and the Schatz future. These are all American. However, no options directly on 2, 5 or 10 year government bonds are traded on the EUREX.

18.6.2   Options on Bond Futures

European options on bond futures are priced—as are all plain vanilla European futures options on the market—using the Black-Scholes formula for options on futures given by Eq. 8.8.

The following consequence of the Black-Scholes differential equations 12.5 and 12.11 are useful in the valuation of *American* options on bond futures: an option on a future is equivalent to an option on an underlying with a dividend yield equal to the risk-free rate and a spot price equal to the future's price.

For such an underlying, the (recombinant) binomial tree can be constructed as described in Chap. 9 to price the option. This is demonstrated very explicitly in Fig. 18.1 (the part of the example for the valuation of the option on the

underlying itself is relevant to our discussion here). Thus, Fig. 18.1 can serve as an explicit demonstration of pricing bond futures options if we use the clean forward bond price in place of the "spot price" of the underlying and the risk-free rate for the dividend yield. In contrast to derivatives referring *directly* to a bond (bond options, bond futures), *clean* prices have to be used for options on bond *futures* since the difference between the clean forward price and the clean strike is due when the option is exercised.

18.6.3   Caps and Floors

A *cap* (*floor*) is a series of *caplets* (*floorlets*), arranged in such a way that the periods of the individual caplets (floorlets) completely cover the time span of interest without overlap. The period length of the individual caplets (floorlets) corresponds to the lifetime of the floating interest to be bounded. A cap, for example, of a 3-month LIBOR for five years consists of a series of 19 (the first period is already fixed) caplets each having a period length of three months. To understand caps and floors it is sufficient to understand their component caplets and floorlets.

A *caplet* is an "interest rate limiting agreement" over a time span from the fixing date T until the payment date T'. The principal N of a caplet is compounded over the future time span $T' - T$ at a rate equaling the difference between the (floating) reference rate $R = R(T, T')$ and a fixed strike K *if* this rate difference is positive. The floating reference rate is fixed at the fixing date T. This construction has the effect that the holder of a caplet pays interest (for instance for a loan) on a principal N at a floating rate over a time span from T until T' but at *most* the strike rate K of the caplet.

A caplet is thus an option with the reference rate as the underlying. The option is *exercised* at the fixing date T since it is at this time that the difference between the strike and the underlying is determined. The option *payment* occurs later, namely at the payment date T'. The value of the option upon *exercise* is of course essential to its valuation. In order to obtain the value of the option upon exercise, its payoff profile must be discounted back to the exercise time T:

$$c^{\text{caplet}}(T, T', K|T) = B(T, T')N\tau(T, T') \max\left[R(T, T') - K, 0\right]$$
$$\tag{18.2}$$

$$p^{\text{caplet}}(T, T', K|T) = B(T, T')N\tau(T, T') \max\left[K - R(T, T'), 0\right] \ .$$

From the condition that a payment is made if the strike of the caplet is greater than the underlying, it follows that a caplet is a *call* option on the reference rate. It is for this reason that the letter c was chosen to denote the caplet in Eq. 18.2. The corresponding *put* on the reference rate is the above mentioned *floorlet*, which represents a lower limit ("floor") on the reference rate. The floorlet is denoted by the letter p in Eq. 18.2. Caps and floors are also referred to as *interest rate guarantees* (IRGs).

Since the discount factors are strictly monotone decreasing functions of the interest rate, conditions on the interest rates can be equivalently interpreted as conditions on the discount factors, i.e.

$$R(T, T') > K \Leftrightarrow B(T, T') < \frac{1}{1 + K\tau(T, T')} \,. \tag{18.3}$$

Valuation as Interest Rate Options (Lognormally Distributed Interest Rates)

These payoff profiles allow the option to be interpreted as a plain vanilla option with a nominal $N\tau(T, T')B(T, T')$, a strike price K and maturity T on an underlying with a price $S(T) = R(T, T')$. The forward price of this underlying as seen from today (time t) is then the forward rate $R(T, T'|t)$. The discount factor $B(T, T')$ appearing in the nominal is not known at time t. Market participant use the discount factor of the forward rate $B(T, T'|t)$ instead. It can be shown [145], that this is consistent with a term structure model. The maximum functions in Eq. 18.2 correspond to payoff profiles of plain vanilla options and can therefore be priced with the Black-76 model, Eq. 8.10. The prices of caps and floors are then given by

$$c^{\text{caplet}}(T, T', K|t)$$
$$= N\tau(T, T')\underbrace{B(T, T'|t)B(t, T)}_{B(t, T')}\left[R(T, T'|t)\text{N}(x) - K\text{N}(x - \sigma\sqrt{\tau(t, T)})\right]$$

$$p^{\text{caplet}}(T, T', K|t)$$
$$= N\tau(T, T')\underbrace{B(T, T'|t)B(t, T)}_{B(t, T')}\left[R(T, T'|t)\text{N}(-x) - K\text{N}(-x + \sigma\sqrt{\tau(t, T)})\right]$$

$$x = \frac{\ln\left(\frac{R(T, T'|t)}{K}\right)}{\sigma\sqrt{\tau(t, T)}} + \frac{1}{2}\sigma\sqrt{\tau(t, T)} \,. \tag{18.4}$$

According to Eq. 2.7, the product of the discount factor and the forward discount factor yield precisely the discount factor from time T' back to time t.

As mentioned at the beginning of this section, a cap is nothing other than a *strip* of caplets. The value of the cap is the sum of the values of the individual caplets (with the analogous relation holding between floorlets and floors):

$$c^{\mathrm{cap}}(T, T', K|t) = \sum_{i=1}^{n} c^{\mathrm{caplet}}(T_{i-1}, T_i, K|t) \tag{18.5}$$

$$p^{\mathrm{cap}}(T, T', K|t) = \sum_{i=1}^{n} p^{\mathrm{caplet}}(T_{i-1}, T_i, K|t)$$

$$\text{with} \quad T = T_0 < T_1 < T_2 < \ \ldots \ < T_n = T' \ .$$

As illustrated in Eq. 8.10, the Black-76 model assumes that the underlying is lognormally distributed at the option's exercise time. The volatility *at that time* is given by a value σ. This means we have to assume that the reference rate $R(T_{i-1}, T_i)$ of each caplet is lognormally distributed.

As described previously, the Black-76 model is functioning for caps and floors more like a quotation tool rather than a valuation model. Caps and floors are liquidly traded products, with prices fixed by offer and demand. Often, implied Black volatilities to be inserted in Eq. 18.5 to get the price are quoted instead of quoting the prices directly. Since caps serve as hedges against interest rate risks, this convention is advantageous since the volatility is a better measure for the current risk of interest rate changes than the cap price itself. For cap quotes the same volatility is used for all caplets of the caplet strips, which is referred to as the *flat volatility* because of the missing term structure.

For ATM (at-the-money) caps, various maturities are quoted as different volatilities. Quotes for non-ATM caps are traded less liquidly. Because of the term structure of interest rate curves, the strike of ATM caps will differ for different maturities. As a consequence, even for ATM caps, the caplets of the corresponding caplet strip won't be ATM caplets. Instead, some of the caplets will be ITM (in-the-money) while others will be OTM (out-of-the-money). In the special case of a flat interest rate term structure, this yields to an apparent inconsistency: Because of the different volatility quoted for caps of different maturity, the same caplet, appearing simultaneously in the strips of two different caps, can have different volatilities and thus have two different prices. For example, a caplet that limits the 6-month LIBOR to 2% in two years occurs in both, a five year and a ten year, 2% cap on the 6-month

LIBOR. Since the flat volatilities differ for 5 and 10-year caps, this caplet will have two different prices corresponding to the two different caps. This inconsistency can be solved by determining the underlying caplet volatilities by means of a bootstrapping method that reproduces the cap prices (calculated with flat cap volatilities inserted into Eq. 18.5), but using a single volatility for each individual caplet, independent of the caplet strip it is a part of. With these implied caplet volatilities it is also possible to calculate prices for caps for maturities and strike for which no quotes are available. For each individual caplet, the volatility to be inserted in Eq. 18.5 is determined by means of inter- and extrapolation.

The valuation of caps and floors described above, i.e., the interpretation as options on the interest rate as the underlying, using of the Black-76 model naturally assumes we are in the Black-Scholes world. In particular that the yield of the underlying (=relative change in the interest rate) is normally distributed, which implies that the underlying itself (the interest rate) is *log*normally distributed.

Valuation as Bond Options (Normally Distributed Interest Rates)

It is also possible to price caps and floors by interpreting them as options on *bonds* and performing the valuation using Black-Scholes for options on bond prices. Then the underlying is the bond price which is therefore implicitly assumed to be lognormally distributed. Thus, its relative changes (the bond *yields* which are by definition nothing other than interest rates) are assumed to be *normally* distributed.

Since the reference interest rate index (as well as the strike) is usually given in terms of simple compounding, we have $B(T, T') = 1/[1 + R(T, T')\tau(T, T')]$ or solved for $R(T, T')$:

$$R(T, T') = \frac{1}{\tau(T, T')} \left(\frac{1}{B(T, T')} - 1 \right) .$$

This allows to write down the pay off profiles of Eq. 18.2 as zero bond option pay offs. As example, we get for the caplet (call):

$$B(T, T')N\tau(T, T') \max \left[R(T, T') - K, 0 \right]$$

$$= B(T, T')N\tau(T, T') \max \left[\frac{1}{\tau(T, T')} \left(\frac{1}{B(T, T')} - 1 \right) - K, 0 \right]$$

$$= N \max \left[1 - B(T, T') - B(T, T') \tau(T, T') K, 0 \right]$$

$$= \left[1 + \tau(T, T') K \right] N \max \left[\frac{1}{1 + \tau(T, T') K} - B(T, T'), 0 \right]$$

$$= N_B \max \left[K_B - B(T, T'), 0 \right] .$$

The derivation for floorlets (put) is analogous. Therefore, it is possible to write the pay off profiles as

$$c^{\text{caplet}}(T, T', K | T) = N_B \max \left[K_B - B(T, T'), 0 \right]$$

$$p^{\text{caplet}}(T, T', K | T) = N_B \max \left[B(T, T') - K_B, 0 \right] \qquad (18.6)$$

with the following interpretation:

- The *caplet* is equivalent to a put option with strike K_B and maturity T on a zero bond with nominal $N_B = N \left[1 + K \tau(T, T' - T) \right]$ and maturity T'.
- The *floorlet* is equivalent to a call option with strike K_B and maturity T on a zero bond with nominal $N_B = N \left[1 + K \tau(T, T' - T) \right]$ and maturity T'.

It is no surprise that the strike $K_B = 1 / \left(1 + \tau(T, T') K \right)$ equals the value of the zero bond if $K = R(T, T')$. With the above, caplets and floorlets could be priced as options on zero bonds as in Sect. 18.6.1. A Cap or Floor can be represented as a portfolio of caplets and floorlets, i.e. as a portfolio of puts or calls on corresponding zero bonds.

Table 18.4 shows the correspondence to bond options valuated according to Eq. 8.10.

Table 18.4 Caplets and floorlets as options on zero bonds

Parameter of the bond option	Corresponding parameter for IRGs
Option	Put for caplet, call for floorlet
Underlying	Zero bond with maturity at the payment date T' of the IRG
Face value of the underlying	$N[1 + \tau(T, T')K]$
Forward price $S(t, T)$	Forward price $B(T, T')$ of the zero bond at the fixing date T of the caplet
Volatility	Volatility of the forward price
Strike of the bond option	$K_B = 1/[1 + \tau(T, T')K]$ equals bond price with $R = K$

Collars and Put-Call Parity for Caps and Floors

A collar is a combination of a cap and a floor having the effect that the collar's holder (if he or she also has a loan, for instance) has to pay the reference rate (for the loan) only as long as it remains within a range bounded below by the strike of the floor and above by the strike of the cap. Should the rate exit this range, the payment that the holder of the collar would receive from the cap (or pay into the floor) would compensate for the higher (or lower) interest payment on the loan. A long collar is thus a portfolio consisting of a long cap and a short floor:

$$c^{\text{collar}}(T, T', K_1, K_2|t) = c^{\text{cap}}(T, T', K_1|t) - p^{\text{cap}}(T, T', K_2|t) \ .$$

Investors often buy collars to provide a safeguard against rising interest rates and to lower the price of this safeguard through the sale of the floor. Often, if the strike of the cap is given, the strike of the floor is chosen such that the floor is worth exactly as much as the cap and that thus the total collar can be acquired without cost (called *zero cost collar*).

If we require instead the strikes of the cap and the floor to be the same, i.e., $K_1 = K_2 = K$, then the holder of the collar effectively pays (for a loan, for example) exactly the strike K instead of the reference rate. This is the same as though he had a forward swap from T to T'. Thus, caps and floors satisfy the following put-call parity.

> If the cap and the floor of a collar have the same strike, the value of the collar is equal to the value of a forward swap over the lifetime of the collar with a fixed rate equal to the strike of the collar.

Or in brief

> Cap minus floor equals forward swap

By definition, the forward swap rate is the coupon rate for the fixed side which makes the swap worthless. The put-call parity now states:

> If a collar is worthless and if its cap and floor both have the same strike, the strike of the collar equals the forward swap rate for the collar's lifetime.

In Table 18.5, a cap, a floor and a collar with term to maturity of 10 years and start in 3 years is priced according to Eqs. 18.4 und Gl. 18.5. The interest rate curve and forward vols are also given in the table. The forward rates have already been calculated in Fig. 18.1. The floor strike (3.24%) was modified (through numerical iteration) until the collar for a given cap of 7% was worthless. Setting the floor strike equal to the cap strike and solving for the strike price again, to make the collar worthless does in fact yield a forward swap rate of 5.28% for the same period.

18.6.4   Swaptions

A *swaption* or swap option is the right to enter into a swap upon maturity of the option, without incurring any further payment. The interest rate K on the fixed side of this underlying swap is the strike price of the option. Depending on whether the swap is a payer or receiver swap, the swaption is referred to as a *payer* or *receiver* swaption, respectively. As opposed to caps and floors, which can be interpreted as options on the floating side of a swap, swaptions are options on the *total* swap. The lifetime of the swap begins at the maturity T of the option. Hence, the swap under consideration is a forward swap. During the lifetime of a forward swap up to its maturity T', a number n of payment dates occur at times T_i, where T_0 is defined as the start of the first coupon period:

$$t < T \leq T_0 < T_i < T_{i+1} < T_n = T' \ \forall \ i = 1, \ldots, n-1 \ .$$

Usually $T = T_0$, which we will set equal in the following therefore. The fixing date $F(T)$ for the swap rate starting at T and terminating at T' lies typically 2 value days before the start date. The exercise date T_{ex}, i.e. the last day until the option buyer requires to notify the option seller that he likes to exercise the option, is usually a bit earlier, typical are 5 or 10 business days before option expiry T.

The payoff profile of a swaption, i.e., its value at maturity T can be determined by considering the following. If we enter into an underlying swap at time T as a consequence of exercising a receiver swaption, then—at each payment date for the entire lifetime of the swap—we receive interest on the swap's principal compounded at K in return for making interest payments on the swap's principal at the floating reference rate. If at time T we enter simultaneously into a *par* payer swap (which by definition of a par swap doesn't incur any additional costs) with the same principal and the same floating side

Table 18.5 Interest rate options on the 12-month floating rate. The values for the cap, floor and collar in $ result from the values in % by multiplication with the principal of $250,000

| | Spot | | 1 Y forward | | | Interest rate options | | | | Principal 250,000 | |
| | | | | | | Cap 700% | | Floor 324% | | Collar | |
Term	Spot rate	Zero bond	Start in years	Forward rate	Forward vol σ	Caplets in %	Caplets in $	Floorlets in %	Floorlets in $	in %	in $
1	2.20%	0.97847	0	2.20%	18.00%						
2	2.50%	0.95181	1	2.80%	18.20%						
3	2.90%	0.91781	2	3.70%	18.40%						
4	3.30%	0.87821	3	4.51%	18.60%	0.062%	156	0.725%	1,813	−0.663%	−1 657
5	3.70%	0.83389	4	5.32%	18.80%	0.256%	640	0.463%	1,158	−0.207%	−518
6	4.00%	0.79031	5	5.51%	19.00%	0.370%	924	0.515%	1,288	−0.146%	−364
7	4.30%	0.74475	6	6.12%	19.20%	0.616%	1,540	0.433%	1,082	0.183%	458
8	4.50%	0.70319	7	5.91%	19.40%	0.585%	1,462	0.569%	1,422	0.016%	40
9	4.70%	0.66142	8	6.31%	19.60%	0.751%	1,877	0.531%	1,328	0.220%	549
10	4.90%	0.61979	9	6.72%	19.80%	0.907%	2,269	0.499%	1,248	0.408%	1,021
11	5.00%	0.58468	10	6.01%	20.00%	0.679%	1,698	0.699%	1,746	−0.019%	−48
12	5.10%	0.55051	11	6.21%	20.20%	0.750%	1,875	0.685%	1,712	0.065%	163
13	5.20%	0.51736	12	6.41%	20.40%	0.812%	2,029	0.669%	1,673	0.143%	357
14	5.25%	0.48853	13	5.90%	20.60%						
15	5.30%	0.46087	14	6.00%	20.80%						
Total value cap, floor, collar:						5.79%	14,470	5.79%	14,470	0.00%	0

as the underlying swap, we pay interest at the swap rate $K_S(T, T')$ valid at this time on the fixed side and receive the floating side. The cash flows on the floating sides of both swap positions (the underlying of the option and the par swap) cancel each other out, and the remaining net cash flows are equal to the difference of the fixed sides. The present value of this difference at time T can be conveniently expressed in terms of the annuity (Eq. 15.7):

$$N \left[K - K_S(T, T') \right] A(T, T, T') .$$

Exercise of a receiver swaption combined with a par swap obtained at no cost thus generates with certainty (i.e., without any market risk) the above cash flows over the entire lifetime of the underlying. Naturally, the receiver swaption will only be exercised if the cash flows thus generated are positive, i.e., if $K > K_S$. It follows for the pay-off profiles of the payer swaption (equivalent to a call on a payer swap) and the receiver swaption (equivalent to a put on a payer swap):

$$c^{\text{swap}}(T, T', K|T) = N A(T, T, T') \max \left[K_S(T, T') - K, 0 \right]$$

$$p^{\text{swap}}(T, T', K|T) = N A(T, T, T') \max \left[K - K_S(T, T'), 0 \right] . \quad (18.7)$$

Valuation as Interest Rate Options (Lognormally Distributed Interest Rates)

It can be seen from the pay off profiles in Eq. 18.7 that the *swap rate* $K_S(T, T')$ is the underlying of a swaption. The maximum functions on the right-hand side can be interpreted as pay off profiles of plain vanilla options with strike price K and expiry date T on an underlying with price $S(T) = K_S(T, T')$. Then, the forward price of this underlying is the forward swap rate $K_S(T, T'|t)$.

The fact that the swaptions' payments do not occur at exercise time T but at the payment dates T_i is already taken into account by multiplication with the annuity $A(T, T, T')$. Alternatively, the option could be settled immediately with a compensation payment equal to the present value of the swap at time T, which is called *cash settlement* instead of *physical settlement*. In case of cash settlement, we need a rule to calculate the compensation payment amount, since the fair value depends on model assumptions and is therefore not properly defined in contrast to a fixed payment. This is solved by the mutual agreement to use the (quoted) swap rate for discounting all future cash

flows to calculate the fair value. This is not an exactly arbitrage free calculation, the prices between cash and physically settled swaptions will differ slightly to compensate for this difference.

The fair values of the plain vanilla options at time t can be valued in the same way as caplet and floorlets with the Black-76 model Eq. 8.10, if the distribution of swap rates at time T is assumed to be lognormal. Note that here, in addition to $K_S(T, T')$, the multiplication factor, the annuity $A(T, T, T')$, does also depend on the future interest rate curve. If we choose the annuity as the numeraire as in Sect. 15.8.3, we get with Eq. 13.1 the following expressions for the option's fair values at time t:

$$\frac{c^{\text{swap}}(T, T', K | t)}{A(t, T, T')} = \text{E}_t^A \left[\frac{N A(T, T, T') \max \left[K_S(T, T') - K, 0 \right]}{A(T, T, T')} \right]$$

$$= N \text{E}_t^A \left[\max \left(K_S(T, T') - K, 0 \right) \right]$$

$$\frac{p^{\text{swap}}(T, T', K | t)}{A(t, T, T')} = N \text{E}_t^A \left[\max \left(K - K_S(T, T'), 0 \right) \right] .$$

Here, the par swap rate $K_S(T, T')$ is the only stochastic term under the expectation operator. The swaption prices in the Black-76 model can now be written as:

$$c^{\text{swap}}(T, T', K | t) = N A(t, T, T') \left[K_S(T, T' | t) \text{N}(x) - K \text{N}(x - \sigma \sqrt{T - t}) \right]$$

$$p^{\text{swap}}(T, T', K | t) = N A(t, T, T') \left[-K_S(T, T' | t) \text{N}(-x) + K \text{N}(-x + \sigma \sqrt{T - t}) \right]$$

$$\text{with} \quad x = \frac{\ln \left(K_S(T, T' | t) / K \right)}{\sigma \sqrt{T - t}} + \frac{1}{2} \sigma \sqrt{T - t} . \tag{18.8}$$

In Table 18.6, the valuation of swaptions using Eq. 18.8 is demonstrated for two different strikes on the forward swap in Table 15.6. The swaptions each have the same lifetime (and the same interest rate curves were used for the valuation) as do the caps, floors and collars in Table 18.5. The underlying, namely the forward swap rate, was calculated in Table 15.7. For this forward swap rate, we assume a volatility of 12.5%. The sum of the spot rate discount factors from Table 15.6 with lifetimes from 4 to 13 years yields 6.884114041. This is the factor $N A(t, T, T')$ to be used in Eq. 18.8. The prices of at-the-money swaptions (i.e., with strike equal to the forward swap rate of 5.82%) are the same for payer (call) and receiver (put) swaptions. This is a special case of the *put-call parity for swaptions* (see Sect. 18.6.4).

Table 18.6 Pricing swaptions. The underlying is a swap fix vs. 12-months floating over ten years, beginning in three years

	Swaptions			
	5.82%	Forward swap rate		
	12.50%	Forward volatility		
	$250 000	Nominal		
	7.00% strike		5.82% strike	
	Call	Put	Call	Put
Value in %	1.03%	9.18%	3.45%	3.45%
Value in $	$2 582	$22 942	$8 630	$8 630

Similar to caps/floors, liquidly traded swaptions are usually quoted as Black-76 volatilities instead of as prices.[4] The quoted vols depend on the underlying swap rate maturity, the option expiry and the strike price, forming a three-dimensional volatility cube. Depending on the market, the quoted vols belong to the most frequently traded swap type, e.g. in the EUR region, these are swaps exchanging semi-annual payments based on 6M-Euribor on the variable leg against annual fixed payments on the fixed leg. In the US-market, swaps exchange quarterly variable payments against semi-annual fixed payments are more frequent. Due to slightly different default risks caused by the different payment frequencies, quoted vols and swap rates for different swap types differ slightly, too. Depending on the use case of these market data, this observed difference might be relevant.

Quoted swaption volatilities show a dependency on the strike which is called *smile* or *skew* (which describes better the actual functional form of the dependency which is different form equity or FX smiles). As mentioned earlier, this skew can't be explained within the Black-76 model. There is a further, more subtle inconsistency. Par swap rates could be represented as a weighted sum of the forward rates for each single interest rate period of the swap. Usually, both rates, par swap rates as well as forward rates, are assumed to be lognormally distributed, if the Black-76 model is used for caplet/floorlet prices as well as for swaption prices. On the other hand, it is well known that the sum of lognormally distributed random numbers itself can't be lognormally distributed. Therefore, it is inconsistent to assume lognormallity for forward rates and par swap rates at the same time.[5] However, both, the strike dependency and the inconsistency of the distribution assumptions for

---

[4]In markets with negative interest rates, quoting of normal vols (i.e. implied Bachelier model vols), shifted lognormal vols or prices have become more common.

[5]In models with normally distributed risk factors, this inconsistency is not present, since the sum of normally distributed random numbers is itself normally distributed.

forward and par swap rates, does not cause problems, if the Black-76 model is not used as a pricing model, but just as a convenient quotation vehicle. For simple use cases, Black-76 is a reasonable useful approximation for modeling the dependencies, as long as one is aware of the limits of the model.

Valuation as Bond Options (Normally Distributed Interest Rates)

To enter into a receiver swap is the same as buying a bond whose coupons and coupon payment dates are given by the fixed side of the swap and in return selling a floater whose floating rate and payment dates are defined by the floating side of the swap. A receiver swaption can thus be interpreted as an option on a bond with a coupon K, whose payment dates and lifetime are defined by the fixed side of the underlying swap. The strike of this bond option is the amount payable for the bond at exercise. This is given by the value at time of exercise T of the floater defined by the floating side of the underlying swap. In general, the value at time T of a floater is not known at time t and as a result we are forced to price a bond option whose strike is unknown. The swaption's underlying swap, however, is usually defined so that it begins at the start of a complete period. This means that the floater is to be priced as if it where just fixed at time T. It's present value, according to Eq. 15.21 is then identically equal to its face value, if basis spread are neglected. With this interpretation, swaptions are options on coupon bonds with strike equal to the principal and therefore can be priced in the same way as was done in Sect. 18.6.1. The correspondence to bond options is summarized in Table 18.7.

Table 18.7 Swaptions as options on coupon bonds

Parameter of the bond option	Corresponding parameter for swaption
Option type	Call for receiver swaption, put for payer swaption
Underlying	Bond with coupon equal to the strike K of the swaption, principal equal to the principal of the swap, maturity and coupon payment dates as on the fixed side of the swap
Forward price $S(t, T)$	Dirty forward price (including accrued interest) of the bond at maturity T of the swaption
Volatility σ	Volatility of this forward price
Strike K of the option	Principal of the swap

Put-Call Parity for Swaptions

In a portfolio consisting of a *short* payer swaption and a *long* receiver swaption with the same underlying lifetime from T to T' and the same strike K, one of the two swaptions will *definitely* be exercised at maturity T. If the current swap rate is greater than the strike K at time T, the receiver swaption expires worthless but the counterparty exercises the payer swaption and, as a result of the short position, a receiver swap is obtained with the counterparty paying the fixed rate of K. On the other hand, if the current swap rate at time T is less than K, the counterparty allows the payer swaption to expire while the receiver swaption is exercised, and consequently a receiver swap with the counter party paying the fixed rate K enters the portfolio. Thus, in any case the portfolio at maturity results in a receiver swap at the fixed rate K. The portfolio must then have the same value as a forward receiver swap for all earlier times $t < T$ as well, since otherwise arbitrage would be possible. This is the put-call parity for swaptions. Exactly the same argument can be used to show that a portfolio consisting of a long payer swaption and short receiver swaption has the same value as a forward payer swap.

> If a payer and a receiver swaption have the same strikes and the same lifetime then the following holds:
>   Short payer swaption and long receiver swaption equals long forward receiver swap.
>   Long payer swaption and short receiver swaption equals long forward payer swap.

If the common strike of payer and receiver swaptions are equal to that of the forward swap rate holding for the corresponding lifetime, then the value of the portfolio must be zero since the corresponding forward swap is worthless. This means that the receiver and payer swaptions with the forward swap rate as their strike have the same value. This was demonstrated in Table 18.6.

Bermudan Swaptions

Bermudan swaptions can only be exercised at certain dates, similar to Bermudan bond options. In general, these exercise dates are equal to a coupon period start date. Bermudan swaptions as well need to be valuated by means of an interest rate term structure model. A lower boundary for the fair value of

a Bermudan swaption can be determined by calculating for each potential exercise date the fair value of the equivalent European swaption (the co-terminal swaption) and than take the maximum of all European co-terminal swaption fair values.

19

Exotic Options

19.1 Payoff Profiles for Selected Exotics

In the world of exotic options payoffs are significantly more complicated than those of plain vanilla options. Selected examples are indicated below. Some more examples will be discussed in the next sections. These kinds of instruments can in principle all be priced using the methods introduced in Part II (for example, finite difference methods or Monte Carlo simulations). In the following, K always denotes the strike, T the maturity of the option and S the price of the underlying.

19.1.1 Power Options

Power options are exotics with payoff profiles similar to those of plain vanilla options, but which involve powers of the underlying, of the strike, or of their difference[1]:

Asymmetric power call:	$\max\{0, (S(T))^n - K^n\}$ mit $n > 0$
Asymmetric power put:	$\max\{0, K^n - (S(T))^n\}$ mit $n > 0$
Symmetric power call:	0 falls $S(T) \leq K$, sonst $[S(T) - K]^n$ mit $n > 0$
Symmetric power put:	0 falls $S(T) \geq K$, sonst $[K - S(T)]^n$ mit $n > 0$
Self quanto call	$S(T) \max\{0, S(T) - K\}$
Self quanto put	$S(T) \max\{0, K - S(T)\}$.

---

[1]Options with (electric) power as underlyings are also called power options

Self quanto options, for example, are traded on FX rates with foreign currencies, which could not be traded freely, and, therefore, the pay-off could not be paid in foreign currency.

19.1.2   Cliquet and Coupe Options

A *cliquet option* settles periodically and resets the strike at the spot level then valid. A cliquet can therefore be thought of as a series of "pre-purchased" at-the-money options, but where the total premium is determined in advance. The payout on each option can either be paid at the final maturity, or at the end of each reset period (in the latter case the payoff at the reset times is compounded until final maturity). For one reset at a previously fixed time t' with $t < t' < T$, the payoff at the final maturity T is given as:

Cliquet call:	$B(t', T)^{-1} \max\{0, S(t') - K\} + \max\{0, S(T) - S(t')\}$
Cliquet put:	$B(t', T)^{-1} \max\{0, K - S(t')\} + \max\{0, S(t') - S(T)\}$.

Here the payoff at time t' is compounded until the final maturity T. The number of reset periods is specified in the contract in advance. More resets make the option more expensive. A cliquet option is always more expensive than the corresponding plain vanilla at-the-money option with the same final maturity.

A *coupe option* settles periodically and resets the strike at the worst (from the option holder's perspective) of either the spot level then valid or the original strike set for the first period. It is a series of options where the total premium is determined in advance. The payout on each option can be paid at final maturity, or paid at the end of each reset period. For one reset at a previously fixed time t' with $t < t' < T$, the payoff is given as:

Coupe call:	$B(t', T)^{-1} \max\{0, S(t') - K\} + \max\{0, S(T) - \max\{S(t'), K\}\}$
Coupe put:	$B(t', T)^{-1} \max\{0, K - S(t')\} + \max\{0, \min\{S(t'), K\} - S(T)\}$.

Here the payoff at time t' is compounded until the final maturity T. The number of reset periods is determined by the buyer in advance. More reset periods make the option more expensive, but because of the "worst of" feature a coupe is always cheaper than the corresponding cliquet.

19.1.3 Look-Back Options

There are in principle two types of *look-back options* also known as *hindsight options*: Either the strike is set at the start and at maturity, the buyer can "look back" over the life of the option and choose the most favorable underlying price achieved during the option's life time to maximize profit between strike and exercise. Or the strike is set at maturity. This means at maturity the buyer can "look back" and choose the most favorable underlying price achieved during the options life *as the option strike* to maximize profit between this strike and the underlying price at maturity. The look-back feature is thus very attractive to investors as it gives the buyer the best possible payout. However, look-back options are generally very expensive. The most common payoff profiles are

Minimum-strike call:	$S(T) - \min\{S(t')\}$
Maximum-strike put:	$\max\{S(t')\} - S(T)$
Maximum-asset call:	$\max\{0, \ \max\{S(t')\} - K\}$
Minimum-asset put:	$\max\{0, K - \ \min\{S(t')\}\}$
Look-back-spread call:	$\max\{0, \max\{S(t')\} - \ \min\{S(t')\} - K\}$
Look-back-spread put:	$\max\{0, K - \ \max\{(S(t')\} + \ \min\{S(t')\}\}$
Look-back-straddle:	$\max\{S(t')\} - \min\{S(t')\}$,

where each of the maxima and minima are taken over the entire lifetime of the option, i.e., for t' such that $t \leq t' \leq T$.

19.1.4 Asian Options

An *average strike Asian option* is an option where the strike rate equals the average of the underlying over the life of the option. The strike rate can therefore only be calculated at maturity of the option. Instead of using the average underlying price as the strike, it can also be used as the price which is to be compared to a fixed strike in the payoff. These kinds of Asian options are then called *average price options*. The payoff profiles are explicitly

Average-price call:	$\max\{\ 0, \ \text{average}[S(t')] - K\ \}$
Average-price put:	$\max\{\ 0, K - \ \text{average}[S(t')]\ \}$
Average-strike call:	$\max\{\ 0, S(T) - \ \text{average}[S(t')]\ \}$
Average-strike put:	$\max\{\ 0, \ \text{average}[S(t')] - S(T)\ \}$.

The function "average[ ]" usually represents either the geometric or arithmetic mean, depending on the option's definition. In fact, an Asian option can use any agreed upon method for averaging (e.g., with different weights,

etc.). The averaging could be taken over the entire lifetime of the option, i.e. $t \leq t' \leq T$, applying daily, weekly, monthly or quarterly observed underlying prices from a specific data source (e.g., a specified page from Bloomberg, Reuters or Telerate). Often, the averaging is applied only for a short period, for example over daily end-of-day prices for the last week before option expiry (this is called *Asian tail*). This could reduce the exposure to short-term extreme price fluctuations. All these details need to be specified in the contract.

19.1.5 Rainbow and Exchange Options

The family of *rainbow options* encompasses options where the payout is based upon the relationship between two or more underlyings. Also known as *multi-factor options*, rainbow options include spread options, "better-of" options, "worst-of" options etc. A rainbow option referring to two underlyings S_1 and S_2 is sometimes called a *two-color rainbow*. The most common payoff profiles are

Rainbow:	$\max\{\, S_1(T),\, S_2(T),\, K\,\}$
Call-on-maximum:	$\max\{\, 0,\, \max\{S_1(T), S_2(T)\} - K\}$
Put-On-maximum:	$\max\{\, 0,\, K - \max\{S_1(T), S_2(T)\}\}$
Call-On-minimum:	$\max\{\, 0,\, \min\{S_1(T), S_2(T)\} - K\}$
Put-On-minimum:	$\max\{\, 0,\, K - \min\{S_1(T), S_2(T)\}\}$
Spread call:	$\max\{\, 0,\, S_2(T) - S_1(T) - K\}$
Spread put:	$\max\{\, 0,\, K - S_2(T) - S_1(T)\}$
Dual strike:	$\max\{\, 0,\, S_1(T) - K_1, S_2(T) - K_2\}$
One-for-another:	$\max\{\, 0, S_1(T) - S_2(T)\,\}$
Outperformance:	$\max\{\, S_1(T), S_2(T)\,\}$
Underperformance:	$\min\{S_1(T), S_2(T)\,\}\,.$

If the payoff contains only the underlyings S_1 and S_2 as in the last three lines of this table, then the option is also called an *exchange* option.

19.1.6 Basket-Options

The options discussed in the previous section are simple special cases of a more general class of options, the so-called *basket options*. All basket options have in common that they depend on more than one underlying. The pay-off profile could be an arbitrary function of the underlying prices. For example, a simple basket option is plain vanilla option on an equity index, since an equity index can be considered as the weighted average of a basket of share prices.

Other basket option replicate rule-based trading strategies. As an example, consider a strategy where the option premium is used to buy a basket of shares. Periodically, the share with the worst performance (keeping only the stocks with the best performance) or best performance (to lock-in the price increase so far) is sold. The cash is either distributed to the option holder or reinvested in the stocks remaining in the basket. Best-of or worst-of certificates are basket options, too. Here, the final pay-off depends on either the underlying with the best or worst performance of all shares in the basket.

19.1.7 Compound and Bermuda Options

Compound options are options whose underlying is also an option (usually a plain vanilla option). Upon exercise at maturity T of a compound option, the option holder obtains (or delivers) a plain vanilla option maturing at a later date $T' > T$ and strike K_2 on the same underlying in exchange for the strike K_1 of the compound option. The payoff profiles at maturity T are

Call on call:	$\max\{0, c_S(T, T', K_2) - K_1\}$
Call on put:	$\max\{0, p_S(T, T', K_2) - K_1\}$
Put on call:	$\max\{0, K_1 - c_S(T, T', K_2)\}$
Put on put:	$\max\{0, K_1 - p_S(T, T', K_2)\}$.

An application of compound options are the *extendible options*. At maturity T, the holder of an extendible option has the right to extend the lifetime to a future date $T' > T$. The payoff profile for extendible options are:

Extendible call:	$\max\{0, S(T) - K_2, c_S(T, T', K_2) - K_1\}$
Extendible put:	$\max\{0, K_2 - S(T), p_S(T, T', K_2) - K_1\}$,

where the strike K_1 is often chosen to be zero. With strike $K_1 = 0$, the extendible option is an example of a *Bermuda option*. A Bermuda option, also called an *Atlantic option*, is an option which can only be exercised at specific times during the lifetime of the option. It is, in a sense, an intermediate form between an American (exercise is possible at any *arbitrary* time during the lifetime of the option) and a European option (exercise is possible at a *single* time, namely at maturity). The extendible option is a Bermuda option with exactly one early-exercise date prior to maturity. Bermuda options are commonly employed as *embedded options* in structured bonds.

Another example of a compound option is the *chooser option*. At inception, the chooser option has a strike, a "chooser" date T, and a final maturity date $T' > T$. At the chooser date, the buyer can choose whether the option from

the chooser date onwards to final maturity is a call or put. The payoff profile at maturity T is thus

$$\text{Chooser:} \quad \max\{c_S(T, T', K_2), p_S(T, T', K_2)\} \, .$$

A chooser option is interesting for investors who expect strong volatility increase in the underlying but who are uncertain about the direction. It is therefore ideal mechanism to take positions on volatility.

Yet another example from the family of compound options is the *caption*. A caption is the right to buy or sell at maturity T an interest rate cap (with maturity T' and strike K_2 on, e.g. 6-months LIBOR) at strike price K_1 (call or put on cap). If at maturity T of the caption representing the right to buy (sell) a cap the underlying interest rate cap is cheaper (more expensive) in the market than the caption strike K_1, the caption holder will let the caption expire. If, however, the underlying cap is more expensive (cheaper) in the market, the caption holder will buy (sell) the cap for the strike described in the caption agreement.

The same agreement, but with reference to a floor, is known as a *floortion*.

19.2 Black-Scholes for Exotics

A closed form Black-Scholes price can be obtained for some exotics by adeptly decomposing them into a combination of options whose prices are known analytically or can be determined using known analytical methods. This will be demonstrated for pay-later and digital options. It is sometimes possible to find Black-Scholes solutions even for options which are path dependent as will be shown for the case of barrier options. Though, it should be noted that, in practice, these Black-Scholes based solutions are used as quotation vehicles only rather than as valuation models, since they neglect significant risk factors like the volatility smile.

19.2.1 Pay-Later Options

A *pay-later option* must be paid for upon exercise of the option and *only if* the option is in fact exercised. Intuitively, such an option is like an insurance premium against an accident which is payable only in event that the accident actually occurs. This may sound too good to be true but such options can actually be priced and are traded. The idea is to "hide" the option price in

the payoff profile so that the holder of the option need pay nothing upon its acquisition. To accomplish this, the cash flows upon exercise must be reduced by a certain amount in comparison with the payoff of the corresponding plain vanilla option. The payoff profiles of the put and the call are

$$c_S^{\text{later}}(T, T, K) = \begin{cases} S(T) - K - \bar{c}\,, & S(T) > K \\ 0 & ,\ S(T) \leq K \end{cases}$$

$$p_S^{\text{later}}(T, T, K) = \begin{cases} K - \bar{p} - S(T)\,, & S(T) < K \\ 0 & ,\ S(T) \geq K\,. \end{cases}$$

The option is exercised exactly when the corresponding plain vanilla would be exercised, namely when $S(T) > K$, or $S(T) < K$. The only difference is that the holder of the call at exercise has to pay a higher amount, $K + \bar{c}$, instead of the strike K, and the holder of the put receives only $K - \bar{p}$ instead of the entire strike K. To price such options, we modify the strikes ($K + \bar{c}$ for calls and $K - \bar{p}$ for puts) in the pricing formulas for plain vanilla options Eq. 8.7 to get the Black-Scholes prices:

$$c_S^{\text{later}}(t, T, K) = \tilde{S}(t, T)\mathrm{N}(x) - (K + \bar{c})\,B(t, T)\mathrm{N}(x - \sigma\sqrt{T - t})$$
$$p_S^{\text{later}}(t, T, K) = -\tilde{S}(t, T)\mathrm{N}(-x) + (K - \bar{p})\,B(t, T)\mathrm{N}(-x + \sigma\sqrt{T - t})\,,$$
$$\tag{19.1}$$

where

$$x(t) = \frac{1}{\sigma\sqrt{T - t}} \ln\left(\frac{\tilde{S}(t, T)}{K\,B_r(t, T)}\right) + \frac{1}{2}\sigma\sqrt{T - t} \tag{19.2}$$

and $\tilde{S}(t, T)$ is the dividend corrected spot price as given in Eq. 2.9. The premium to be paid at exercise, $\bar{c}$ and $\bar{p}$ for the call and put respectively, are established when the option is written, i.e., at t_0, in such a way that the option is worthless at this time. From the condition that the option price in Eq. 19.1 is zero at t_0, we can derive

$$\bar{c} = \frac{\tilde{S}(t_0, T)\mathrm{N}(x_0)}{B(t_0, T)\,\mathrm{N}(x_0 - \sigma\sqrt{T - t_0})} - K$$

$$\bar{p} = \frac{-\tilde{S}(t_0, T)\mathrm{N}(-x_0)}{B(t_0, T)\,\mathrm{N}(-x_0 + \sigma\sqrt{T - t_0})} + K\,,$$

where x_0 stands for the value of x at $t = t_0$.

19.2.2   Digital Options

Digital options, also referred to as *binary options*, are options paying a fixed *amount* where the underlying only determines whether a payoff *occurs* at all. The option holder receives the same payout irrespective of how far in the money the option closes. The payoff thus has a binary (digital) character: either a fixed amount or nothing at all is paid. We will now show how to price such options using Black-Scholes. Thus we have to assume the options to be *European*. There are however also options called *one touch digitals* in the market, which also give the buyer a fixed payout profile. Unlike ordinary digitals, one-touch digitals pay out a fixed amount if the underlying reaches the strike *at any time* during the options life time. They can therefore be considered as *American* style digital options.

Cash-or-Nothing

The holder of a *cash-or-nothing option* receives upon exercise a fixed amount of money (for example, 1 euro) if the underlying price is higher (call) or lower (put) than the strike of the option. The payoff profiles of the call and put are

$$c_S^{\text{con}}(T, T, K) = \begin{cases} 1 \, , & S(T) > K \\ 0 \, , & S(T) \leq K \end{cases}$$

$$p_S^{\text{con}}(T, T, K) = \begin{cases} 1 \, , & S(T) < K \\ 0 \, , & S(T) \geq K \, . \end{cases}$$

where "con" stands for "cash or nothing". Pricing these options is accomplished by decomposing them into options whose prices can be determined using conventional methods (*option stripping*). The payoff profile of the *pay-later* option corresponds to that of the *plain vanilla* option where the holder pays an additional fixed amount $\bar{c}$ or $\bar{p}$ upon exercise. These fixed amounts correspond to the payoff profiles of *digital* options. A pay-later call thus has the same payoff profile as a portfolio consisting of a plain vanilla call (long) and $\bar{c}$ digital options (short). An analogous relation holds for put options:

$$c_S^{\text{later}}(t, T, K) = c_S^{\text{plain}}(t, T, K) - \bar{c} \, c_S^{\text{con}}(t, T, K)$$

$$p_S^{\text{later}}(t, T, K) = p_S^{\text{plain}}(t, T, K) - \bar{p} \, p_S^{\text{con}}(t, T, K) \, .$$

Solving for $c_S^{\mathrm{con}}(t, T, K)$ and $p_S^{\mathrm{con}}(t, T, K)$ and using Eqs. 8.7 and 19.1 for the plain vanilla and the pay-later options yields the Black-Scholes prices for digital options (with x as in Eq. 19.2):

$$c_S^{\mathrm{con}}(t, T, K) = B(t, T)\mathrm{N}(x - \sigma\sqrt{T - t}) \qquad (19.3)$$

$$p_S^{\mathrm{con}}(t, T, K) = B(t, T)\mathrm{N}(-x + \sigma\sqrt{T - t}) \ .$$

These equations can be interpreted as follows: the value of the option is the discounted payoff upon exercise (1 euro) multiplied by the risk-neutral probability that the option is exercised.

An alternative approach yields a simple static hedge for a European cash-or-nothing option. A Call-Spread, consisting of a bought call option with strike K and a sold call option with strike $K + \varepsilon$ has almost everywhere, i.e. for all $S(T) > K + \varepsilon$ and for all $S(T) < K$ the same pay-off as a cash-or-nothing call paying the amount ε. Both pay-off profiles differ only in the small region $K < S(T) < K + \varepsilon$. The smaller ε, the better the approximation of the cash-or-nothing call. Therefore, the cash-or-nothing call with nominal 1 could be approximately statically replicated by buying $1/\varepsilon$ call spreads. In the limit $\varepsilon \to 0$ both pay-off profiles are identical:

$$c_S^{\mathrm{con}}(T, T, K) = \lim_{\varepsilon \to 0} \frac{\max(S(T) - K, 0) - \max(S(T) - K - \varepsilon, 0)}{\varepsilon}$$

$$= -\frac{\partial}{\partial K}\max(S(T) - K, 0) \ .$$

Therefore, the pay-off profile is identical to the (negative) partial derivative of the plain vanilla pay-off function with respect to the strike! For a call-or-nothing put an analogous consideration holds true. Since the derivative is taken with respect to the strike, not to the underlying, the derivative can be calculated outside the expectation function, i.e. the relation is valid for the present value functions, too:

$$c_S^{\mathrm{con}}(t, T, K) = -\frac{\partial}{\partial K}c_S^{\mathrm{plain}}(t, T, K)$$

$$p_S^{\mathrm{con}}(t, T, K) = -\frac{\partial}{\partial K}p_S^{\mathrm{plain}}(t, T, K) \ .$$

Comparison with Eq. 19.3 yields immediately

$$\frac{\partial}{\partial K} c_S^{\text{plain}}(t, T, K) = -B(t, T)\text{N}(x - \sigma\sqrt{T - t})$$

$$\frac{\partial}{\partial K} p_S^{\text{plain}}(t, T, K) = -B(t, T)\text{N}(-x + \sigma\sqrt{T - t}) .$$

In fact, European call-or-nothing options are often hedged and represented by means of such call spreads.

It can be observed that the implied volatilities traded liquidly in the market show a strong dependency on the strike (volatility smile). Therefore, it follows that cash-or-nothing or more generally all digital options which are sensitive to the smile or especially its slope with respect to K should always be priced with models taking smile effects into account.

Asset-or-Nothing

The holder of an *asset-or-nothing option* receives the underlying (or its cash value $S(T)$) if the underlying price is greater than (call) or less than (put) the strike of the option. Seen as such, the amount obtained upon exercise is certainly dependent on the underlying but remains binary (digital) in the sense that the option holder either receives the underlying or nothing at all. The payoff profiles of a call and put are

$$c_S^{\text{aon}}(T, T, K) = \begin{cases} S(T) , & S(T) > K \\ 0 , & S(T) \leq K \end{cases}$$

$$p_S^{\text{aon}}(T, T, K) = \begin{cases} S(T) , & S(T) < K \\ 0 , & S(T) \geq K , \end{cases}$$

where "aon" stands for, "asset or nothing". This resembles the payoff profile of the corresponding plain vanilla options with the exception that the holder of the call, for example, need not pay the strike price K in order to obtain the underlying. Were the holder to pay the strike and simultaneously receive the amount of the strike, the payoff profile would remain the same. Thus, the asset-or-nothing call can be interpreted as a plain vanilla call and K cash-or-nothing calls. An analogous decomposition can be produced for the asset-or-

nothing put. The *option stripping* of the asset-or-nothing options thus yields the expressions

$$c_S^{\text{aon}}(t, T, K) = K\, c_S^{\text{con}}(t, T, K) + c_S^{\text{plain}}(t, T, K)$$

$$p_S^{\text{aon}}(t, T, K) = K\, p_S^{\text{con}}(t, T, K) - p_S^{\text{plain}}(t, T, K)\ .$$

Inserting the Black-Scholes equations for the already known option prices discussed above gives

$$c_S^{\text{aon}}(t, T, K) = \tilde{S}(t)\mathrm{N}(x) \tag{19.4}$$

$$p_S^{\text{aon}}(t, T, K) = \tilde{S}(t)\mathrm{N}(-x)\ ,$$

where again, $\tilde{S}(t, T)$ is the spot price adjusted for dividends as given in Eq. 2.9.

Range Floater

An application of digital options is the *range floater*. A range floater is a deposit or note that accrues interest on days when the underlying S is within a predefined range between S_1 and S_2 and accrues zero on days when the underlying is outside that range. Range floaters are usually principal guaranteed so that the investor is assured of at least receiving the principal back. A range floater is thus a series of daily digital options (plus a zero bond to ensure the redemption of the principal at maturity). On each day during the life of the range floater *two* digital options mature: a long digital call with strike S_1 and a short digital call with strike S_2, with $S_1 < S_2$. Each day the underlying is above S_1 the long digital call is automatically exercised. If the underlying is above S_2 both the long and the short call will automatically be exercised, resulting in zero net payout. The range floater will therefore accrue interest equal to the digital payoff each day the underlying is within the range. Range floaters can be designed with any underlying including interest rates and FX rates. Range floaters are also known as *fairway bonds* or *fairway floaters* or *daily range accruals*.

19.2.3 Barrier Options

Barrier options have become so common (primarily in foreign exchange markets) that many investors no longer think of them as exotic. Barrier options

are European options and come in the form of *knock-in* and *knock-out* options. A knock-out option corresponds to a plain vanilla option with the additional condition that the option becomes worthless (*knocks out*) if the underlying price breaks through a specified barrier H. Depending on whether the barrier is smaller or larger than the current price of the underlying, we speak of *down-and-out* and *up-and-out* options. A knock-in option is completely analogous to the knock-out option. In this case however, the additional condition is that the option remains worthless as long as the underlying does *not* attain the barrier. As soon as the barrier has been attained, the option becomes plain vanilla option. The option is thus first "brought to life" (*knocks in*) when the barrier is attained. We speak of a *down-and-in* and *up-and-in* options, depending on whether the barrier has to be attained from above or below.

To determine whether the barrier has been hit or not, the underlying is usually measured once a day at a fixed time (say 12:00 London time) and with reference to an agreed upon source, for instance a Reuters, Bloomberg or Telerate page.

Sometimes a knock-out option is defined in such a way that it does not completely lose its value if the underlying breaks through the barrier, but a small previously agreed upon amount, called the *rebate*, is paid.[2] Analogously, a rebate is sometimes paid after the maturity of a knock-in option if the option has not been activated in its lifetime. For the sake of clarity, however, we will not enter into a further discussion of these, since such option components could be easily stripped of. A rebate is nothing else as an extra digital option with American exercise often called *touch* option (see Sect. 19.2.2).

The payoff profiles at maturity T of knock-out barrier options with strike K and barrier H on an underlying with a spot price $S(t)$ are:

Up-and-out call:	$\max\{0, S(t) - K\}$ if $S(t') < H$
Down-and-out call:	$\max\{0, S(t) - K\}$ if $S(t') > H$
Up-and-out put:	$\max\{0, K - S(t)\}$ if $S(t') < H$
Down-and-out put:	$\max\{0, K - S(t)\}$ if $S(t') > H$,

where the conditions must hold for *all* t' such that $t < t' \leq T$. The payoff profiles of the corresponding knock-in barrier options are:

Up-and-In Call:	$\max\{0, S(t) - K\}$ if $S(t') \geq H$
Down-and-In Call:	$\max\{0, S(t) - K\}$ if $S(t') \leq H$
Up-and-In Put:	$\max\{0, K - S(t)\}$ if $S(t') \geq H$
Down-and-In Put:	$\max\{0, K - S(t)\}$ if $S(t') \leq H$,

---

[2]This concept is useful in the valuation of companies using option pricing theory.

where the conditions must in each case hold for *at least one t'* with $t < t' \leq T$.

A portfolio consisting of a knock-out option and the corresponding knock-in option, for example a down-and-out call and a down-and-in call, yields the same payoff as the corresponding plain vanilla option. If the barrier is attained then the knock-out option "dies" and the knock-in becomes a plain vanilla call. Conversely, if the barrier is never attained in the lifetime of the options, the knock-out pays the same as a plain vanilla call and the knock-in is worthless. Thus, a portfolio consisting of a knock-out option and the corresponding knock-in option is worth exactly the same as the corresponding plain vanilla option for all times before and at maturity:

$$c_s^{\text{plain}}(t, T, K) = c_s^{\text{di}}(t, T, K, H) + c_s^{\text{do}}(t, T, K, H) \tag{19.5}$$
$$= c_s^{\text{ui}}(t, T, K, H) + c_s^{\text{uo}}(t, T, K, H)$$
$$p_s^{\text{plain}}(t, T, K) = p_s^{\text{di}}(t, T, K, H) + p_s^{\text{do}}(t, T, K, H)$$
$$= p_s^{\text{ui}}(t, T, K, H) + p_s^{\text{uo}}(t, T, K, H) \,,$$

where the superscripts "di", "do", etc. stand for "down and in", "down and out", etc. It is therefore sufficient to determine the prices of, for example, the knock-out options. The knock-in prices then follow immediately from Eq. 19.5. In addition to the four different barrier option variants, we must further distinguish whether the strike is above or below the barrier. Hence, we need to find equations for 8 different cases. With an appropriate choice of notation, these 8 cases can be summarized in a single Black-Scholes prices for knock-out options (the derivation of which will not be given here):

$$V^{\text{out}} = \delta\, e^{-q(T-t)} S(t) \left\{ N\left(\frac{\beta - x_+}{\sigma\sqrt{T-t}}\right) - N\left(\frac{\alpha - x_+}{\sigma\sqrt{T-t}}\right) \right. \tag{19.6}$$
$$\left. -e^{-\lambda h}\left[N\left(\frac{\beta - x_+ + 2h}{\sigma\sqrt{T-t}}\right) - N\left(\frac{\alpha - x_+ + 2h}{\sigma\sqrt{T-t}}\right) \right] \right\}$$
$$- \delta\, B(t, T) K \left\{ N\left(\frac{\beta - x_-}{\sigma\sqrt{T-t}}\right) - N\left(\frac{\alpha - x_-}{\sigma\sqrt{T-t}}\right) \right.$$
$$\left. -e^{(2-\lambda)h}\left[N\left(\frac{\beta - x_- + 2h}{\sigma\sqrt{T-t}}\right) - N\left(\frac{\alpha - x_- + 2h}{\sigma\sqrt{T-t}}\right) \right] \right\} \,.$$

Table 19.1 Parameters for the valuation of knock-out options using Eq. 19.6

	$H \leq K$	$H > K$
Up-and-out call	Pointless	$\delta = +1$
		$\alpha = -\ln\left(\frac{S(t)}{K}\right)$
		$\beta = -h$
Up-and-out put	$\delta = -1$	$\delta = -1$
	$\alpha = -\infty$	$\alpha = -\infty$
	$\beta = -h$	$\beta = -\ln\left(\frac{S(t)}{K}\right)$
Down-and-out call	$\delta = +1$	$\delta = +1$
	$\alpha = -\ln\left(\frac{S(t)}{K}\right)$	$\alpha = -h$
	$\beta = +\infty$	$\beta = +\infty$
Down-and-out put	$\delta = -1$	Pointless
	$\alpha = -h$	
	$\beta = -\ln\left(\frac{S(t)}{K}\right)$	

In this equation, the following parameters are defined to hold for all knock-out options:

$$x_+ = \ln\left(\frac{e^{-q(T-t)}}{B(t,T)}\right) + \frac{\sigma^2}{2}(T - t) \tag{19.7}$$

$$x_- = \ln\left(\frac{e^{-q(T-t)}}{B(t,T)}\right) - \frac{\sigma^2}{2}(T - t)$$

$$\lambda = 1 + \frac{\ln\left(\frac{e^{-q(T-t)}}{B(t,T)}\right)}{\frac{\sigma^2}{2}(T - t)}, \quad h = \ln\left(\frac{S(t)}{H}\right).$$

The parameters α, β and δ, on the other hand, are not identical for all knock-out options. These three parameters are presented in Table 19.1 for the different types of knock-out options. Together with the Eqs. 19.6 and 19.7, this table provides all the necessary information for determining the Black-Scholes price of all knock-out options.[3]

The case $H \leq K$ for an up-and-out call does not make sense since for the call to pay out anything the underlying must be larger than the strike K. But for this to happen, it must cross the barrier H since $H \leq K$. But then the up-and-out call "dies". Therefore an up-and-out call with $H \leq K$ never pays out

---

[3]The following well-known properties of the normal distribution are useful for the valuation: $N(\infty) = 1$, $N(-\infty) = 0$, $1 - N(x) = N(-x)$.

anything. Such an option is always worth zero. Inserting this result in Eq. 19.5 shows that the corresponding up-and-in call is worth exactly as much as the plain vanilla call. This makes sense since for the call to pay out anything the underlying must be larger then the strike K and therefore cross the barrier. Thus, an up-and-in call with $H \leq K$ will be "alive" in *every* situation where the corresponding plain vanilla call pays out anything. Therefore it has the same value as the plain vanilla call.

The same reasoning leads to the result that a down-and-out put with $H > K$ never pays out anything (and is therefore worth zero) and the down-and-in put with $H > K$ is worth exactly as much as the corresponding plain vanilla put.

As an example, we present the explicit equation for the price of a down-and-out call for the case that the barrier is less than or equal to the strike. Referring to Table 19.1 for this case, we find that $\delta = +1, \alpha = -\ln(S(t)/K)$, $\beta = +\infty$. Substituting these parameters into Eq. 19.6 and making use of the properties of the normal distribution yields

$$c_S^{\text{do}}(t, T, H \leq K) =$$

$$e^{-q(T-t)} S(t) \left[N\left(\frac{x_+ + \ln\left(\frac{S(t)}{K}\right)}{\sigma\sqrt{T-t}} \right) - e^{-\lambda h} N\left(\frac{x_+ + \ln\left(\frac{S(t)}{K}\right) - 2h}{\sigma\sqrt{T-t}} \right) \right]$$

$$- B(t, T) K \left[N\left(\frac{x_- + \ln\left(\frac{S(t)}{K}\right)}{\sigma\sqrt{T-t}} \right) - e^{(2-\lambda)h} N\left(\frac{x_- + \ln\left(\frac{S(t)}{K}\right) - 2h}{\sigma\sqrt{T-t}} \right) \right] .$$

Here, the argument $x_+ + \ln(S(t)/K)$ is precisely equal to the x known from the plain vanilla options (see for example Eq. 8.5). Substituting this and all other parameters as defined in Eq. 19.7 (with the exception of λ) finally leads to the equation for the price of a down-and-out call, in agreement with that found by Cox and Rubinstein in [43], for example:

$$c_S^{\text{do}}(t, T, H \leq K) = \overbrace{e^{-q(T-t)} S(t) N(x) - B(t, T) S(t) N(x - \sigma\sqrt{T-t})}^{\text{plain vanilla}}$$

$$- \left(\frac{S(t)}{H} \right)^{-\lambda} \left[e^{-q(T-t)} S(t) N(y) - B(t, T) K \left(\frac{S(t)}{H} \right)^2 N(y - \sigma\sqrt{T-t}) \right]$$

$$(19.8)$$

where the following new abbreviations have been introduced

$$y = x - \frac{2}{\sigma\sqrt{T-t}}\ln\left(\frac{S(t)}{H}\right) = \frac{1}{\sigma\sqrt{T-t}}\ln\left(\frac{e^{-q(T-t)}H^2}{B(t,T)K\,S(t)}\right) + \frac{1}{2}\sigma\sqrt{T-t}\,.$$

The price of the down-and-out call is thus equal to the price of the corresponding plain vanilla call less an amount generated by the barrier. It is thus always cheaper than the plain vanilla call. This makes sense, since its payoff in the most favorable case (when the barrier is not attained) equals that of the plain vanilla call. The barrier increases the risk that a payoff will not be made and thus the barrier option must always be less expensive than its plain vanilla counterpart. They are therefore the appropriate instruments for investors who have an exact opinion as to future developments in the price of an underlying and, based on this opinion, wish to decrease the costs of a hedge, for example. The speculative element should not be overlooked here: when the barrier is attained, a hedging portfolio consisting of knock-out options disappears!

The price of the corresponding down-and-in call follows immediately from Eq. 19.5. The value of this call is just the difference between the price of the plain vanilla and the down-and-out call:

$$c_S^{di}(t, T, H \le K) \tag{19.9}$$

$$= \left(\frac{S(t)}{H}\right)^{-\lambda}\left[e^{-q(T-t)}S(t)\mathrm{N}\,(y) - B(t,T)K\left(\frac{S(t)}{H}\right)^2\mathrm{N}\left(y - \sigma\sqrt{T-t}\right)\right]\,.$$

Double Barrier Option

In order to receive the option payout of a *double barrier option* the underlying must not breach either of *two* defined barriers at any time during the life of the transaction. Generally, the barriers are used to define a range inside which the underlying must remain for the life of the transaction in order for the option payout to be made. If the underlying as measured on any day is outside the defined range, the payout is zero, i.e., the holder has been knocked out. The payout for the double barrier option is usually a fixed amount and hence this option type is also known as a *binary knockout range*.

A double barrier option is *not* the addition of two knock-out options. If this were the case, once one barrier had been breached the other option would still be "alive". A double barrier option is the addition of two "*contingent*" Knockout Options, i.e., the "survival" of each barrier option is contingent

upon the other barrier not having being breached. Should one barrier be breached, the other barrier option also dies.

Double barrier options are suitable to take advantage of markets expected to trade within a range. Most common in foreign exchange, they are available in most underlying markets, including commodities, interest rate and equity markets.

19.2.4 Ladder Options

With a *ladder option*, the strike is periodically reset when the underlying hits specified trigger levels, at the same time locking in the profit between the old and the new strike. The trigger strikes appear as rungs on a ladder. This behavior of the option can be summarized by the following payout function: at expiry the option will pay out either the difference between the underlying spot price $S(T)$ and the original strike K_0, or the difference between the highest rung reached and the original strike, whichever is greater. The payoff profile in its simplest form is therefore

$$\max \left\{ S(T) - K_0, \ \max_i \{K_i - K_0\}, \ 0 \right\} \quad \text{with} \quad K_0 < K_1 < K_2 < \cdots < K_n \,,$$

where $K_1, K_2, \ldots, K_n$, represent the "rungs on the ladder" established in the contract and K_0 the original strike of the option. The maximum function for these ladder steps refers only to *those* rungs K_i attained by the underlying during the lifetime of the option. Upon exercise of such an option, the holder receives the same payoff as that of a plain vanilla option if $S(T)$ is greater than each of the previously attained steps. However, if the underlying price has declined toward the end of its lifetime, the holder receives the difference between the original strike K_0 and the highest rung attained during the lifetime of the option. Ladder options can be structured to reset the strike in either one or both directions. Also a "Ladder Put" could be constructed along the same lines (if we consider the above payoff profile as a "call"). Ladder options are also known as a *ratchet options* or *lock-in options*. A ladder option with the above payoff profile can be constructed from a plain vanilla call and as a series of long and short knock-in puts (*option stripping*). Since knock-in options can be priced using Eqs. 19.5 and 19.6, a Black-Scholes price for a ladder option can be obtained by pricing the corresponding portfolio of plain

vanilla and barrier options. Explicitly, the portfolio which has the same payoff profile as the ladder option consists of

- one *long* plain vanilla call with strike K_0.
- for each rung K_i (with $i = 1, \ldots, n$) one *long* knock-in put with knock-in barrier $H_i = K_i$ and strike K_i.
- for each rung K_i (with $i = 1, \ldots, n$) one *short* knock-in put with knock-in barrier $H_i = K_i$ and strike K_{i-1}.

All these options have maturity T. To check that this portfolio indeed replicates a ladder option we consider a ladder option with only one rung K_1 and an initial strike $K_0 < K_1$. With one rung and one strike, five different situations can arise at maturity. These are displayed in the following table:

	K_1 reached	K_1 not reached
$S(T) > K_1$	case a)	impossible
$K_0 < S(T) < K_1$	case b)	case d)
$S(T) < K_0$	Fall c)	case e).

The replicating portfolio consists of the plain vanilla call with strike K_0, a long knock-in put with barrier K_1 and strike K_1 and a short knock-in put with barrier K_1 and strike K_0. We will now show that in each case the portfolio has the same payoff as the ladder option:

Case a
- payoff of ladder option: $S(T) - K_0$
- payoff of options portfolio: payoff from plain vanilla call is $S(T) - K_0$; the puts are both knocked in but expire worthless.

Case b
- payoff of ladder option: $K_1 - K_0$
- payoff of options portfolio: payoff from plain vanilla call is $S(T) - K_0$; the puts are both knocked in; payoff from long put is $K_1 - S(T)$; short put expires worthless; altogether: $S(T) - K_0 + K_1 - S(T) = K_1 - K_0$

Case c
- payoff of ladder option: $K_1 - K_0$
- payoff of options portfolio: plain vanilla call expires worthless; the puts are both knocked in; payoff from long put is $K_1 - S(T)$; payoff from short put is $-[K_0 - S(T)]$; altogether: $K_1 - S(T) - [K_0 - S(T)] = K_1 - K_0$

Case d

- – payoff of ladder option: $S(T) - K_0$
- – payoff of options portfolio: payoff from plain vanilla call is $S(T) - K_0$; the puts are never knocked in.

Case e

- – payoff of ladder option: 0
- – payoff of options portfolio: plain vanilla call expires worthless; the puts are never knocked in.

For more than one rung we simply repeat the above idea: each time a rung K_i is reached for the first time by the underlying, a put is knocked in which guarantees that the underlying can be sold for at least K_i. That locks in a profit $K_i - K_0$ with certainty since at maturity the underlying can be bought for K_0 because of the call and then immediately sold for K_i because of the put. However, when the next rung K_{i+1} is reached, one more put is knocked in and we would then have two puts; the "old" one with strike K_i and the "new" one with strike K_{i+1}. That is one put too many. Therefore the "old" put (the one with strike K_i) now has to be "destroyed". This is achieved by the short knock-in put with strike K_i which comes into being exactly when K_{i+1} is hit. The payoff of this short put then exactly compensates for the payoff of the "old" long put. The portfolio is then left with the payoffs of the new put and the plain vanilla call and thus, a payoff $K_{i+1} - K_0$ is locked in with certainty, just as it should be to replicate the ladder option.

19.3 Numerical Pricing Methods for Exotics

The list of exotics described above is far from complete and an analytic price determination—even under the assumptions of the Black-Scholes world—is not possible for the majority of exotics. The multitude of exotic instruments can be roughly assigned to four different classes as indicated in Table 19.2. The options are differentiated by whether the payoff profile is dependent solely on current values of the underlying or on earlier values as well and whether or not early exercise is allowed. We speak of a *path dependent* option if the option's value is dependent on the history of the underlying, i.e., if the price evolution (the *path*) of the underlying during the lifetime of the option affects the option's payoff. Assuming the availability of adequate computer capacity and performance, some common numerical pricing methods leading to reasonable prices for exotic derivatives are indicated in Table 19.2. The fundamental

Table 19.2 Classification and "brute force" valuation methods for exotic options

	European	American
Path-independent	Grid	Grid
	trees	trees
	Monte-Carlo	
Path-dependent	(Grid)	(Grid)
	Monte-Carlo	(Non-recombining trees)
		American Monte-Carlo

assumption behind these methods is that the underlying is governed by a random walk as given in Eq. 2.17, or more generally by an Ito process as defined by Eq. 2.19.

In this table, "grid" means that the price can be determined by solving a partial differential equation under certain boundary conditions (determined by the option to be priced) on a grid. These include numerical methods such as the finite difference methods (see Chap. 10) or finite elements. These methods approximate the solution on a fixed grid or lattice. Strictly speaking, tree methods fall also in the class of grid methods, though we consider tree methods here as distinct from grid methods, since they are usually implemented without relating explicitly to the underlying PDE. For path dependent problems, grid methods can only be applied in special cases, of which barrier options are an example. Another special case is a problem, where the path dependency can be summarized in a single parameter and the method of *extension of dimensions* can be applied (here, an additional virtual dimension representing this additional parameter is introduced, such that many different paths can be calculated in parallel on the grid). Therefore, we have put the word "grid" in brackets for this class of problems. Non-recombining (or "bushy") trees have such a bad performance that they play a minor role only since they have no obvious advantage over Monte-Carlo methods when path-dependent problems need to be solved (see also Sect. 11.5 or Sect. 14.10).

Another important aspect is the number of risk factors. Grid and tree methods are used mostly for low-dimensional problems, i.e. with one or two, rarely with three or for dimensions. The number of dimensions is determined by the number of stochastic factors. A plain vanilla call on a stock is an example for a one-dimensional problem. If the volatility is also being considered as stochastic, this becomes a two- (or more) dimensional problem. Long term options on FX rates require the simulation of both involved interest rate curve an the FX rate itself, yielding to at least three dimensions, etc.

For high-dimensional problems, Monte-Carlo is the method of choice. E.g., implementing a 10-dimensional problem in Monte-Carlo is only marginally

more complex than implementing a 3-factor model, in stark contrast to grid methods—every additional dimension increases the complexity significantly. Still, it is easy to reach the computational limit with Monte-Carlo methods, too, mainly because of the rapidly increasing computation time for more complex problems. Also, the difficulties to calibrate a high-dimensional model are often underestimated. Though it very easy to produce number with a badly calibrated model, the results are of limited use, since they often don't have the required accuracy or stability. Another problem is the computation of sensitivities resp. the Greeks required for hedging, if Monte-Carlo methods are used, which is significantly more difficult than computing present values.

The discussed methods so far are very generic in the sense that they could be implemented as very general procedures. The very same code can than be applied to a broad variety of products, and, depending on the implementation, also for different models or model variants. In addition to these generic models, there are many valuation models and techniques which can be applied only to very specific products and could not be generalized to a broader variety of situations. Examples for these specific methods are replication approaches (e.g. for CMS-Swaps, see Sect. 17.7) or Fast Fourier Transformations. These specialized approaches are often more accurate, cover more relevant features or increase computational performance dramatically compared to generic methods. If such special methods exists for a special product, it's often the first choice. However, for the remainder of the chapter we concentrate on generic methods.

19.3.1 Monte Carlo for European Exotics

A method which, in principle (neglecting potential problems regarding CPU time) can be applied to all derivatives with a European payoff mode is the Monte Carlo method introduced in Chap. 11. Using a "Pseudo-Code", we will demonstrate how *Asian options* can be priced with this method. The schematic algorithm offered here explicitly shows how the chapters on Monte Carlo simulation and on random walks can be applied to option pricing. The algorithm is suitable for the valuation of all European options depending on a single underlying (with or without dividends). Multiple underlyings can be simulated using the methods discussed in Sect. 11.3.2 the for the case of two underlyings. The generalization of this method for more than two underlyings can be realized using Eqs. 21.41 and 23.1.

I the following concrete example, the algorithm has been explicitly formulated for Asian options to demonstrate the general procedure. The complete

algorithm can be found in the form of an executable Visual Basic program contained in the Excel workbook MONTECARLOSIMULATION.xlsm from the download section [50]. The variable $x = \ln(S(t')/S(t))$ is simulated as a random walk where $S(t')$ is the price (denoted by S in the algorithm) at the current simulation time $t' = t + i\,dt$ and $S(t)$ (denoted by S0 in the algorithm) is the price at the start of the simulation, i.e., at the value date t. Since the Monte Carlo simulations are based on random walks and, as such, on random numbers, this method will generate a somewhat different option price with each simulation run. To demonstrate the calculation of the error in accordance with Sect. 31.2, the statistical error (denoted by EPrice) of the option price generated by the simulation will also be determined.

```
1    '* Initialization of the payoff profile:
2    '* ======================================
3    Payoff1 = 0
4    Payoff2 = 0
5    '* The loop over i is the loop over the paths:
6    '* ===========================================
7    For i = 1 to Path
8    '* Initialization at the start of each path:
9    '* =========================================
10     SamplePath = 0
11     x = 0
12   '* The loop over j generates one underlying path:
13   '* ==============================================
14     For j = 1 to Steps
15   '* The next step on the path of the underlying's random walk:
16   '* =========================================================
17       dx = drift * dt + vola * randomnumber() * Squareroot(dt)
18       x = x + dx
19       S = S0 * Exp(x)
20   '* The evaluation after each step in a path:
21   '* =========================================
22       SamplePath = SamplePath + S
23     Next j
24   '* The evaluation after each path:
25   '* ==============================
26   '* For Average-Price Call:
27     SamplePath = SamplePath / Steps
28     If SamplePath > Strike then
29       Payoff1 = Payoff1 + (SamplePath - Strike)
30       Payoff2 = Payoff2 + (SamplePath - Strike)^2
31     End If
32   '*For Average-Price Put:
33   '* SamplePath = SamplePath / Steps
34   '* If Strike > SamplePath Then
35   '*    Payoff1 = Payoff1 + (Strike - SamplePath)
36   '*    Payoff2 = Payoff2 + (Strike - SamplePath)^2
37   '* End If
38   '*For Average-Strike Call:
39   '* SamplePath = SamplePath / Steps
40   '* If S > SamplePath Then
41   '*    Payoff1 = Payoff1 + (S - SamplePath)
42   '*    Payoff2 = Payoff2 + (S - SamplePath)^2
43   '* End If
44   '*For Average-Strike Put:
```

```
45  '* SamplePath = SamplePath / Steps
46  '* If SamplePath > S Then
47  '*    Payoff1 = Payoff1 + (SamplePath - S)
48  '*    Payoff2 = Payoff2 + (SamplePath - S)^2
49  '* End If
50  Next i
51  '* final evaluation of the price and its error after the simulation
52  '* =================================================================
53  Payoff1 = Payoff1 / Paths
54  Payoff2 = Payoff2 / Paths
55  Price = Payoff1
56  EPrice = Squareroot((Payoff2-Payoff1)^2)/Squareroot(Paths-1)
57  '* Discounting for Options which are not future-styled:
58  Price = Exp(-InterestRate * Time) * Price
59  EPrice = Exp(-InterestRate * Time) * EPrice
```

According to Eq. 12.17, the option sensitivities ("*Greeks*") can be determined by running the simulation twice, having modified the parameter according to the sensitivity under consideration while using the *same* random numbers as in the previous simulation run. The difference in the price thus obtained, divided by the parameter change, yields the corresponding sensitivity. Intuitively, this is the slope of the line passing through two measured prices. Two effects contribute to the error of this slope induced by the error of the option prices:

- *Statistical Error* This error is generated by the measurement error in both prices. The error associated with the slope of the line passing through two points, whose positions are known only up to an error term, increases the closer these points lie to one another. This is illustrated in Fig. 19.1. In order to minimize this error, the measured points should therefore lie as far from one another as possible. This means that the parameter change should be as large as possible.
- *Systematic Error* The sensitivities are defined as the derivatives of the price function with respect its different parameters, i.e., they represent the change of the price with respect to a *small* change in the parameter, divided by this
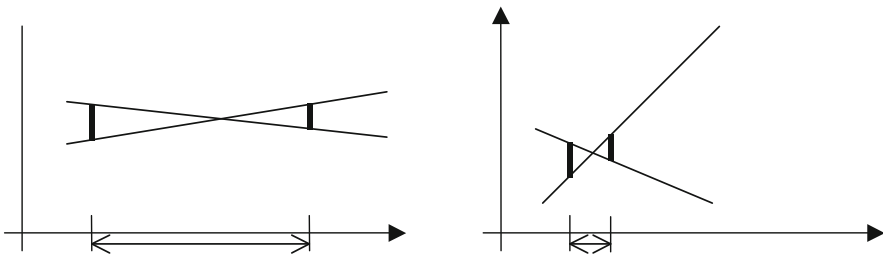
Fig. 19.1 The range of possible straight lines through two points where the points are known only up to a certain error represented by the bars

change in the parameter. Thus, choosing a large change in the parameter to reduce the statistical error means moving away from the actual definition of the sensitivities and thereby committing a *systematic* error.

Both of these effects compete with one another. The "most favorable" choice for the size of the parameter change is always a compromise between a large parameter change minimizing the effect of the statistical error and a small parameter change minimizing the systematic error. Alternative methods for the calculation of sensitivities could be found, e.g., in [4].

As the commentary in the algorithm suggests, only minor modifications need to be made when other option types are priced. The locations in the code where these modifications are to be made are written in bold print.[4] The lines printed in bold are those which serve to measure the payoff profile and its error and generate the values of "Payoff1" and "Payoff2".

If, for example, a barrier option (with the additional input parameter Barrier) were to be priced instead of an Asian, the lines from 20 to 31 would have to be replaced by:

```
1   '* Initialization at the start of each path:
2   '* =======================================
3   SamplePath = 0
4   '* The evaluation after each step of the path:
5   '* ==========================================
6   If S > Barrier
7      SamplePath = 1
8   End If
9   '* The evaluation at the end of each path:
10  '* ======================================
11  If SamplePath = 1 And S > Strike Then
12     Payoff1 = Payoff1 + (S - Strike)
13     Payoff2 = Payoff2 + (S - Strike)^2
14  End If
```

That is all! By modifying these few lines measuring the payoff profile, barrier options can be priced instead of Asian options together with all their respective Greeks an error terms.

This example should serve to illustrate that many European options, regardless of how complicated, can be priced with the Monte Carlo method if the payoff profile is known. This means that as soon as the option is defined (through its payoff) the valuation via Monte Carlo simulation is as good as complete!

Some of the options listed in Sect. 19.1 require extensions of the above given algorithm which will not be presented here. Rainbow and exchange

---

[4]In addition there may of course be changes in the input parameters.

options, for example, can be priced through a simulation of correlated prices. An *extendible option*, for example, an extendible call, is priced by simulating its underlying's price $S(T)$ at the first maturity date T. The payoff of the extendible call is then the maximum of a) the payoff of a plain vanilla call with strike K_2 and maturity T and b) the value of a plain vanilla call with strike K_2 and maturity T' as of the time point T (calculated with Black-Scholes, for instance) less the strike, K_1, of the extendible option.

19.3.2 The Binomial Model for American Exotics

Though it is possible to price American or Bermudan style options with Monte-Carlo simulations (see Sects. 11.5 and 14.10), it is significantly simpler and more accurate to implement grid based methods for this kind of options.

All grid methods have in common that option values are only calculated for special points in the parameter space (where time and underlying price are the parameters), which is called discretization. The set of these points in the parameter space actually form the grid. Then, the solution on each grid point is calculated by either solving the partial differential equation (finite difference method), integration of the stochastic equation (finite elements) or by simulating the stochastic process (binomial or trinomial tree) by applying adequate boundary conditions, and linked to the solutions for neighboring grid nodes.

This yields many equations which need to be solved either iteratively or as a system of equations. The binomial tree is as well intuitive and comparatively simple to implement. Therefore, we concentrate in the following on a schematic implementation of a binomial tree as described in Chap. 9. The complete algorithm can be found as an executable Visual Basic program in the Excel spreadsheet BINOMIALTREE.xlsm from the download section [50].

Trinomial trees show a slightly better convergence performance than binomial trees, though finite differences method offer significantly more flexibility (e.g. by choice of various discretization schemes) compared to tree methods and show a superior convergence behavior. For an efficient implementation of grid methods, finite difference should be preferred over tree methods.

Recombinant Trees for Path-Independent Options

As emphasized in Chap. 9, recombinant trees are only suitable for path-independent options since the path taken to arrive at a particular node is not

unique if the branches of the tree recombine. The payoff profile of an (exotic)
option is then defined on the boundary of the tree, in particular at maturity T
of the option. We subsequently calculate backward in time applying Eq. 9.8.
The algorithm is actually a mere reformulation of Fig. 18.1. It will be applied
here to price an American *self quanto option*. The payoff profile of this option,
as was noted in Sect. 19.1.1, is $S(T) \times \max\{0, S(T) - K\}$. Naturally, any
other path-dependent payoff profile could have been inserted instead. The
algorithm can be generalized for options with multiple underlyings but such
a generalization is quite involved as the binomial tree takes on an additional
dimension with the addition of each new underlying [167].

The time to maturity $T - t$ is denoted by "t" in the algorithm, the number
of binomial steps by "n". The computations are performed with continuous
compounding.

```
1   dt = t / n
2   discount = Exp(-r * dt)
3   up = Exp(vola * Squareroot(dt))
4   down = 1 / up
5   p = (Exp((r - q) * dt) - down) / (up - down)
6   '* Inserting the boundary values for the option:
7   '* =============================================
8   For j = 0 to n
9     S = S0 * up ^ j * down ^ (n - j)
10  '* Payoff Profile:
11    exercise = S * ( S - Strike )
12    If Optiontype = Put Then
13      exercise = -exercise
14    End If
15    If exercise > 0 Then
16      f(n; j) = exercise
17    End If
18  Next j
19  '* Calculate backwards for each node in the tree:
20  '* =============================================
21  For i = n - 1 To 0 stepsize -1
22   For j = i To 0 stepsize -1
23      S= S0 * up ^ j * down ^ (n - j)
24  '* Intrinsic Value:
25  '* ================
26      exercise = S * ( S - Strike )
27      If Optiontype = Put Then
28        exercise = -exercise
29      End If
30  '* Option price:
31  '* ============
32      f(i; j) = discount * (p * f(i + 1; j + 1)
33               + (1 - p) * f(i + 1; j))
34      If exercise > f(i; j) And Optionstyle = American
```

```
35      Then
36         f(i; j) = exercise
37      End If
38    Next j
39  Next i
40  Price(c) = f(0; 0)
```

As in the above algorithm for the Monte Carlo simulation, only a few lines need to be changed to price a different type of options. These are the locations in the code where the payoff profile is defined in the program, i.e., where both the values at the tree's boundary (at time T) are specified and where the intrinsic value is computed for the purpose of deciding whether to exercise early. For example, a power option can be priced by simply replacing the lines 11 and 26 by "exercise = (S - Strike)^P".

Further boundary conditions, for example for specific *underlying* values (as for barrier options, for example) can be quite easily accounted for as well by setting the option's values at the nodes lying at the boundary equal to the given boundary values. For reasons of stability the grid should always be constructed so that the boundaries lie *directly on the nodes* (this applies to all grid methods, also for instance for finite differences schemes).

Binomial Bushy Trees for Path-Dependent Options

The schematic algorithm presented here is suitable for the valuation of *path-dependent* American (and European) options. So-called *bushy trees* will be constructed in which each path of the underlying (denoted by "u") is generated without regard to recombination. Information on the entire history of the underlying is then accessible. Each individual underlying path is then evaluated to obtain the variable (denoted by "x" in the algorithm) determining the price of the option. For Asian options, for example, this is the average price of the underlying over the lifetime of the option. This will change from path to path, even if the paths arrive at the same end point. Once the path-dependent variables are determined, the payoff profile of the option on the boundary of the tree can be implemented, finally using Eq. 9.8 to calculate backwards in time along the tree. The procedure is demonstrated here for Asian average price options, where the lines corresponding to the Average Strike options are indicated in the commentary. The right to early exercise is accounted for in the algorithm.

```
1  dt = t / n
2  discount = Exp(-r * dt)
3  up = Exp(Vola * Squareroot(dt))
4  down = 1 / up
```

```
 5  p = (Exp((r - q) * dt) - down) / (up - down)
 6  If Dividendpayment Then
 7    divstep = Rounddown(Dividendtime * n / t; 0)
 8  End If
 9  '* Construction of the Underlying:
10  '* =============================
11  u(0; 0) = S0
12  x(0; 0) = S0
13  For i = 0 To n - 1
14    For j = 0 To 2^i - 1
15      u(i + 1; 2 * j + 1) = u(i; j) * up
16      u(i + 1; 2 * j) = u(i; j) * down
17      If Dividendpayment And i = divstep Then
18        u(i + 1; 2 * j + 1) = u(i + 1; 2 * j + 1) - Dividends
19        u(i + 1; 2 * j) = u(i + 1; 2 * j) - Dividends
20      End If
21  '* Evaluation of the underlying after each step in the path:
22  '* ==========================================================
23  '* Asian:
24      x(i + 1; 2 * j + 1) = u(i + 1; 2 * j + 1) + x(i; j)
25      x(i + 1; 2 * j) = u(i + 1; 2 * j) + x(i; j)
26    Next j
27  Next i
28  '* Inserting the boundary values for the option at maturity:
29  '* ==========================================================
30  For j = 0 To 2^n - 1
31  '* Average Price:
32    exercise = x(n; j) / (n + 1) - Strike
33  '* Average Strike:
34  '* exercise = u(n; j) - x(n; j) / (n + 1)
35    If Optiontype = Put Then
36      exercise = -exercise
37    End If
38    If exercise > 0 Then
39      f(n; j) = exercise
40    End If
41  Next j
42  '* Valuation of the option:
43  '* =======================
44  For i = n - 1 To 0 Steplength -1
45    For j = 0 To 2^i - 1
46  '* Intrinsic value:
47  '* ================
48  '* Average Price:
49      exercise = x(i; j) / (i + 1) - Strike
50  '* Average Strike:
51  '*     exercise = u(i; j) - x(i; j) / (i + 1)
52      If Optiontype = Put Then
53        exercise = -exercise
54      End If
55  '* Option price:
56  '* ============
57      f(i; j) = discount * (p * f(i + 1; 2 * j + 1)
58              + (1 - p) * f(i + 1; 2 * j))
59      If exercise > f(i; j) And Optionstyle = American
60        f(i; j) = exercise
```

```
61       End If
62     Next j
63   Next i
64   Price(c) = f(0; 0)
```

Again, only a few lines need to be modified if an option of another type is to be priced. These are the locations where the payoff profile and the procedure for calculating the particular path-dependent property of the underlying enter into the code. If we are interested in pricing, for example, an American *look-back option* (minimum-strike call) with a payoff profile $S(T) - \min\{S(t')\}$, the lines 23 to 25 should be replaced by

```
1   '* Determination of the minimal price:
2   If x(i + 1; 2 * j + 1) > u(i + 1; 2 * j + 1) Then
3     x(i + 1; 2 * j + 1) = u(i + 1; 2 * j + 1)
```

and lines 32 and 49 by:

```
1   exercise = u(n;j) - x(n; j)
```

and

```
1   exercise = u(i;j) - x(i; j)
```

These slight modifications accomplish the valuation of a look-back option rather than the original Asian option. The difference between the current price and the minimum attained up to that point is computed at maturity as well as when deciding whether to exercise early.

Naturally, the numerical methods just introduced are quite computation intensive. The computation of *bushy trees* in particular is extremely time consuming. Let n denote the number of binomial steps. The computation time and memory requirements are then proportional to $(n + 1)2^{(n+1)}$. Each additional step more than doubles the computation time and memory required! Depending on the hardware, more than 14 steps can hardly be computed. For example, 14 steps give $(14 + 1)2^{(14+1)} = 491,520$, for 15, we already have $1,048,576$. The algorithms introduced here have not been optimized in any way with respect to reducing the computation time. They were written to ensure the clarity of the pricing algorithm in each case. Despite these difficulties, a Visual Basic Excel program based on these algorithms can price an exotic option on a PC in a few seconds.

20

Check for updates

Credit Risk

Up to now we didn't take any credit risk into account but have been looking exclusively at market risk which, in the case of an interest rate instrument, is simply the risk that the default-risk free discount curve changes. However, the biggest risk a bond investor faces is in fact the risk that the bond issuer doesn't pay his or her obligations—the worst case being the complete loss of the invested capital due to the default of the obligor. On the other hand, the largest possible gain is the sum of all interest payments.[1] Dependent on its specific default-risk a bond is traded at a market price lower than its default-risk-free present value. We will now present the most important techniques to incorporate such default-risks in the pricing and risk management of financial instruments. A more detailed introduction to these topics can be found for instance in the book of Schönbucher [172].

20.1 Expected Positive Exposure, Probability of Default and Loss Given Default

To take default risk into account the following questions have to be answered:

- How large is the amount receivable from the issuer at the time of default?

---

[1]At least if interest rates are not negative, which is not always the case (e.g., CHF in 2012). However, as long as interest rates are not too strongly negative, more precisely, as long as the yield to maturity is still positive, the possible gain is still approximately the sum of all interest payments.

- What is the probability for the issuer to default within a certain future time span?
- What will be the realized loss (as a percentage of the amount receivable) at the time of default?

The answers to these questions are formally denoted as Positive Exposure PE, *Probability of Default PD*, and *Loss Given Default LGD*. *LGD* is the percentage of PE which is actually lost by the investor at default. Instead of LGD one also uses the *Recovery Rate R*, denoting the percentage of PE the investor can still secure despite the default, i.e., $R = 1 - LGD$.

PE is the future present value of all future payments between the investor and the counterparty if this value is positive (for the investor), i.e. this present value nets into an amount receivable by the investor. Otherwise, PE is zero. That's because an amount *payable* to the counterparty is still payable even if the counterparty defaults, it becomes part of the bankruptcy assets. Therefore $PE = V^+(t, T) = \max(V(t, T), 0)$ where $V(t, T)$ is the present value at a future time T as seen from the (current) time t.

In real life, there will be a bankruptcy trustee who decides, how the assets of the defaulted entity are to be distributed among the creditors and how the amounts receivable of each creditor are to be calculated. This process is by no means unambiguous. For instance, the PE of a bond is often simply set to its residual debt without any discounting or considerations of future interest payments. It could very well be that the residual debt is larger than the present value of all open future cash flows. In such a scenario, the investor could even make an additional profit out of the default, at least if there are still enough assets available to satisfy the residual debt payment.[2] This simple approach doesn't work for swaps, however, since there is no sensible way to define a residual debt for a swap. Thus, the only reasonable way is to calculate PE as the present value of the swap—which can very well be negative. Since at the end of the day the only relevant number is the realized loss in case of default $LGD \cdot PE$, it is only important that the reference of LGD (present value or residual debt) is consistent with the definition of PE. In the following we will always determine the exposure based on the future present value $PE(T) = V^+(t, T)$, since we can avoid above mentioned inconsistencies and the present value is well defined for all financial instruments.

In general, PE, PD and LGD are all time-dependent. The default probability always refers to a time span as opposed to a single point in time

---

[2]Especially for bonds with priority of claim before other debts, such a scenario is quite plausible.

(for which the default probability is zero). Nonetheless can we define the default probability over an infinitesimally short time span dt and denote it as $PD(t)\,dt$. This is the probability that the counterparty will default within the (infinitesimal) time span between t and $t+dt$. Therefore, $PD(t)$ is the default probability *density* (see Appendix A). Integrating this density over dt from T to maturity T' yields the probability $P(T, T'|T)$ that the counterparty will default during the future time span between T and T', given he did survive until T:

$$P(T, T'|T) = \int_T^{T'} PD(s)ds = 1 - Q(T, T'|T) \,. \qquad (20.1)$$

Here we have also introduced the survival probability $Q(T, T'|T)$, i.e., the probability that the counterparty survives the time span from T to T', given he survived until T. Those are all *conditional* probabilities, since they are dependent on the condition that the counterparty survives until the (future) time T. We arrive at the *un*conditional probabilities $P(T, T')$ and $Q(T, T')$ by multiplying with the probability to survive from t (i.e. today) until T:

$$
\begin{aligned}
P(T, T') &= Q(t, T)P(T, T'|T) = Q(t, T)(1 - Q(T, T'|T)) \\
&= Q(t, T) - Q(t, T') \\
Q(T, T') &= Q(t, T)Q(T, T'|T) = Q(t, T') \,. \qquad (20.2)
\end{aligned}
$$

Thus, the unconditional probability for a default between T and T' is the probability that the counterparty survives until T but not until T', while the unconditional probability for a survival between T and T' is simply the same as the survival probability from t (now) until T'. Survival probabilities $Q(t, T)$ are mathematically very similar to discount factors $B(t, T)$: Both are always lower[3] (or equal) to 1 and both fulfill the 'product rule' Eq. 2.6 and Eq. 20.2 for Forward-Zerobonds and Forward-Survival-Probabilities. Therefore, like we wrote the discount factor as a function of the instantaneous forward rate $f(t)$ in Eq. 14.3

$$B(t, T) = e^{-\int_t^T f(s)ds}$$

---

[3] For positive interest rates.

we can now write the survival probability in the same form, thereby defining the so-called *Hazard-Rate* $h(t)$:

$$Q(t, T) = e^{-\int_t^T h(s)ds} \tag{20.3}$$

Differentiating Eq. 20.1 w.r.t. T' now yields the default probability density as a function of the hazard rate

$$PD(s) = \frac{\partial P(t, s)}{\partial s} = -\frac{\partial Q(t, s)}{\partial s} = h(s)e^{-\int_t^s h(u)du} \,.$$

This analogy between discount factors and survival probabilities will prove to be quite useful later on.

After these preparations we can now determine the present value of credit risk, the so-called *Credit-Value-Adjustment*:

$$\begin{aligned}
CVA(t) &= E\left[\mathbf{1}_{\{T_D < T\}} B(t, T_D) LGD(T_D) V^+(t, t_D)\right] \\
&= E\left[\int_t^T \mathbf{1}_{\{s = T_D\}} B(t, s) LGD(s) V^+(t, s)\, ds\right] ,
\end{aligned} \tag{20.4}$$

where T_D denotes the time of default and $\mathbf{1}_{\{.\}}$ equals 1 only if the condition in the brackets is true, and equals 0 otherwise. *CVA* is the present value of the expected loss due to a default of the counterparty. For $T_D = s$ the present value of the loss is just $B(t, s) LGD(s) V^+(t, s)$. The expectation in Eq. 20.4 has to be calculated w.r.t. all risk factors on which the quantities in the integral depend on. Those are (besides default and interest rate risks) all other risks influencing the exposure $V^+(t, s)$. If these risks were independent from the default risk, we could simplify the expression substantially. However, in general such independence cannot be assumed. For instance, $V^+(t, s)$ can very well be dependent on interest rates and thus on $B(t, s)$. Also, a default might be more probable with high interest rates (or vice versa). Such dependencies then also imply a correlation between exposure and default probability. A situation where the default probability and the exposure both have a positive correlation with a risk factor is called *Wrong-Way-Risk*, since the more probable a default becomes, the higher the exposure gets. The opposite situation, where exposure and default-probability are anti-correlated, is called *Right-Way-Risk*.

For the special case that the default risk is the only risk factor, the expectation simply becomes an integral over the default probability density $PD(t)$

$$CVA(t) = \int_t^T PD(s)B(t,s)LGD(s)V^+(t,s)\,ds \;, \qquad (20.5)$$

The (default-)risky present value of a financial instrument is the default-risk-free present value minus the *CVA*. Looking at the transaction from the other side, one could of course also calculate the *CVA* of the transaction as seen by the counterparty. The *CVA* of the counterparty is called *Debt-Value-Adjustment* or *DVA* and involves ones own default risk. Nowadays, the consideration of *CVA* and *DVA* is a necessity required by the supervisory authorities for calculating capital requirements. The topic is discussed in many publications [22, 33, 79].

20.2 Measures To Reduce Default Risk

With every business transaction runs the risk that the counterparty won't fulfill her obligations. That risk cannot be eliminated. Even though one might not be able to influence the probability that a counterparty will default, there are nonetheless means to reduce the impact of such a default on ones own situation.

20.2.1 Collateral Management

One such measure is the allocation of collateral, similar to what we have for futures contracts in Sect. 3.3.2. There, the margin account was the collateral. In case of a default, the exchange could use the money in margin account of the defaulting counterparty to (almost) completely negate any financial impact of that default. Also in interbank trading it is by now common practice to sign bilateral Collateral Service Agreements (CSAs) about the exchange of *collateral*, be it as cash collateral or as security papers. To determine which counterparty has to provide collateral, and to what amount, both counterparties regularly (daily or weekly) perform portfolio evaluations to determine the present values of all open transactions between the counterparties. If the counterparties can not agree on the present value (and thus on the necessary collateral), they have a so-called *Margin Dispute* triggering an escalation procedure which might even become reportable to the authorities.

20.2.2   Central Counterparties

It is the goal of the regulators to institutionalize this collateral process between banks for the most important (by volume) derivatives. *Central Counterparties (CCPs)* have been set up for this purpose. Interest rate swaps must now be transferred to a mandatory CCP. If Bank A swaps with Bank B, that swap will be replaced by two trades, one from Bank A with the CCP and another from Bank B with the CCP. The CCP thus acts as an intermediary and ensures strict compliance to the provision of collateral. How much collateral must be provided is determined by the CCP. This will reduce the uncertainties of bilateral collateral management.

20.2.3   Netting Agreements

Another important measure to reduce the risk of default is entering into a *Netting Agreement*. All trades of a specific portfolio (the *netting set*) between two counterparties can then be offset against each other based on their present values. It often happens that a bank A has both liabilities and receivables from derivatives transactions with bank B. If B defaults and a netting agreement between the two banks exist, all receivables and liabilities be offset against each other. The exposure of A to B is thus reduced by A's liabilities. Without the netting agreement, all receivables would be lost, but the liabilities would remain in full.

20.2.4   Hedging

A third way to reduce default risks is to hedge against the default by entering into trades with opposite counterparty default risk. In particular, credit default swaps are commonly used (see Sect. 20.6).

20.3   Bonds with Default Risk

In the following we will use Eq. 20.4 to calculate the present value of a coupon bond taking default risk into account. We assume interest rates to be deterministic so we can use Eq.20.5. We also observe that the (default-)risk-free present value of a bond is always positive from the perspective of the bond

investor, i.e., $V^+(t, T) = V(t, T)$. With Eq. 15.6 and Eq. 15.7 we find the Credit-Value-Adjustment $CVA_{\text{Bond}}(t)$ of the bond:

$$
\begin{aligned}
CVA_{\text{Bond}}(t) &= \int_t^T B(t, s)PD(s)LGD(s)\,(NKA(s, t_0, T) + NB(t, T))\,ds \\
&= N \int_t^T PD(s)LGD(s) \left(K \sum_{i=m+1}^{n} B(t, t_i)\tau(t_{i-1}, t_i)\mathbf{1}_{s < t_i} + B(t, T) \right) ds \\
&= NK \sum_{i=m+1}^{n} \tau(t_{i-1}, t_i)B(t, t_i) \int_t^{t_i} PD(s)LGD(s)ds \\
&\quad + NB(t, T) \int_t^T PD(s)LGD(s)ds .
\end{aligned}
\tag{20.6}
$$

Here, we used $B(t, s)B(s, t_i) = B(t, t_i)$ and moved the sum in front of the integral since $\mathbf{1}_{s < t_i}$ is the only factor in the sum which depends on s and ultimately only amounts to an adjustment of the integral's upper limit. Without the sum the bond becomes a zero bond and we obtained from Eq. 20.6 the credit value adjustment of a zero bond:

$$
CVA_{\text{Zerobond}}(t) = B(t, T) \int_t^T PD(s)LGD(s)ds .
\tag{20.7}
$$

This expression still applies if discount factors are not assumed to be deterministic, but still independent of the default probabilities. For the zero bond with default risk (or risky zerobond) $\bar{B}(t, T)$ we get:

$$
\begin{aligned}
\bar{B}(t, T) &= B(t, T) - CVA_{\text{Zerobond}}(t) \\
&= B(t, T) - B(t, T) \int_t^T PD(s)LGD(s)ds \\
&= B(t, T) \left(1 - \int_t^T PD(s)LGD(s)ds \right) .
\end{aligned}
\tag{20.8}
$$

The value of a risky coupon bond can now be very generally expressed in terms of risky zero bonds:

$$
\bar{V}_{\text{Bond}}(t) = V_{\text{Bond}}(t) - CVA_{\text{Bond}}(t) \qquad = NK\bar{A}(t, t_m, t_n) + N\bar{B}(t, t_n) .
\tag{20.9}
$$

The risky annuity $\bar{A}(t, t_0, t_n)$ follows from Eq. 15.7 by replacing all risk-free zero bonds with risky zero bonds. Similarly, replacing all zero bonds in Eq. 15.6 with the corresponding risky zero bonds yields the present value of a coupon bond with default risk. In other words, default risk is taken into account by simply using the risky yield curve for discounting instead of the risk-free yield curve. Strictly speaking, the risky yield curve is not a yield curve in the true sense of the word since it contains default risk components.

Often, a time-independent $LGD(t) = LGD = $ const. is used. Equation 20.8 then simplifies with Eq. 20.1 to:

$$\bar{B}(t, T) = B(t, T)\,(1 - LGD\,P(t, T))\;.$$

This is a very intuitive result: The value of the risky bond is simply the risk-free value minus the amount lost in case of default weighted with the probability of that default. Expressed in terms of the recovery rate $R = 1 - LGD$ and the survival probability $Q(t, T) = 1 - P(t, T)$ the equation reads

$$\bar{B}(t, T) = B(t, T)\,(R + (1 - R)Q(t, T))\;. \qquad (20.10)$$

This is also very intuitive: The value of the risky bond is the risk free present value of the recovery rate (which you get in any case—even in case of default) plus the amount $1 - R$, which you only get if the counterparty survives, weighted with the probability of that survival.

For the special case of a vanishing recovery rate $R = 0$ we introduce $\tilde{B}(t, T)$ as

$$\tilde{B}(t, T) = B(t, T)Q(t, T) \qquad (20.11)$$

Because of the similar formal structures of zerobonds and survival probabilities mentioned above, explicitly expressed in Eq. 2.6 and 20.2, a simple forward-price-rule is also valid for $\tilde{B}(t, T)$:

$$\tilde{B}(T, T'|T) = B(T, T'|T)Q(T, T'|T)\;.$$

The relationship between $\bar{B}(t, T)$ and $\tilde{B}(t, T)$ is simply

$$\bar{B}(t, T) = B(t, T)R + \tilde{B}(t, T)(1 - R) = \tilde{B}(t, T) + R(B(t, T) - \tilde{B}(t, T))\;.$$

Thus, if the survival rate is greater than zero, the value of a risky zero bond is increased by $R(B(t,T) - \tilde{B}(t,T))$.

20.4 Credit Spreads

Because of $0 \leq R + (1-R)Q(t,T) \leq 1$ for the second factor in Eq. 20.10, $\bar{B}(t,T) \leq B(t,T)$ is always satisfied. Therefore, any value of $\bar{B}(t,T)$ can be represented by a Zerobond for which the risk-free interest rate used to calculate $B(t,T)$ is increased by the so-called *Credit Spread*.[4] For example, if $r(t,T)$ is the risk free continuous compounding interest rate from t to T with $B(t,T) = e^{-r(t,T)\tau(t,T)}$, then the present value $\bar{B}(t,T)$ can always be represented by a time-dependent credit spread $s(t,T)$:

$$\bar{B}(t,T) = e^{-(r(t,T)+s(t,T))\tau(t,T)} = B(t,T)e^{-s(t,T)\tau(t,T)} \ .$$

Equation 20.10 can be used to determine the relationship between $s(t,T)$ on one side and R and $Q(t,T)$ on the other side.:

$$e^{-(r(t,T)+s(t,T))\tau(t,T)} = e^{-r(t,T)\tau(t,T)}(R + (1-R)Q(t,T))$$

$$\Rightarrow s(t,T) = -\frac{1}{\tau(t,T)}\ln(R + (1-R)Q(t,T)) \ .$$

The credit spread can also be expressed as a spread over interest rates in other compounding methods. For instance in discrete annual compounding

$$(1 + r(t,T) + s(t,T))^{-\tau(t,T)} = (1 + r(t,T))^{-\tau(t,T)}(R + (1-R)Q(t,T))$$

$$\Rightarrow 1 + r(t,T) + s(t,T) = (1 + r(t,T))(R + (1-R)Q(t,T))^{-\frac{1}{\tau(t,T)}}$$

$$\Rightarrow s(t,T) = (1 + r(t,T))\left((R + (1-R)Q(t,T))^{-\frac{1}{\tau(t,T)}} - 1\right) \ .$$

The credit spread now also depends on the risk-free interest rate, and is therefore no longer a sole function of recovery rate and survival probability. In practice, depending on the application, very different quantities are referred to as credit spreads. One should therefore always insure in each case which

---

[4]Strictly speaking, the limit $\bar{B}(t,T) = 0$ in the case of $Q(t,T) = 0$ and $R = 0$ is excluded here. However, this borderline case of a sure default is not of practical relevance.

definition is meant. Let us stick to the definition of credit spreads in continuous compounding. For the special case $R = 0$, the survival probability $Q(t, T)$ depends on the credit spread $s(t, T)$ in the same way as the zero bond on the risk-free rate:

$$Q(t, T) = e^{-s(t,T)\tau(t,T)} \ .$$

The credit spread $s(t, T)$ in this form can be represented as function of the hazard rate $h(t)$

$$s(t, T) = \frac{1}{\tau(t, T)} \int_t^T h(u)du \ .$$

This is analogous to the representation of zero interest rates as a function of the instant forward rates in Eq. 14.4. This connection is used in particular if the default risk is modelled as a stochastic process within the framework of a hazard rate model.

20.5 Credit Spread Risk

If the default risk of a counterparty is expressed by a credit spread instead of survival probability and recovery rate, information is lost. For example, two different bonds of the same issuer might have the same default probability, but the credit spreads are nonetheless different if the recovery rates for both bonds are different. Trying to deduct credit spreads from bond prices traded in the market, runs into further difficulties like, e.g., different liquidities for different emissions. Nevertheless, credit spreads are a useful tool to estimate the credit risk of bonds and loans. Since a credit spread is simply a premium on the risk-free interest rate, modified duration and convexity of a bond or loan are well suited to estimate credit risk (understood as *Spread-Risk*). Because present values can be determined by simply replacing risk-free zero bonds with risky zero bonds, the same sensitivities to estimate interest rate risk can also be used to estimate spread risk.

The present value of a loan changes only if the risky yield curve used to value that loan changes. In contrast to a risk-free valuation, there is now additionally the influence of the *credit rating*: If the credit rating of the debtor deteriorates, the *spread* between the risk-free yield curve (which carries pure interest rate risk) and the risky yield curve for the bond increases, i.e., the rates used for discounting increase and the present value decreases according to Eq. 20.9. As

with the interest rate risk, the spread risk can therefore be attributed solely to the sensitivity to changes in a yield curve. This change in the yield curve, however, now has two causes:

- Change in risk-free interest rate
- Change in creditworthiness and thus in the credit spread.

The (credit) risk is therefore regarded here as a spread risk. Note, that for this credit risk no *default* (bankruptcy) of the debtor is necessary. A deterioration in creditworthiness, e.g. downgrade from *rating* AA to B, can significantly increase the spread and means a correspondingly large present value loss for the loan. The *counterparty default risk*, i.e. the debtor's actual inability to pay, may therefore be viewed as an extreme special case of a deterioration in creditworthiness. Often, however, credit spread risks are attributed to the market risk (because market prices are directly affected by the credit spreads) and are clearly distinguished from the actual default risk (which is considered credit risk).

A private investor following a *Buy-And-Hold-Strategy* might now argue that a deterioration in credit rating (e.g., from AA to B) is not relevant to him, as long as the debtor continues to make all interest and principal payments on time, and that therefore only the actual default represents a risk. However, if the investor changes his mind at a later date and wants to close the position prematurely, he will be affected by the deteriorated creditworthiness, as it will have a correspondingly negative impact on the market price of the position.

Institutions that are subject to supervisory regulations (e.g. banks, insurance companies and fund managers), have to reserve capital for the default risk in their positions. And they have to reserve the more capital, the worse the credit ratings of their debtors. (see also Sect. 21.1). Spread risks are therefore of importance even for the classical lending business.

If an increased expectation of losses, expressed by an increased credit spread, is realised, a bank with a large bond or credit portfolio will on average realise losses due to defaults which are an average exactly equal to the present value losses incurred by the deterioration in creditworthiness. The higher credit spread after a deterioration in creditworthiness is therefore nothing else than a reflection of the increased probability of default.

Although the credit spread is time-dependent, it is often assumed that the credit spread is either constant or has a very simple parametric form, such as a straight line $s(t, T) = a + m\tau(t, T)$. For example, a constant credit spread (i.e. $m = 0$ or $s(t, T) = a$) corresponds to an increase of the whole risk-free yield curve by the constant spread, i.e. a parallel shift of the whole risk-free

yield curve. These simple approaches are often necessary, because for many issuers there is not enough liquid, independent market data (e.g. bond prices) available to construct a reliable term structure of the credit spreads. Often, however, the dominant effects are already represented well enough by such simple parameterizations.

20.6 Credit-Default-Swaps

Credit-Default-Swaps represent the most important instrument for trading and hedging default risks. In spite of its name a credit default swap or CDS is more of an insurance policy than a standard swap. The buyer of a swap insures himself against the risk of default of a specific issuer or counterparty, often simply called *reference address*, *address* or *name*. For this purpose, he regularly pays a fixed premium (the CDS premium or the CDS spread). If the reference address defaults, the CDS ends and the buyer receives a compensation payment, either in the form of a predefined fixed amount (Digital CDS) or, more frequently, the LGD amount of a reference bond, which was previously agreed upon in the CDS contract. Such a reference bond can, for example, be a bond emitted by the reference address. In return, the buyer pays the pro rata CDS premium for the current premium period up until the default event. The buyer of a CDS is also referred to as a *protection buyer*, the seller correspondingly as *protection seller*.

Credit-Default-Swaps fulfill two different purposes. On the one hand, for the purpose of diversification, the guarantor can selectively take credit risks on counterparties that do not belong to its business partners. The other more important purpose is the hedging of credit risk arising from other positions, such as a bond or swap, the issuer or counterparty of which is the reference address of the CDS. This possibility of hedging default risks makes it possible to assess credit risks on a risk-neutral basis, as this is the only way to replicate a position with default risk using a self-financing strategy (see Sect. 13.3).

What is considered a *default* for a CDS, must be contractually agreed. Typically, these are the following *credit events*:

- Initiation of bankruptcy proceedings
- Payment default (of the reference bond or another public bond)
- Restructuring of the company (provided that the reference address is a company)
- Takeover of the reference address by another company
- Rating change.

Bankruptcy and payment default are always relevant in practice, the other credit events listed are rather optional. Often, it takes a while until the actual damage, i.e. the LGD, can be determined. Therefore, a *Grace Period* is usually provided for the payment of the compensation after a credit event. This period might, for example, end 90 days after the credit event becomes known.

When closing a CDS, the buyer pays an *upfront fee*. This is one of the most important conventional changes introduced with the *big bang* in April 2009 in order to facilitate premature closing of a CDS position. Unlike a standard swap, for example, or before the big bang, the CDS premium is not set so that the CDS has a present value of 0 when entering the contract. Instead, only certain premiums are traded. As a result, the CDS has a present value not equal to 0 upon conclusion, which is compensated by the upfront fee. In fact, this fee may also become negative, in which case it will be paid by the seller of the CDS. Typical premium levels are 100 and 500 basis points for the North American market and 25, 100, 500 or 1000 basis points for the European market. Since a credit event does not necessarily become public immediately, the period for a filed bankruptcy proceeding to be relevant for the CDS has been brought forward to 60 days *before* the date of closing. For takeovers there is even a deadline of 90 days prior to the CDS conclusion. Otherwise, the insurance cover applies from the first day after the conclusion of the transaction.

In connection with a CDS, the default risk generally refers to the default risk of the reference address. In fact, however, a CDS involves three different default risks because the reference address and either of the two CDS counterparties may fail. For the buyer of the CDS, in particular the scenario that first the seller of the CDS and then (or simultaneously) the reference address defaults (*Double Default*) is associated with considerable losses. However, it is immediately plausible that a buyer of a CDS will only enter into a CDS with counterparties whose default risks (at least upon entry into the contract) are significantly lower than that of the reference address. In a first approximation, therefore, the default risk of the CDS counterparties can be neglected. The counterparty risk of the CDS seller is limited in any case to the *replacement risk*, as the seller is only liable in case of a default of the reference address and only when the premium has been paid in full. In interbank trading, it has become standard practice for CDS transactions to always be secured by collateral and in the near future they will even have to be processed via *Central Counterparties* (CCP) (see also Sect. 20.2). That is why in the following we will neglect the default risk of the two CDS counterparties.

20.6.1   Cashflows and Present Value

Even before the big bang, the CDS market was much more standardized than, for example, the interest rate swap market. Since 2003, coupon and premium periods have always started and ended on the 20th of the months March, June, September and December, with a fixed period length of 3 months (3M). The term of a CDS always counts from the next period end. If you enter a 6M-CDS on 19 September, it shall end on 20 March of the following year (20 September plus 6 months). If that deal is done one day later, i.e. on 20 September, the CDS ends only on 20 June of the following year (20 December + 6 months). Thus, the first period is normally shorter than 3 months and called *short stub*.[5]

As for coupon bonds (see Sect. 15.3.1), we denote the dates for the premium payment with t_i with $i = 0, 1, 2, \ldots, n$, where t_0 is the start date of the CDS. The latest premium date before t is denoted by t_m with $t_m < t < t_m+1$. Thus, $m = 0$ applies for a CDS just entered into. For a nominal N and a CDS spread (or premium rate) K we can then calculate a single premium $C_P(t_i)$ as

$$C_P(t_i) = N K \tau(t_{i-1}, t_i)$$

For the last premium $C_P(t_n)$ the following special feature applies

$$C_P(t_n) = N K \tau(t_{n-1}, t_n + 1\mathrm{D}) ,$$

since a default on day t_n is still insured and therefore the premium for this day still has to be paid. For interest calculations, on the other hand, the first day of the period is counted, but the last day is not. Accordingly, the various day count conventions are defined. In the following, we will write $\tau_{\mathrm{cds}}(t_i, t_{i+1})$ for this special calculation of the day count conventions.

The present value of a single premium payment of a CDS is dependent on the time of default T_D of the reference address. If T_D is before the start of the period t_{i-1} then the present value is zero. If the default occurs within the coupon period, $t_{i-1} \leq T_D \leq t_i$, then the present value equals $N K B(t, T_D) \tau_{\mathrm{cds}}(t_{i-1}, T_D)$. If the reference address does not default before the end of the period, $T_D > t_i$, the present value corresponds to the full

---

[5]Before the big bang things were a little more complicated. If the next coupon date was less than 30 days in the future, no premium was paid on that date. Instead, the premium for the first (short) period and the first full period was paid at the end of the first full period (*long stub*).

premium $NKB(t, t_i)\tau_{\text{cds}}(t_{i-1}, t_i)$. In summary, the present value V_i of the i-th premium payment can be written as

$$V_i = E\left[NKB(t, \min(t_i, T_D))\tau_{\text{cds}}(t_{i-1}, \min(t_i, T_D))\mathbf{1}_{\{T_D \geq t_{i-1}\}}\right] \ .$$

With the simplified assumption that interest rates and default probabilities are independent of each other, the expected value can be calculated as an integral over the default probability density.:

$$V_i = NK \int_{t_{i-1}}^{\infty} B(t, s)\tau_{\text{cds}}(t_{i-1}, s)PD(s)\,ds$$

$$= NK \underbrace{\int_{t_{i-i}}^{t_i} B(t, s)\tau_{\text{cds}}(t_{i-1}, s)PD(s)\,ds}_{\text{accrued premium}} + \underbrace{NKB(t, t_i)\tau_{\text{cds}}(t_{i-1}, t_i) \int_{t_i}^{\infty} PD(s)\,ds}_{\text{full premium}}$$

Because of Eq. 20.11 and $\int_t^{\infty} PD(s)\,ds = 1$ and

$$Q(t, t_i) = 1 - P(t, t_i) = \int_t^{\infty} PD(s)\,ds - \int_t^{t_i} PD(s)\,ds = \int_{t_i}^{\infty} PD(s)\,ds \ ,$$

the term for the full coupon simplifies to:

$$V_i = NK \int_{t_{i-i}}^{t_i} B(t, s)\tau_{\text{cds}}(t_{i-1}, s)PD(s)\,ds + NK\tilde{B}(t, t_i)\tau_{\text{cds}}(t_{i-1}, t_i) \ .$$

The present value of all premiums of the CDS is then the sum:

$$V_{\text{premium}} = NK \sum_{i=1}^{n} \tilde{B}(t, t_i)\tau_{\text{cds}}(t_{i-1}, t_i)\mathbf{1}_{\{t < t_i\}}$$

$$+ NK \sum_{i=1}^{n} \int_{t_{i-i}}^{t_i} B(t, s)\tau_{\text{cds}}(t_{i-1}, s)PD(s)\,ds \ .$$

With the risky CDS annuity $\tilde{A}_{\text{cds}}(t, t_0, t_n)$ given as

$$\tilde{A}_{\text{cds}}(t, t_0, t_n) = \sum_{i=1}^{n} \tau(t_{i-1}, t_i)\tilde{B}(t, t_i)\mathbf{1}_{\{t < t_i\}}$$

the first term can also be represented in terms of this annuity:

$$V_{\text{premium}} = NK\tilde{A}_{\text{cds}}(t, t_0, t_n) + NK\sum_{i=1}^{n}\int_{t_{i-i}}^{t_i} B(t, s)\tau_{\text{cds}}(t_{i-1}, s)PD(s)\,ds\ .$$

The payment of the CDS seller in the event of a default of the reference address is a one-time payment in the amount of LGDs as a percentage of the nominal N. But then the present value of this payment just corresponds to the CVA of a constant PE of N. With Eq. 20.4 it now follows that:

$$V_{\text{default}} = \text{E}\left[B(t, T_D)LGD(T_D)N\mathbf{1}_{\{T_D \leq T\}}\right]\ .$$

If the LGD is independent of default probability and interest rates, the expected value of the LGD can be split off and expressed by the expected value R of the recovery rate $R(t)$:

$$V_{\text{default}} = N\text{E}[LGD(T_D)]\,\text{E}\left[B(t, T_D)\mathbf{1}_{\{T_D \leq T\}}\right]$$
$$= N(1 - R)\int_{t}^{T} B(t, s)PD(s)\,ds\ .$$

In the second step, we again used the assumption of independence from interest rates and default probabilities.

Overall, from the perspective of the protection buyer, the CDS has the following value:

$$V_{\text{CDS}} = N(1 - R)\int_{t}^{T} B(t, s)PD(s)\,ds - NK\tilde{A}_{\text{cds}}(t, t_0, t_n)$$
$$- NK\sum_{i=1}^{n}\int_{t_{i-i}}^{t_i} B(t, s)\tau_{\text{cds}}(t_{i-1}, s)PD(s)\,ds\ . \tag{20.12}$$

20.6.2 Approximations

The integrals in 20.12 cannot be solved analytically without additional approximations or assumptions. A frequently used approximation is, for example, that a default always occurs in the middle of an interest period and the settlement payment is made at the end of the period in which the default

occurs. With these additional assumptions, an approximation for the present value of the CDS can be determined analytically:

$$V_{\text{CDS}} \approx N(1-R) \sum_{i=1}^{n} B(t, t_i)(Q(t, t_{i-1}) - Q(t, t_i))\mathbf{1}_{\{t < t_i\}} - NK\tilde{A}_{\text{cds}}(t, t_0, t_n)$$

$$- \frac{NK}{2} \sum_{i=1}^{n} B(t, t_i)\tau_{\text{cds}}(t_{i-1}, t_i)(Q(t, t_{i-1}) - Q(t, t_i))\mathbf{1}_{\{t < t_i\}} \,.$$

$$(20.13)$$

The ISDA CDS Standard Model, which is also used by some market data providers to quote CDS spreads, works with fixed conventions for the presentation of discount factors and survival probabilities. Specifically, the model is based on the assumption that yield curves are represented as piecewise constant forward curves. Analogous to this, survival probabilities are modeled as piecewise constant Hazard rates (see also Sect. 29.9.1). Then the integrals can be calculated analytically. For this model there is a freely accessible program source code available in the programming language C [99, 189].

20.6.3 Par CDS Rate

Before the big bang, CDSs were always quoted by specifying the par CDS rate, i.e. the premium for which the present value is zero. Nowadays, only certain fixed premiums are traded, so that the CDS already has a non-zero present value at the time of conclusion, which is offset by an up-front fee (see above). However, the Par CDS rate is often still provided as additional information. Setting the present value in Gl. 20.12 to zero and solving for K yields the par CDS rate K_{par} as:

$$K_{\text{par}} = \frac{(1-R)\int_t^T B(t, s)PD(s)\,ds}{\tilde{A}_{\text{cds}}(t, t_0, t_n) + \sum_{i=1}^{n} \int_{t_{i-i}}^{t_i} B(t, s)\tau_{\text{cds}}(t_{i-1}, s)PD(s)\,ds} \,,$$

$$(20.14)$$

or with the approximations of Eq. 20.13:

$$K_{\text{par}} \approx \frac{(1-R)\sum_{i=1}^{n} B(t, t_i)(Q(t, t_{i-1}) - Q(t, t_i))\mathbf{1}_{\{t < t_i\}}}{\tilde{A}_{\text{cds}}(t, t_0, t_n) + \frac{1}{2}\sum_{i=1}^{n} B(t, t_i)\tau_{\text{cds}}(t_{i-1}, t_i)(Q(t, t_{i-1}) - Q(t, t_i))\mathbf{1}_{\{t < t_i\}}} \,.$$

$$(20.15)$$

Part IV

Risk

21

Fundamentals

In general, the term *risk* signifies the possibility that some future event might have some negative consequences. Since the future is uncertain, the term risk is tightly connected with the probability or likelihood that an uncertain future event actually becomes real. The magnitude of risk depends besides the probability of this event also on the magnitude of the damage done if the event actually becomes real. *Risk management*, i.e. controlling and minimizing risk, is an important task in any corporation. This is true especially for financial institutions, not only because they face very significant and large risks, but also because it is indeed a substantial part of their business model. This is, for example, apparent for an insurance company, but it is also the core business of a bank to take over risks, especially *market price risk* (brief market risk) and *credit risk*, which we will consider in the following in more detail. As potential damage caused by a "negative event" we will consider only the potential financial loss. Such risks caused by risk factors connected to specific single trades may be reduced or eliminated by entering into hedge trades sensitive to the same risk factors, but with opposite sign. The total risk of a portfolio, and therefore also the potential financial loss, could be largely reduced.

Market price risk is caused by the uncertainty of the future development of *market risk factors*. These include prices of shares, bonds or commodities, but also other parameters with a direct impact on the fair value of financial products, like FX or interest rates, inflation, volatilities or correlations between other market risk factors. Depending on one's portfolio, the risk could be that a certain risk factor (e.g. a share price) rises, falls, or stays constant. A long position in a risk factors indicates that the value of the portfolio increases, if the

risk factor increases. Vice versa, with a short position the value of the portfolio decreases if the risk factor increases, e.g. a short sell, where the seller didn't own the shares and needs to repurchase them at some point in the future.[1] A risk factor that stays constant could also be risky, e.g. if the portfolio consists of a put option with strike K_P and a call option with strike K_C where $K_P < K_C$ on the same underlying asset. Only if the asset price at option maturity is greater than K_C or lower than K_P, a total loss of the previously paid option premiums could be avoided.

Therefore, it is not correct to identify risk with the possibility of market risk factor fluctuations or the volatility of these factors as is done occasionally. Risk depends always on the concrete positions within a given portfolio of assets. Even, the strategy of buying a put and a call with the same strike and same maturity is important enough to have a special name: it's called a *straddle*.

Credit risk denotes the potential loss caused by the failure to pay or bankruptcy of a trading partner (*counter party risk*) or issuer of securities (*issuer credit risk*). Both types of risk are summarized as credit risk or default risk. On the other hand, risk caused by a mere diminishing of some party's credit worthiness, i.e. *credit spread risk* is usually considered as just another type of market risk.

Credit spread risk, that is the risk associated with a changing credit spread, i.e. the add-on a party has to pay over risk free interest rate for a loan or bond with correspondence level of creditworthiness, is most often considered as market price risk rather than credit risk. In this view, the credit spread is just an additional market risk factor. On the other hand, the risk associated with a change in the credit rating determined by some credit rating agency (*credit migration risk*) is seen as credit risk, though it could also be reflected in a jump of credit spreads, especially if it wasn't expected. More often, credit spreads change long before a down- or upgrade of the credit rating is pronounced. Here, the border between market and credit risk is blurred. Sometimes, the credit spreads are even modeled based on credit ratings.

Though market risk and credit risk used to be largely independent concepts (on one hand trading of financial instruments with market risk, on the other the classic loan business with credit risk), it is increasingly difficult to differentiate both areas clearly. On reason is that derivatives bear substantial credit risks (i.e. counter party default risk), too. But loans are also not free of market risk (e.g. caused by fixed interest rates and embedded options,

---

[1]Naked short selling, i.e. selling of shares without owning or borrowing (or at least ensuring that they can be borrowed) has been largely restricted in many countries, though it is still possible to sell shares that have been borrowed beforehand, e.g. by a security lending trade.

either explicitly agreed upon or enforced by law). In addition, by means of instruments like credit default swaps it has become possible to directly trade credit risk, either to hedge against credit risk from loans or other financial instruments or to actively invest in credit risk on names to which the bank would have no exposure as part of a diversification strategy. As a consequence, in risk management, it is more and more common to consider market and credit risk together.

Besides market and credit risk, there are other risk classes like *liquidity risk*, *operational risk*, or *legal risk*. Liquidity risk[2] are caused by insufficient liquidity of traded assets (i.e. the impossibility to trade large volumes of assets at any time). Liquidity is essential since it is a required assumption for risk neutral valuation (see item 3 in Chap. 4). In practice, it is often difficult to differentiate between liquidity risk and spread risk, since both risk classes have similar impacts (e.g., modeled as spread over the risk-free interest rate).

Operational risk deals with disruptions of required business processes, e.g., failure of IT systems, insufficient or malfunctioning processes. Legal risk denotes the risk, that contracts closed between two parties (e.g. to close a derivative trade) may not hold in court if either of the parties takes legal action. Here, in addition to financial losses, legal risk may induce loss of reputation, too. In the following we will consider market risk and credit risk only.

The actual financial loss of an asset at risk varies depending on the definition of the reference value the loss is referred to. A loss could be defined in terms of a loss in net present value (i.e. the total present value of all shares, securities, derivatives, cash account in given portfolio), a reduction of the capital (i.e. if the close-out value is lower than the capital invested originally), as relative loss in comparison with some benchmark (commonly used to measure performance of equity or fixed income funds), or as a reduction in book value (after an adjustment of balance sheet items, e.g. because of required depreciations). Though present values of financial assets have an increasingly direct impact on balance sheets, it is in general not possible, depending on the accounting rules, to identify changes in balance sheet items with respective changes of theoretical present values of assets. It is also common to differentiate between realized and unrealized.

A realized gain or loss could be illustrated as an "irreversible" gain or loss, because, for example, the loss position has been sold out already, where in contrast an unrealized loss could diminish, if the value of the position "recovers" over time. In practice, different accounting regulations and there

---

[2]Here, the term liquidity risk means *market liquidity risk* rather than the risk not to be able to pay.

interpretation could yield to very different split of a gain or loss into its realized and unrealized part. Also, it is possible to sell and buy back a position immediately at the same price to realize a gain or loss. Therefore, risk should always be related to potential total loss, regardless if its realized or unrealized (gains are not considered to be a risk at all).

Parameters, which have a significant impact on the reference value, are denoted as *risk factors*. Which risk factors are considered as consistent depends on the context. In the following, we will use the total current present value of a portfolio (in a specified currency) as our reference value. The portfolio may contain securities (shares, bonds, certificates, etc.), loans, derivatives, commodity and cash positions in various currencies. In this context, all risk factors are relevant which have an influence on the present values of the positions within the portfolio. For example, risk factors could be the 3 months LIBOR rate, the US$/Euro FX rate or the Dow Jones index.

The *risk* of a portfolio could be expressed by means of various key figures. Commonly used risk figures are *value at risk* (VaR), *credit value at risk* (CVaR), *expected shortfall* and *potential future exposure*. A portfolio's value at risk, is defined as the upper boundary of a potential loss over a (finite) time period δt, such that this limit is not violated with a probability of c. This probability is called *confidence level*. *Stress testing* is another important tool for risk analysis. Here, the change of the portfolio's present value under a couple of extreme scenarios is analyzed, where a certain scenario is specified by providing a value for each selected risk factor.

21.1 Regulatory Requirements

Financial services is a strictly regulated economic sector. Because of the complexity of this topic with all their regional specialties, and the constant reformation of the regulations, we can provide here only a rough sketch. Due to the global interconnections of the financial markets, regulations are essentially driven by a supra-national organization, the *Basel Committee on Banking Supervision* of the *Bank for International Settlements* or brief BIS. In reaction to the financial crises that began 2007, a new package of reformed regulations as been pushed forward in 2013 under the name "Basel III" that modified and extended the earlier Basel II standard. It was implemented in EU law in terms of the Capital Requirement Regulation (CRR) and the Capital Requirements Directive (CRD IV), both have been settled as of July 17th, 2013 in their current form. The CRR has become legally binding as of January 1st, 2014, while CRD IV needs to be implemented into national law. The

European regulatory authority, i.e. the *European Banking Authority* or EBA, is responsible for the concretisation of CRD IV and CRR requirements.

In Germany, CRD IV was, among other regulations, implemented by the *Solvabilitätsverordnung* (SolvV), supplementing the *Kreditwesengesetzt* (KWG), and by the *Mindestanforderungen an die Risikomessung* (MaRisk). While the SolvV regulates the amount of regulatory capital financial institutions are obliged to deposit as reserve for the risks they face (Basel III, pillar I), the MaRisk regulates the methodical, technical and procedural implementation of the internal risk controls (Basel III, pillar II). Besides the regulatory authorities' requirements, financial institutions must also obey accounting standards, either according to national accounting rules (in Germany, these are defined in the *Handelsgesetztbuch* or HGB) or international accounting rules (as specified by the International Accounting Standards Board or IASB and known as the International Financial Reporting Standards or IFRS).

The regulations for risk management and control are constantly modified adjusted, modified and extended. Therefore, we will abstain here from a detailed discussions of the regulations. Up-to-date information can be found on the homepages of the regulatory authorities, e.g. the Bundesamt für das Finanzwesen (BaFin) (Germany), the Finanzmarkaufsicht (FMA) (Austria), the Eidgenössische Finanzmarktaufsicht (FINMA), or the Financial Services Authority (FSA) (UK).

21.2 Confidence, Percentile and Risk

The value of a portfolio is a function of stochastic processes (the risk factors) and is thus itself a stochastic process. A *confidence interval* of an arbitrary stochastic variable X with density $p(X)$ is defined as the subset of the variable's range (the interval) attained with a previously specified probability c, called the *confidence*, or *level of confidence.* A level of confidence of 95% ($c = 0.95$) for example, implies that the probability that the value realized by the random variable lies outside the confidence interval is only 5%. Or equivalently, we can be 95% certain that a realization of the random variable will lie within the interval. Should the fluctuations about a mean value μ be of interest, the upper and lower bounds, a and $-a$, of a *symmetric confidence interval* can be determined as the values satisfying the condition

$$c \overset{!}{=} P(\mu - a < x < \mu + a) = \int\limits_{\mu-a}^{\mu+a} p(x)dx \ .$$

If, on the other hand, we want to know with a previously specified probability that the random variable will not fall below a certain value (the portfolio will not lose more than a certain amount of its value), we consider the *one-sided confidence interval* and determine the boundary a as that value satisfying the condition

$$c \overset{!}{=} P(x > a) = 1 - P(x \le a) = 1 - \int_{-\infty}^{a} p(x)dx \ . \qquad (21.1)$$

The *percentile* or *quantile* associated with a previously specified probability c is defined as the value Q_c such that the probability is c, that a random value x is less than or equal to Q_c:

$$P(x \le Q_c) = c \quad \Longleftrightarrow \quad \int_{-\infty}^{Q_c} p(x)dx = c \ .$$

This definition defines the percentile implicitly. The percentile can be determined by "inverting" the cumulative distribution function P:

$$Q_c = P^{-1}(c)$$

The boundary a of a one-sided confidence interval is thus precisely the $(1-c)$ percentile

$$c \overset{!}{=} P(x > a) = 1 - P(x \le a)$$
$$P(x \le a) = 1 - c$$
$$\Longrightarrow \quad a = Q_{1-c} = P^{-1}(1 - c) \ .$$

Using these definitions of confidence, confidence interval and percentile, the *value at risk* (VaR) of a financial instrument or a portfolio with value V can be precisely defined for a specified c. As mentioned above, the value at risk of a portfolio in a time span δt is intuitively the upper bound for the depreciation of the portfolio's value in this time period which will not be exceeded with a specified confidence c. Here, the random variable is the change in the value δV (not the change in the underlying risk factors!) and VaR is the (value of the) *loss* which will not be exceeded with confidence c.

$$c \overset{!}{=} P(\delta V > -\text{VaR}(c)) \ . \qquad (21.2)$$

The definition of the VaR can be expressed more explicitly if we use the *cumulative probability function* $\mathrm{cpf}_{\delta V}$ of the random variable δV instead of P.

$$c \stackrel{!}{=} 1 - \mathrm{cpf}_{\delta V}(-\mathrm{VaR}(c)) = 1 - \int_{-\infty}^{-\mathrm{VaR}(c)} \mathrm{pdf}_{\delta V}(x)dx , \qquad (21.3)$$

where $\mathrm{pdf}_{\delta V}(x)$ denotes the *probability density function* of the random variable δV. Now, we see that the negative VaR is precisely the $(1 - c)$ percentile of the distribution of δV:

$$-\mathrm{VaR}(c) = Q_{1-c}^{\mathrm{cpf}_{\delta V}}$$

or, equivalently,

$$\mathrm{VaR}(c) = -Q_{1-c}^{\mathrm{cpf}_{\delta V}} = -\mathrm{cpf}_{\delta V}^{-1}(1 - c) . \qquad (21.4)$$

In the discussion up to now, we have emphasized that the probability of the change in the portfolio's value δV (and *not* the risk factors and their probability distributions) defines the value at risk. However, changes in the portfolio's value are induced by the risk factors. In addition, it is not the usual practice to model the stochastic process δV or its distribution directly. Instead, the risk factors (as in pricing) are modeled as stochastic processes and the methods derived for pricing (see, for example Part II) are then used to calculate the value change δV and its distribution (and percentiles) induced by the changes in the risk factors.

Using the probability distribution of the *risk factor* instead of the probability distribution of the *portfolio* for the determination of the value at risk is only possible when V is a *monotonous* function of the risk factor process S. If, for example, V is a monotonously increasing function of S, we have

$$V(S) \le V(a) \Longleftrightarrow S \le a ,$$

and V lies below the confidence interval bound for the *portfolio* if and only if S lies below the confidence interval bound for the risk factor. The value at risk of the portfolio is then the difference between the current portfolio value and the value of the portfolio when the risk factor lies at the lower bound of the confidence interval for the *risk factor*. Likewise, for the case where the portfolio's value is a monotonously decreasing function, the value at risk of the portfolio is the difference between the current value of the portfolio and

the portfolio's value when the *upper boundary* of the confidence interval of the *risk factor* is attained. The upper bound $\tilde{a}$ of a one-sided confidence interval is defined analogously to Eq. 21.1 through

$$c \overset{!}{=} P(x < \tilde{a}) = \int_{-\infty}^{\tilde{a}} p(x)dx \ . \tag{21.5}$$

Thus, if V is a monotonous function of the risk factor, the value at risk of the portfolio is

$$\text{VaR}_V(c) = V(S) - \min\left\{V(S=a), V(S=\tilde{a})\right\} \tag{21.6}$$
$$= \max\left\{V(S) - V(S=a), V(S) - V(S=\tilde{a})\right\} \ .$$

The minimum function guarantees that precisely *that* boundary of the confidence interval is chosen which results in the greatest *loss* for the portfolio. Hence, the one-sided confidence interval for the risk factor which is bounded from *above* determines the value at risk for instruments (or portfolios) whose value *declines* with increasing underlying prices. Analogously, the one-sided confidence interval with a *lower* bound determines the value at risk for those instruments (or portfolios) whose prices *rise* with the rising price of the underlying. The values of many instruments like bonds, futures, and most options are monotonous functions of their risk factors. Instruments (or portfolios) which are *not* monotonous functions of their underlyings must be either stripped into their component elements which are monotonous functions of their underlyings or the distribution of V itself and its confidence interval must be determined directly. The confidence intervals of the *risk factors* are no longer of assistance in this case.

21.2.1   Other Risk Measures

Besides the Value at Risk there are other concepts used in the market to quantify risk. Two risk measures mainly used in the asset management industry are *shortfall probability* and *expected shortfall*. Expected Shortfall is also relevant for credit risk measurement and becomes increasingly important due to recent changes of regulatory capital requirements for managing market risk.

Value at Risk has the disadvantageous property that no statement about the potential loss outside the confidence level is made. In contrast, Expected Shortfall is directly linked to the potential loss.

The expected shortfall for a given confidence level c is defined as the expected loss, given the loss is larger or equal to the Value at Risk. It is therefore a conditional expectation:

$$\text{Expected Shortfall}(c) = -\int_{-\infty}^{-\text{VaR}(c)} \text{pdf}_{\delta V}(x)\, x\, dx \ .$$

The *shortfall probability* is the probability, that the change in portfolio value is below a pre-specified amount. In a way, this is complementary to the Value at Risk: for VaR we are *given* the *probability* and calculate the loss, which will not be exceeded with this given probability; for Shortfall Probability we are *given* the *loss* and calculate the probability for this (or a larger) loss to occur. In other words: the Shortfall Probability belonging to a given loss equal to the VaR is simply 1 minus the confidence level belonging to that VaR, or with Eq. 21.3

$$\text{Shortfall Probability}(\text{VaR}) = \int_{-\infty}^{-\text{VaR}} \text{pdf}_{\delta V}(x)dx \ .$$

21.3 The Value at Risk of a Single Risk Factor

In order to illustrate the practical application of the concepts just introduced, we now consider a single risk factor S whose (infinitesimal) changes are governed by a *geometric Brownian motion* (*GBM*) as given in Eqs. 2.17 and 2.25. In other words, the risk factor satisfies the stochastic differential equation

$$d \ln S(t) = \mu dt + \sigma dW \text{ mit } dW \sim \text{N}(0, dt) \ . \tag{21.7}$$

The evolution of S in a *finite* time interval of length δt is the solution to the stochastic differential equation 21.7. This solution has already been given in Eq. 2.28. It is a stochastic process which describes the distribution of the market parameter $S(t + \delta t)$ a finite time step δt later:

$$S(t + \delta t) = S(t) \exp\left(\mu \delta t + \sigma \delta W\right) \text{ mit } \delta W \sim \text{N}(0, \delta t) \ . \tag{21.8}$$

The volatility σ of the risk factor appearing here is *theoretically* the same volatility as appears in the differential equation, Eq. 21.7. In practice, it is usually calculated as the standard deviation of the relative price changes over a finite time span δt (for example, through the analysis of historical time series),

$$\frac{S(t + \delta t) - S(t)}{S(t)} \approx \delta \ln(S(t))$$

or obtained from the volatility values made available by commercial data providers or in internet sites. In contrast to e.g. the drift, the volatilities in the risk neutral world are in theory identical to those in the real world. In practice, this is not the case since theoretical models only approximate the real world. As mentioned before (e.g., see Sect. 14.9.3), quoted volatilities are often just convenient terms to talk about prices calculated from these volatilities by means of the Black '76 formula. Within this context, (implicit) volatilities used for the arbitrage free valuation of derivatives can be considered as mere market parameters, resulting just like equity share prices from offer and demand in markets with active trading. Therefore, implicit volatilities differ from realized real volatilities of, e.g., observed real equity share prices. As a consequence, portfolios with a significant portion of volatility dependent derivatives are exposed to volatility risk, i.e. the risk caused by changing implicit volatilities. Based on historical time series of implicit volatilities, in analogy to share prices, the (real) volatilities of (implicit) volatilities are calculated.

There is another reason why risk calculations are based on real volatilities and therefore on real probability distributions of the parameters. The arbitrage-free valuation of derivatives is accurately applied in cases, in which every risk position is hedged and therefore neutralized, i.e. the risk vanishes. Without any risk, calculating the risk does not make much sense. Instead, it is the purpose of the risk measurement to estimate the risk from those positions and risk factors that could not (or just aren't) be fully hedged. Therefore, risk calculation is performed in the real world rather then in the arbitrage-free world. A (currently) statically hedged position or risk factor would not show any risk in the risk calculation. For the valuation of derivatives within the risk calculation, arbitrage-free methods are still required in order to estimate the real change in valuation of the derivative after a real change of market parameters required for the valuation (e.g. share prices, implicit volatilities, etc.).

As opposed to Eq. 21.7 which only holds for infinitesimal time steps dt, the finite time span δt in Eq. 21.8 can be taken to be arbitrarily long. Obviously, since the variance of the normally distributed stochastic component δW of

the risk factor equals δt, the risk can be kept small by keeping the length of this time interval small. δt should, however, be chosen large enough so that the position concerned can be liquidated within this time span. For this reason δt is sometimes referred to as the *liquidation period*. To comply with regulations, a liquidation period of $\delta t = 10$ days is required for internal models. Sometimes, liquidation periods of $\delta t = 1$ day or $\delta t = 25$ days are used for other purposes, too.

We now consider the value at risk with respect to a specified confidence c of a portfolio consisting of a single position in N of one risk factor S. The value of this portfolio at time t is $V = NS(t)$. The change in the value $\delta V(t)$ caused by the change in the risk factor is

$$\delta V(t) = N\delta S(t) \quad \text{with} \quad \delta S(t) = S(t + \delta t) - S(t) \ .$$

This case is by no means as special as it may seem. The change in S induces a change in V amplified by the *constant* factor N. The factor N is, so to speak, the *sensitivity* of V with respect to S and δV is a *linear* function of δS. The interpretation of N as the "number of instruments" in the portfolio is not essential in our deliberations. The same results hold for any portfolio whose change in value is a linear function of the change in the risk factor (or at least approximately so), as is the case for the delta-normal approximation in the variance-covariance method introduced below.

The value change δV can be explicitly derived from Eq. 21.8 as

$$\begin{aligned} \delta V &= NS(t + \delta t) - NS(t) \\ &= NS(t)\left[\exp\left(\mu\delta t + \sigma\delta W\right) - 1\right] \ . \end{aligned} \tag{21.9}$$

This allows us to express the probability required in Eq. 21.3, i.e., the probability that $\delta V > -\text{VaR}(c)$, as

$$P(\delta V > -\text{VaR}) = P\left(NS(t)\left[e^{\mu\delta t + \sigma\delta W} - 1\right] > -\text{VaR}\right) \ .$$

Next, we need to determine this unknown probability. The only stochastic variable involved is the Brownian motion whose distribution is given by

$$\delta W \sim \text{N}(0, \delta t) \Rightarrow \delta W \sim X\sqrt{\delta t} \quad \text{with} \quad X \sim \text{N}(0, 1) \ . \tag{21.10}$$

We begin by rewriting the condition that $\delta V > -\text{VaR}$ with the purpose of isolating the stochastic component δW to see what this condition means in terms of δW.

$$P(\delta V > -\text{VaR})$$

$$= P\left(e^{\mu \delta t + \sigma \delta W} - 1 > -\frac{\text{VaR}}{N S(t)}\right)$$

$$= P\left(e^{\mu \delta t + \sigma \delta W} > 1 - \frac{\text{VaR}}{N S(t)}\right)$$

$$= P\left(\mu \delta t + \sigma \delta W > \ln\left(1 - \frac{\text{VaR}}{N S(t)}\right)\right)$$

$$= P\left(\delta W > \frac{1}{\sigma}\ln\left(1 - \frac{\text{VaR}}{N S(t)}\right) - \frac{\mu}{\sigma}\delta t\right)$$

$$= P\left(\sqrt{\delta t}\, X < \frac{1}{\sigma}\ln\left(1 - \frac{\text{VaR}}{N S(t)}\right) - \frac{\mu}{\sigma}\delta t\right)$$

$$= P\left(X > a\right) ,$$

where we have used Eq. 21.10 and the abbreviation

$$a = \frac{1}{\sigma\sqrt{\delta t}}\ln\left(1 - \frac{\text{VaR}}{N S(t)}\right) - \frac{\mu}{\sigma}\sqrt{\delta t} . \qquad (21.11)$$

Since the probability distribution of X is just the standard normal distribution $N(0, 1)$, which has the well-known density $\text{pdf}_X(x)$ (e.g., see Sect. A.4.3, Gl. A.52) with

$$\text{pdf}_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} .$$

Then, the required probability can be calculated as

$$P(\delta V > -\text{VaR}) = P\left(X > a\right)$$

$$= \int_a^\infty \text{pdf}_X(x)dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_a^\infty \exp(-x^2/2)dx .$$

The value at risk with respect to the confidence c is the value satisfying Eq. 21.3. Inserting the above result into this requirement now yields

$$c \stackrel{!}{=} P(\delta V > -\text{VaR}) = \frac{1}{\sqrt{2\pi}} \int_a^\infty \exp(-x^2/2) dx .$$

The parameter a introduced in Eq. 21.11 is thus the $(1-c)$ percentile of the standard normal distribution

$$a = Q_{1-c}^{\text{N}(0,1)} . \tag{21.12}$$

This value can be determined for any arbitrary confidence level. Examples of the boundary of a one-sided confidence interval of the standard normal distribution are

$$c = 95\% = 0.95 \Rightarrow a = Q_{1-c}^{\text{N}(0,1)} \approx -1.645$$
$$c = 99\% = 0.99 \Rightarrow a = Q_{1-c}^{\text{N}(0,1)} \approx -2.326 \tag{21.13}$$

Both of these confidences are frequently used in practice. Regulators require $c = 99\%$ for the calculation of the capital requirement, whereas for other purposes like internal risk management $c = 95\%$ is also used.

Having computed the percentile a we are now in a position to apply Eq. 21.11 to isolate the value at risk for a previously specified confidence level c:

$$a = \left[\ln \left(1 - \frac{\text{VaR}}{NS(t)} \right) - \mu \delta t \right] / (\sigma \sqrt{\delta t})$$

$$a\sigma \sqrt{\delta t} + \mu \delta t = \ln \left(1 - \frac{\text{VaR}}{NS(t)} \right)$$

$$\exp \left(a\sigma \sqrt{\delta t} + \mu \delta t \right) = 1 - \frac{\text{VaR}}{NS(t)}$$

$$\frac{\text{VaR}}{NS(t)} = 1 - \exp \left(a\sigma \sqrt{\delta t} + \mu \delta t \right)$$

$$\text{VaR} = NS(t) \left[1 - \exp \left(a\sigma \sqrt{\delta t} + \mu \delta t \right) \right] .$$

Now using our main result Eq. 21.12 we finally obtain the value at risk with respect to the confidence level c of a position consisting of N of the same instruments depending solely on a single underlying risk factor S:

$$\text{VaR}(c) = N S(t) \left[1 - \exp \left(\mu \delta t + Q_{1-c}^{\text{N(0,1)}} \sigma \sqrt{\delta t} \right) \right] . \qquad (21.14)$$

Thus, under the assumption that S behaves as the random walk in Eq. 21.7, the probability is c that any loss due to this investment over the time span δt is smaller than this value at risk.

Now consider the risk of a *short* position consisting of N of the risk factor. The change in the portfolio's value is given by an equation analogous to Eq. 21.9, namely

$$\delta V = -N S(t) \left[\exp \left(\mu \delta t + \sigma \delta W \right) - 1 \right]$$

and the cumulative distribution function needed in Eq. 21.3 is now

$$P(\delta V > -\text{VaR}) = P \left(-N S(t) \left[e^{\mu \delta t + \sigma \delta W} - 1 \right] > -\text{VaR} \right)$$
$$= P \left(N S(t) \left[e^{\mu \delta t + \sigma \delta W} - 1 \right] < \text{VaR} \right) ,$$

where in the last step, multiplication by (-1) has reversed the inequality to a "lower-than" in the above condition. It follows that

$$P(\delta V > -\text{VaR}) = P \left(\delta W < \frac{1}{\sigma} \ln \left(1 + \frac{\text{VaR}}{N S(t)} \right) - \frac{\mu}{\sigma} \delta t \right)$$
$$= P \left(X < \frac{1}{\sigma \sqrt{\delta t}} \ln \left(1 + \frac{\text{VaR}}{N S(t)} \right) - \frac{\mu}{\sigma} \sqrt{\delta t} \right) .$$

Analogous to Eq. 21.11, we introduce the following definition[3]:

$$\tilde{a} = \frac{1}{\sigma \sqrt{\delta t}} \ln \left(1 + \frac{\text{VaR}}{N S(t)} \right) - \frac{\mu}{\sigma} \sqrt{\delta t} . \qquad (21.15)$$

It then follows that the definition given in 21.2 for the value at risk becomes

$$c \overset{!}{=} \text{cpf}_{\delta V}(\delta V > -\text{VaR}) = \text{cpf}_X (X < \tilde{a}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\tilde{a}} \exp(-x^2/2) dx$$

---

[3]The difference to a in Eq. 21.11 lies in the sign of the VaR.

The parameter $\widetilde{a}$ is thus the c percentile of the standard normal distribution

$$\widetilde{a} = Q_c^{N(0,1)} .$$

The symmetry A.54 of the standard normal distribution

$$N(-x) = 1 - N(x)$$

implies that the percentile obeys the relation[4]

$$Q_c^{N(0,1)} = -Q_{1-c}^{N(0,1)} .$$

It follows that we can use the same percentiles as given for instance in Eq. 21.13 for two commonly used confidence levels. Inserting $\widetilde{a} = -Q_{1-c}^{N(0,1)}$ and solving Eq. 21.15 for the value at risk of a short position yields

$$\begin{aligned}
\mathrm{VaR}(c) &= -NS(t)\left[1 - \exp\left(\widetilde{a}\sigma\sqrt{\delta t} + \mu\delta t\right)\right] \\
&= -NS(t)\left[1 - \exp\left(\mu\delta t - Q_{1-c}^{N(0,1)}\sigma\sqrt{\delta t}\right)\right] .
\end{aligned}$$

We can now write down the two formulas for the risks involved in both a long and a short position in one and the same risk factor:

$$\begin{aligned}
\mathrm{VaR}_{\mathrm{long}}(c) &= NS(t)\left[1 - \exp\left(\mu\delta t + Q_{1-c}^{N(0,1)}\sigma\sqrt{\delta t}\right)\right] \\
\mathrm{VaR}_{\mathrm{short}}(c) &= -NS(t)\left[1 - \exp\left(\mu\delta t - Q_{1-c}^{N(0,1)}\sigma\sqrt{\delta t}\right)\right] . \qquad (21.16)
\end{aligned}$$

These two key figures, because the underlying is assumed to belong to a lognormal distribution, which is other than a normal distribution not symmetric (see Fig. 2.5). The effect of the drift term $\mu\delta t$ differs also: for a long position the VaR is reduced, while it increases the VaR of a short position,

---

[4]This can be seen quite easily: Let $c = N(x)$. By definition, x is the percentile associated with c. Or equivalently, the inverse of the cumulative distribution function gives the percentile $Q_c = x = N^{-1}[N(x)] = N^{-1}(c)$. Applying N^{-1} to the symmetry equation $N(-x) = 1-N(x)$ gives

$$N^{-1}[N(-x)] = N^{-1}[1 - N(x)]$$

$$\Leftrightarrow \qquad -x = Q_{1-N(x)} .$$

Substituting c for $N(x)$ and Q_c for x immediately yields $-Q_c = Q_{1-c}$.

since N is negative. The approximations required to make the risks of a long and short position look equal will be discussed in greater detail in the next section.

21.4 Approximations in the Distribution of Risk Factors

For short liquidation periods (for example, $\delta t = 10$ days $= 0.0274$ years), the exponential function in Eq. 21.8 is frequently approximated by its Taylor series[5] expanded up to and including its linear term. In addition, the contribution of the drift is often assumed to be negligible in comparison to the volatility in such short periods and is frequently excluded from consideration. A third possibility, also frequently observed in practice is to combine the above two approximations, neglecting the drift and approximating the exponential function with its linear Taylor series expansion as well. These three approximations correspond to a change in the risk factor, $\delta S(t) = S(t + \delta t) - S(t)$, in the following three ways:

$$\delta S(t) \approx \begin{cases} S(t)\left[e^{\sigma \delta W} - 1\right] & \mu \text{ neglected} \\ S(t)\left[\mu \delta t + \sigma \delta W\right] & \text{linear approximation for exp} \\ S(t)\sigma \delta W & \mu \text{ neglected and linear approximation .} \end{cases}$$
$$(21.17)$$

We now show what effect these different approximations have on the value at risk of the portfolios treated above consisting of a long or a short position in a risk factor, respectively. The change in portfolio value δV induced by the risk factor changes are

$$\delta V_{\text{long}} = N \delta S(t), \ \delta V_{\text{short}} = -N \delta S(t) \ ,$$

for the long and short position, respectively. The approximations for $\delta S(t)$ thus induce an analogous approximation in δV. Consequently, the

---

[5]$\exp(x) \approx 1 + x + \cdots$

approximate value at risk of, for example, the long position can be derived as was done above yielding[6]

$$
\mathrm{VaR}_{\mathrm{long}}(c) \approx
\begin{cases}
N S(t) \left[1 - \exp\left(Q_{1-c}^{\mathrm{N}(0,1)} \sigma \sqrt{\delta t} \right) \right] & \mu \approx 0 \\
N S(t) \left[-\mu \delta t - Q_{1-c}^{\mathrm{N}(0,1)} \sigma \sqrt{\delta t} \right] & \text{exp linear} \\
-N S(t) Q_{1-c}^{\mathrm{N}(0,1)} \sigma \sqrt{\delta t} & \mu \approx 0 \,, \ \text{exp linear} \,.
\end{cases}
\tag{21.18}
$$

This follows also—and more directly—by approximating the exponential function in Eq. 21.16. The value at risk for the short position can be approximated in much the same way, the only difference being that the signs of N and $Q_{1-c}^{\mathrm{N}(0,1)}$ are now negative. This implies that, even in linear approximation, the VaRs of a long and short position are not exactly equal:

$$
\mathrm{VaR}_{\mathrm{long}}(c) \approx -N S(t) Q_{1-c}^{\mathrm{N}(0,1)} \sigma \sqrt{\delta t} - N S(t) \mu \delta t
$$
$$
\mathrm{VaR}_{\mathrm{short}}(c) \approx -N S(t) Q_{1-c}^{\mathrm{N}(0,1)} \sigma \sqrt{\delta t} + N S(t) \mu \delta t \,.
\tag{21.19}
$$

From these two linear approximations for the value at risk it is clear that the two VaRs can only then be equal when the drift is neglected in the linear approximation!

And only in this case is the well-known *square root of time law* valid. This law states that the value at risk with respect to one liquidation period δt can be calculated from the value at risk with respect to another liquidation period $\delta t'$ by multiplying by the square root of the ratio of the liquidation periods. Correspondingly, only in this approximation can the value at risk at a given confidence level a be converted into a VaR with respect to another confidence a' by simply multiplying by the ratio of the boundaries of the respective confidence intervals. Both conversions find their expression in the following equation:

$$
\mathrm{VaR}(c', \delta t') \approx \frac{Q_{1-c'}}{Q_{1-c}} \sqrt{\frac{\delta t'}{\delta t}} \ \mathrm{VaR}(c, \delta t) \,.
\tag{21.20}
$$

---

[6]Note that $Q_{1-c}^{\mathrm{N}(0,1)} < 0$ for all reasonable confidence levels c, see for instance Eq. 21.13. Since for all time spans δt usually considered and for all reasonable values of μ the drift term is smaller than the volatility term, the value at risk is a positive number.

On the basis of this equation, the value at risk computed for given data obtained from some commercial data providers, with, for example, 95% confidence and a liquidation period of 1 day, can be converted to the value at risk corresponding to a different confidence level and/or liquidation period (for instance, 99% and 10 days for capital requirements). Equation 21.20 can be easily derived from the third approximation of the value at risk listed in 21.18 (for this approximation, the risk for a long and short position are equal).

$$
\begin{aligned}
\text{VaR}(c', \delta t') &\approx -NS(t)\,Q_{1-c'}\sigma\sqrt{\delta t'} \\
&= -NS(t)\frac{Q_{1-c'}}{Q_{1-c}}Q_{1-c}\sigma\sqrt{\frac{\delta t'}{\delta t}}\delta t = \underbrace{-NS(t)Q_{1-c}\sigma\sqrt{\delta t}}_{\text{VaR}(c,\delta t)}\frac{Q_{1-c'}}{Q_{1-c}}\sqrt{\frac{\delta t'}{\delta t}} \ .
\end{aligned}
$$

Equation 21.20 is valid *only* in linear approximation *and* when the drift is neglected. Furthermore, Eq. 21.20 holds only for portfolios whose value changes are linear functions of the change in the underlying risk factor or when the linearity holds at least in good approximation. Examples are (as in the case currently under discussion) portfolios consisting of a single risk factor (the constant of proportionality is N) or the value at risk obtained by means of the delta-normal approximation of the variance-covariance method (see Chap. 22) where the constant of proportionality is given by the portfolio delta.

21.5 The Covariance Matrix

In general, the value of the portfolio V depends on not only one but on several (often hundreds of) risk factors. We assume that these risk factors $S_i(t)$, $i = 1, \ldots, n$ can be modeled by random walks as in Eqs. 21.7, 2.17 or Eq. 2.25 and obey the *coupled* stochastic differential equation

$$
d \ln S_i(t) = \mu_i dt + dZ_i \ , \tag{21.21}
$$

where the dZ_i denote Brownian motions which are *correlated* with each other. If the random walks governing the risk factors were uncorrelated, their behavior could be completely determined through the specification of two parameters per risk factor, namely the drift and the volatility. The fluctuation (the variance) of a risk factor, which is of particular importance in risk management, could be described by the volatility alone. Since underlying prices, however, are in general correlated and the variance of the entire

portfolio (a linear combination of the risk factors) is of interest, it is insufficient to specify a single variance for each risk factor. It is evident from Eq. A.17, for example, that the *covariance* is needed to incorporate both the variance of each risk factor as well as the *correlation* between each pair of risk factors into a model which completely describes the value changes of the portfolio. These variances and correlations of the logarithmic changes of each of the risk factors can be quite clearly represented by a matrix whose ijth element is composed of the product of the correlation between the ith and jth risk factor with the standard deviations of both these risk factors:

$$\delta\mathbf{\Sigma} = \begin{pmatrix} \delta\Sigma_{11} & \delta\Sigma_{12} & \cdots & \cdots & \delta\Sigma_{1n} \\ \delta\Sigma_{21} & \ddots & & \ddots & \delta\Sigma_{2n} \\ \vdots & & \delta\Sigma_{ij} & & \vdots \\ \vdots & \ddots & & \ddots & \vdots \\ \delta\Sigma_{n1} & \delta\Sigma_{n2} & \cdots & \cdots & \delta\Sigma_{nn} \end{pmatrix}$$

$$\text{where } \delta\Sigma_{ij} = \underbrace{\rho_{ij}}_{\substack{\text{Correlation} \\ \text{of } \ln S_i \text{ with } \ln S_j}} \underbrace{\sigma_i\sqrt{\delta t}}_{\substack{\text{Std. Dev.} \\ \text{of } \ln S_i}} \underbrace{\sigma_j\sqrt{\delta t}}_{\substack{\text{Std. Dev.} \\ \ln \text{of } S_j}} \text{for } i, j = 1, 2, \ldots, n \ .$$

$$(21.22)$$

Here, δt is the reference time interval for the change in the risk factors, i.e., $\delta t = 1$ day for a daily change in the risk factors, $\delta t = 25$ days for a monthly change, etc. Matrices of this sort are called *covariance matrices*. With this matrix we can write

$$d\mathbf{Z} = \begin{pmatrix} dZ_1 \\ dZ_2 \\ \vdots \\ dZ_n \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, d\mathbf{\Sigma}) \ ,$$

where the following general (vector) notation has been introduced

$$\mathbf{X} \sim \mathrm{N}(\mathbf{R}, \mathbf{V})$$

$$\Longleftrightarrow \quad X_i \text{ are normally distributed with}$$

$$\mathrm{cov}[X_i, X_j] = V_{ij}, \ \mathrm{E}[X_i] = R_i \ .$$

The notation $d\mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma} dt)$ denotes that the vector $d\mathbf{Z}$ consists of the stochastic differentials dZ_i and that these are multivariate normally distributed with expectation 0 and covariance matrix $d\boldsymbol{\Sigma}$:

$$d\mathbf{Z} \sim N(\mathbf{0}, d\boldsymbol{\Sigma}) \implies \text{cov}[dZ_i, dZ_j] = d\Sigma_{ij}, \;\; E[dZ_i] = 0$$
$$\text{with} \quad d\Sigma_{ij} = \sigma_i \rho_{ij} \sigma_j dt \;. \qquad (21.23)$$

As was the case for a single risk factor, the S_i are solutions to the associated stochastic differential equations 21.21. Analogously to Eq. 2.28, theses solutions are given by the following stochastic processes

$$S_i(t + \delta t) = S_i(t) \exp\left(\mu_i \delta t + \delta Z_i\right) \text{ with } \delta\mathbf{Z} \sim N(\mathbf{0}, \delta\boldsymbol{\Sigma}) \qquad (21.24)$$

This equation specifies the distribution of the market parameter $S_i(t + \delta t)$ after a finite time interval δt has passed. The value $S_i(t)$ is known at time t as is the distribution of the random vector $\delta\mathbf{Z}$. The δZ_i are multivariate normally distributed with the covariance matrix

$$\delta\Sigma_{ij} = \sigma_i \rho_{i,j} \sigma_j \delta t$$

as given in Eq. 21.22.

As opposed to Eq. 21.21, which holds for infinitesimal time steps dt, the finite time interval of length δt is arbitrary and can be taken to be the liquidation period used for the computation of the VaR, for example. For short time intervals δt, the changes in the risk factors $\delta S_i(t) = S_i(t + \delta t) - S_i(t)$ can be approximated analogously as was done in Eq. 21.17:

$$\delta S_i(t) \approx \begin{cases} S_i(t) \left[e^{\delta Z_i} - 1 \right] & \text{drift neglected} \\ S(t) \left[\mu_i \delta t + \delta Z_i \right] & \text{linear approximation for exp} \\ S(t) \delta Z_i & \text{drift neglected and linear approximation .} \end{cases}$$
$$(21.25)$$

It is common practice to work with the latter and simplest of the three approximations above

$$\delta S_i(t) \approx S_i(t) \, \delta Z_i \text{ mit } \;\; \delta\mathbf{Z} \sim N(\mathbf{0}, \delta\boldsymbol{\Sigma}) \;. \qquad (21.26)$$

21.5.1 Logarithmic Changes Versus Returns

It is important (for instance when determining the covariance matrix from historical time series) to understand, that the covariance matrix $\delta \Sigma_{ij}$ is defined as the covariance of the *logarithmic changes*, and *not* as the covariance of the *returns*. The risk factor returns $r_i(t)$ over a time period δt are defined via

$$S_i(t)e^{r_i(t)\delta t} := S_i(t + \delta t)$$

$$\iff \quad r_i(t)\delta t = \ln \left(\frac{S_i(t + \delta t)}{S_i(t)} \right) = \ln S_i(t + \delta t) - \ln S_i(t) = \delta \ln S_i(t) \ .$$

Thus, the difference between returns and logarithmic changes is a factor δt:

$$\delta \ln S_i(t) = r_i(t)\delta t \ . \tag{21.27}$$

Therefore the covariance matrix $\delta \Sigma_{ij}$ differs from the covariance of the returns by a factor of δt^2:

$$\delta \Sigma_{ij} \equiv \text{cov}\left[\delta \ln S_i, \delta \ln S_j\right] = \text{cov}\left[r_i \delta t, r_j \delta t\right] = \delta t^2 \, \text{cov}\left[r_i, r_j\right] \ . \tag{21.28}$$

In particular

$$\sigma_i^2 \delta t \equiv \delta \Sigma_{ii} \equiv \text{var}\left[\delta \ln S_i\right] = \delta t^2 \, \text{var}\left[r_i\right] \ . \tag{21.29}$$

Thus, the risk factor volatility can be expressed in terms of the variance of the logarithmic changes or in terms of the variance of the returns in the following way

$$\sigma_i = \frac{1}{\sqrt{\delta t}}\sqrt{\text{var}\left[\ln S_i\right]} = \sqrt{\delta t}\sqrt{\text{var}\left[r_i\right]} \ . \tag{21.30}$$

This relation is of utmost importance at any *annualization* of volatilities obtained from historical analysis!

From Eq. 21.28 it follows, that the Betas (which measures to what extend the variance of a risk factor could be explained by correlation with another one) as defined in Eq. A.20 are the same for returns and for logarithmic changes:

$$\beta\left[\delta \ln S_i, \delta \ln S_j\right] = \frac{\text{cov}\left[\delta \ln S_i, \delta \ln S_j\right]}{\text{var}\left[\delta \ln S_j\right]} = \frac{\delta t^2 \, \text{cov}\left[r_i, r_j\right]}{\delta t^2 \, \text{var}\left[r_j\right]}$$

$$= \beta\left[r_i, r_j\right] \equiv \beta_{ij} \ .$$

Similarly, the correlations as defined in Eq. A.14 are the same for returns and for logarithmic changes:

$$\text{corr}\left[\delta \ln S_i, \delta \ln S_j\right] \equiv \frac{\text{cov}\left[\delta \ln S_i, \delta \ln S_j\right]}{\sqrt{\text{var}\left[\delta \ln S_i\right]\text{var}\left[\delta \ln S_j\right]}} = \frac{\delta t^2 \, \text{cov}\left[r_i, r_j\right]}{\sqrt{\delta t^2 \, \text{var}\left[r_i\right] \delta t^2 \, \text{var}\left[r_j\right]}}$$

$$= \text{corr}\left[r_i, r_j\right] \equiv \rho_{ij} \ .$$

Expressing the covariances in terms of correlation and volatilities we find

$$\text{cov}\left[\delta \ln S_i, \delta \ln S_j\right] = \rho_{ij}\sqrt{\text{var}\left[\delta \ln S_i\right]}\sqrt{\text{var}\left[\delta \ln S_j\right]} = \rho_{ij}\sigma_i \sqrt{\delta t}\; \sigma_j \sqrt{\delta t}$$

$$\text{cov}\left[r_i, r_j\right] = \rho_{ij}\sqrt{\text{var}\left[r_i\right]}\sqrt{\text{var}\left[r_j\right]} = \rho_{ij}\frac{\sigma_i}{\sqrt{\delta t}}\frac{\sigma_j}{\sqrt{\delta t}}\ .$$

In summary, the relations involving volatilities, correlations and Betas with the correct factors δt are

$$\beta_{ij} \equiv \beta\left[\delta \ln S_i, \delta \ln S_j\right] = \beta\left[r_i, r_j\right]$$

$$\rho_{ij} \equiv \text{corr}\left[\delta \ln S_i, \delta \ln S_j\right] = \text{corr}\left[r_i, r_j\right] \qquad (21.31)$$

$$\rho_{ij}\sigma_i\sigma_j = \frac{1}{\delta t}\,\text{cov}\left[\delta \ln S_i, \delta \ln S_j\right] = \delta t\,\text{cov}\left[r_i, r_j\right]\ .$$

21.5.2 Covariance Matrices of Data Providers

A price of the ith risk factor is, of course, 100% correlated with itself. Observe that the correlation of two risk factors, say the ith and jth risk factors, is symmetric in the sense of Eq. A.15. This symmetry makes almost half the information appearing in Eq. 21.22 redundant. Only the "triangle" below the diagonal is needed. The matrix elements in the diagonal, i.e., those whose row index equals their column index, are composed of the volatilities since, as was mentioned above, the correlation of a risk factor with itself is always 1. Some commercial data providers only make the data in the "triangle" below the main diagonal of the correlation matrix (without the volatilities) available rather than the entire covariance matrix. Volatilities are not placed directly at the user's disposal. Instead, the standard deviations of the logarithmic change in the risk factors over $\delta t = 1$ day or $\delta t = 25$ days multiplied by 1.645 are

given. Also, the current risk factor levels (called the *current levels*) are available. In summary, the following data can be obtained for both 1 day and 25 days:

$$
\mathbf{K} = \begin{pmatrix}
1 & 0 & \cdots & 0 & \cdots & 0 \\
\rho_{2,1} & \ddots & \ddots & & \ddots & \vdots \\
\vdots & \ddots & 1 & 0 & & 0 \\
\rho_{i,1} & & \rho_{i,i-1} & 1 & \ddots & \vdots \\
\vdots & \ddots & & & \ddots & \ddots & 0 \\
\rho_{n,1} & \cdots & \rho_{n,i-1} & \cdots & \rho_{n,n-1} & 1
\end{pmatrix}
\tag{21.32}
$$

$$
\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_{n-1} \\ Y_n \end{pmatrix}, \quad
\mathbf{S} = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ \vdots \\ S_{n-1} \\ S_n \end{pmatrix}.
$$

$\mathbf{Y}$ is a vector whose components are not the volatilities but the *standard deviations* of either daily ($\delta t = 1$ day) or monthly ($\delta t = 25$ days) logarithmic risk factor fluctuations multiplied by 1.645 given explicitly by

$$
Y_i = -Q_{1-c}\sqrt{\text{var}[x_i]} = -Q_{1-c}\sqrt{\delta t}\,\sigma_i
\tag{21.33}
$$

with $Q_{1-c} = -1,645 \Rightarrow c = 95\%$.

In order to obtain the actual volatilities, these Y_i need to be divided by 1.645 and $\sqrt{\delta t}$. The variables Y_i are referred to by some commercial data providers as the "value at risk" of the underlying (risk factor) S_i. In fact, neglecting the drift as in Eq. 21.18, Y_i can be used to directly calculate the value at risk of a risk factor at a confidence level of $c = 95\%$ (see Eq. 21.13):

$$
\text{VaR}_i(95\%, \delta t) \approx S_i(t) - e^{-Y_i} S_i(t) \approx S_i(t)Y_i .
\tag{21.34}
$$

Thus interpreted, Y_i is a linear approximation (also neglecting the effect of the drift) of the *relative* value at risk of the ith risk factor:

$$
Y_i = \frac{\text{VaR}_i(95\%, \delta t)}{S_i(t)} .
$$

A linear approximation of the value at risk with respect to *another* confidence level and/or *another* liquidation period can be obtained (neglecting the drift) via 21.20.

Instead of relying on the volatilities and correlations of an external provider, these parameters can be determined by analyzing historical time series. In general, volatilities implied by traded options will not be used, even in the presence of liquidly traded standard options. Instead, "real world" volatilities should be preferred. since implicit volatilities are mere quotation vehicles. On the other side, implicit volatilities are themselves treated as pricing relevant risk factors in the risk management of portfolios with options. The volatility of these implicit volatilities can be calculated from historical time series of implicit volatility quotes.

21.5.3   Cholesky Decomposition of the Covariance Matrix

As we will soon see, it is often quite convenient to work with a matrix which, when multiplied by itself, yields the covariance matrix 21.22. The need for such a matrix arises when applying several of the most important methods used in computing risk, such as Monte Carlo or the delta-gamma methods. This "square root" of the covariance matrix will be denoted by $\mathbf{A}$ and satisfies the condition

$$\mathbf{A}\mathbf{A}^T = \delta\mathbf{\Sigma} \ . \tag{21.35}$$

Here the notation $\mathbf{A}^T$ denotes the *transpose* of the matrix $\mathbf{A}$. We obtain the transpose by writing the row vectors of $\mathbf{A}$ as column vectors, i.e.

$$\left(\mathbf{A}^T\right)_{ij} = A_{ji} \ .$$

Transformation of Uncorrelated Random Variables into Correlated Random Variables

A useful property of $\mathbf{A}$ (particularly for Monte Carlo simulations, see Sect. 23.1) is that it transforms uncorrelated random variables into correlated

ones (with covariance $\delta\boldsymbol{\Sigma}$). Let $X_i, i = 1, \ldots, n$ be uncorrelated standard normally distributed random variables

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \text{ with } \mathbf{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{1}) \ . \tag{21.36}$$

The symbol $\mathbf{1}$ is used to denote the *identity matrix*

$$\mathbf{1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \ , \tag{21.37}$$

or equivalently,[7] $(\mathbf{1})_{ij} = \delta_{ij}$. The notation in Eq. 21.36 means explicitly:

$$\mathrm{cov}[X_i, X_j] = \delta_{ij}, \ \mathrm{E}[X_i] = 0 \ \forall i, j = 1, \ldots n \ .$$

Applying the matrix $\mathbf{A}$ to $\mathbf{X}$ generates new random variables $\mathbf{Y}$ defined by

$$\mathbf{Y} = \mathbf{AX} \Longleftrightarrow Y_i = \sum_k A_{ik} X_k \ .$$

The covariances of these new random variables are

$$\mathrm{cov}\left[Y_i, Y_j\right] = \mathrm{cov}\left[\sum_k A_{ik} X_k, \sum_m A_{jm} X_m\right]$$

$$= \sum_k A_{ik} \sum_m A_{jm} \underbrace{\mathrm{cov}[X_k, X_m]}_{\delta_{km}}$$

$$= \sum_k A_{ik} A_{jk}$$

---

[7] Here, we make use of the so called *Kronecker delta* often appearing in the science. It is defined as

$$\delta_{ij} = \begin{cases} 1 \text{ for } i = j \\ 0 \text{ for } i \neq j \end{cases}$$

$$= \left(\mathbf{A}\mathbf{A}^T \right)_{ij}$$
$$= \delta \Sigma_{ij} \ ,$$

where the defining property of the matrix **A,** Eq. 21.35, was used in the last step. The expectations of these random variables are

$$\mathrm{E}[Y_i] = \mathrm{E}\left[\sum_k A_{ik} X_k \right] = \sum_k A_{ik} \underbrace{\mathrm{E}[X_k]}_{0} = 0 \ .$$

This means that the random vector with components Y_i is multivariate normally distributed with covariance matrix $\delta \boldsymbol{\Sigma}$ and zero expectation:

$$\mathbf{A}\mathbf{X} = \mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \delta \boldsymbol{\Sigma}) \ . \tag{21.38}$$

Transformation of Correlated Random Variables into Uncorrelated Random Variables

Conversely, it is often convenient (for example, when using the delta-gamma method, see Sect. 22.3) to transform correlated random variables into uncorrelated ones. As we might suspect, this is accomplished through the mapping given by the inverse $\mathbf{A}^{-1}$ of the matrix **A**. The proof is somewhat more involved than that given in the previous section: Let $Y_i, i = 1, \ldots, n$ be correlated multivariate normally distributed random variables with covariance $\delta \boldsymbol{\Sigma}$,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ with } \mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \delta \boldsymbol{\Sigma})$$

or more explicitly

$$\mathrm{cov}[Y_i, Y_j] = \delta \Sigma_{ij} \ , \ \ \mathrm{E}[Y_i] = 0 \ \forall i, j = 1, \ldots n$$

The random variables X_i obtained by applying the inverse matrix $\mathbf{A}^{-1}$ to the random vector **Y** are given by

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y} \Longleftrightarrow X_i = \sum_k (A^{-1})_{ik} Y_k$$

The covariance of these X_i can be calculated as follows

$$\text{cov}[X_i, X_j] = \text{cov}[\sum_k (A^{-1})_{ik} Y_k, \sum_m (A^{-1})_{jm} Y_m]$$

$$= \sum_k (A^{-1})_{ik} \sum_m (A^{-1})_{jm} \underbrace{\text{cov}[Y_k, Y_m]}_{\delta \Sigma_{km}}$$

$$= \sum_k \sum_m (A^{-1})_{ik} \delta \Sigma_{km} (A^{-1})^T_{mj}$$

This, however, is precisely the *matrix product*[8] of the matrices $\mathbf{A}^{-1}$, $\delta \mathbf{\Sigma}$ and $(\mathbf{A}^{-1})^T$. Decomposing the matrix $\delta \mathbf{\Sigma}$ as in Eq. 21.35 yields

$$\text{cov}[X_i, X_j] = \left(\mathbf{A}^{-1} \delta \mathbf{\Sigma} (\mathbf{A}^{-1})^T \right)_{ij}$$

$$= \left(\mathbf{A}^{-1} \mathbf{A} \mathbf{A}^T (\mathbf{A}^{-1})^T \right)_{ij} .$$

The product $\mathbf{A}^{-1}\mathbf{A}$ of first two matrices appearing on the right hand side is by definition of *inverse matrices*, of course, the identity matrix given in Eq. 21.37. Furthermore, the following property holds for every invertible matrix[9]

$$(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1} . \tag{21.39}$$

allowing the covariances to be calculated as

$$\text{cov}[X_i, X_j] = \left(\mathbf{A}^T (\mathbf{A}^{-1})^T \right)_{ij}$$

$$= \left(\mathbf{A}^T (\mathbf{A}^T)^{-1} \right)_{ij}$$

$$= (\mathbf{1})_{ij}$$

$$= \delta_{ij} .$$

---

[8]A short and simple overview of *matrix algebra* can be found in [78], for example.

[9]Since $\mathbf{A}$ is invertible we have: $(\mathbf{A}\mathbf{A}^{-1})^T = \mathbf{1} \Rightarrow (\mathbf{A}^{-1})^T \mathbf{A}^T = \mathbf{1} \Rightarrow (\mathbf{A}^{-1})^T \underbrace{\mathbf{A}^T (\mathbf{A}^T)^{-1}}_{1} = \mathbf{1}(\mathbf{A}^T)^{-1}$.

The expectations of the new random variables remain zero since, by the linearity of the expectation,

$$E[X_i] = E\left[\sum_k (A^{-1})_{ik} Y_k\right] = \sum_k (A^{-1})_{ik} \underbrace{E[Y_k]}_{0} = 0 .$$

Since matrix multiplication is a *linear* transformation, the form of the distribution remains the same as well. In summary, the new random variables X_i are uncorrelated and all have the same standard normal distribution:

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y} \sim N(\mathbf{0}, \mathbf{1}) . \tag{21.40}$$

However, uncorrelated random variables are in general not *independent*. For independence, the joint distribution of the variables has to be equal to the product of the distributions of each individual variable, see Eq. A.63. Fortunately for uncorrelated standard normally distributed random variables this is indeed the case, see Eq. A.62. Therefore, the X_i are *independent*. They are also all governed by the *same* distribution (in this case the standard normal distribution). Therefore the X_i are called *independent, identically distributed* random variables or *iid,* for short.

The Cholesky Decomposition

Having shown the usefulness of a matrix $\mathbf{A}$ satisfying the property in Eq. 21.35, we proceed with the explicit construction of this matrix. In Eq. 11.4 for $n = 2$, it has already been shown how the square root of a matrix can be calculated by means of the *Cholesky decomposition*. For arbitrarily large matrices ($n > 2$), the components A_{ij}, $i, j = 1, \ldots, n$ of the matrix $\mathbf{A}$ are obtained by iteratively solving the following system of equations:

$$A_{ji} = \begin{cases} 0 & \text{for } j < i \\ \sqrt{\delta \Sigma_{ii} - \sum_{k=1}^{i-1} A_{ik}^2} & \text{for } j = i \\ \frac{1}{A_{ii}}\left(\delta \Sigma_{ji} - \sum_{k=1}^{i-1} A_{ik} A_{jk}\right) & \text{for } j > i \end{cases} \tag{21.41}$$

with $\delta\Sigma_{ji}$ as given in Eq. 21.22:

$$\delta\Sigma_{ji} = \begin{cases} \sigma_i^2 \delta t & \text{for } j = i \\ \rho_{ij}\,\sigma_i\sigma_j\delta t & \text{for } j \neq i \end{cases} .$$

In solving the above system, we begin with $i = 1$, $j = 1$. Within a given Index i we first solve for all possible indices j before increasing i. This procedure is illustrated here for $n = 3$:

$$A_{11} = \sqrt{\delta\Sigma_{11}} , \quad A_{21} = \frac{\delta\Sigma_{21}}{A_{11}} = \frac{\delta\Sigma_{21}}{\sqrt{\delta\Sigma_{11}}}, \quad A_{31} = \frac{\delta\Sigma_{31}}{A_{11}} = \frac{\delta\Sigma_{31}}{\sqrt{\delta\Sigma_{11}}}$$

$$A_{22} = \sqrt{\delta\Sigma_{22} - A_{21}^2} = \sqrt{\delta\Sigma_{22} - \delta\Sigma_{21}^2/\delta\Sigma_{11}}$$

$$A_{32} = \frac{\delta\Sigma_{32} - A_{21}A_{31}}{A_{22}} = \frac{\delta\Sigma_{32} - \delta\Sigma_{21}\delta\Sigma_{31}/\delta\Sigma_{11}}{\sqrt{\delta\Sigma_{22} - \delta\Sigma_{21}^2/\delta\Sigma_{11}}}$$

$$A_{33} = \sqrt{\delta\Sigma_{33} - A_{31}^2 - A_{32}^2} = \sqrt{\delta\Sigma_{33} - \frac{\delta\Sigma_{31}^2}{\delta\Sigma_{11}} - \frac{(\delta\Sigma_{32} - \delta\Sigma_{21}\delta\Sigma_{31}/\delta\Sigma_{11})^2}{\delta\Sigma_{22} - \delta\Sigma_{21}^2/\delta\Sigma_{11}}} .$$

Thus, all elements of the matrix **A** have been expressed in terms of the elements of the covariance matrix. Finally, substituting for $\delta\Sigma_{ij}$ as defined in Eq. 21.22 yields

$$A_{11} = \sigma_1\sqrt{\delta t}$$

$$A_{21} = \sigma_2\sqrt{\delta t}\,\rho_{21} , \quad A_{22} = \sigma_2\sqrt{\delta t}\,\sqrt{1 - \rho_{21}^2}$$

$$A_{31} = \sigma_3\sqrt{\delta t}\,\rho_{31} , \quad A_{32} = \sigma_3\sqrt{\delta t}\,\frac{\rho_{32} - \rho_{31}\rho_{21}}{\sqrt{1 - \rho_{12}^2}}$$

$$A_{33} = \sigma_3\sqrt{\delta t}\,\sqrt{1 - \rho_{31}^2 - \frac{(\rho_{32} - \rho_{31}\rho_{21})^2}{1 - \rho_{12}^2}} .$$

The first equation above holds when only a single risk factor is involved. The first *two* equations correspond to Eq. 11.4 for *two* correlated random walks. The third and fourth rows complete the decomposition for an additional *third* correlated random walk.

22

Check for updates

The Variance-Covariance Method

The *variance-covariance method* makes use of covariances (volatilities and correlations) of the risk factors and the sensitivities of the portfolio values with respect to these risk factors with the goal of approximating the value at risk. This method leads directly to the final result, i.e., the portfolio's value at risk, based on the properties of the assumed portfolio value's probability distribution; no simulation of market data scenarios is involved. The variance-covariance method utilizes linear approximations of the risk factors themselves throughout the entire calculation, often neglecting the drift as well. In view of Eq. 21.25, we have

$$\delta S_i(t) \approx S_i(t) \left[\mu_i \delta t + \delta Z_i \right] \approx S_i(t) \delta Z_i . \tag{22.1}$$

The main idea characterizing this method, however, is that the portfolio value V is expanded in its Taylor series as a function of its risk factors $S_i, i = 1, \ldots n$, and approximated by breaking off after the first or second order term. Let

$$\mathbf{S}(t) = \begin{pmatrix} S_1(t) \\ \vdots \\ S_n(t) \end{pmatrix}$$

denote the vector of risk factors. The Taylor expansion for the change in portfolio value $\delta V(\mathbf{S})$ up to second order is

$$\delta V(\mathbf{S}(t)) = V(\mathbf{S}(t) + \delta \mathbf{S}(t)) - V(\mathbf{S}(t))$$

$$\approx \sum_{i}^{n} \frac{\partial V}{\partial S_i} \delta S_i(t) + \frac{1}{2} \sum_{i,j}^{n} \delta S_i(t) \frac{\partial^2 V}{\partial S_i \partial S_j} \delta S_j(t) \tag{22.2}$$

$$= \sum_{i}^{n} \Delta_i \delta S_i(t) + \frac{1}{2} \sum_{i,j}^{n} \delta S_i(t) \Gamma_{ij} \delta S_j(t)$$

$$\approx \sum_{i}^{n} \widetilde{\Delta}_i \left[\mu_i \delta t + \delta Z_i \right] + \frac{1}{2} \sum_{i,j}^{n} \left[\mu_i \delta t + \delta Z_i \right] \widetilde{\Gamma}_{ij} \left[\mu_j \delta t + \delta Z_j \right] \tag{22.3}$$

$$\approx \sum_{i}^{n} \widetilde{\Delta}_i \, \delta Z_i + \frac{1}{2} \sum_{i,j}^{n} \delta Z_i \widetilde{\Gamma}_{ij} \, \delta Z_j . \tag{22.4}$$

In Eq. 22.2, the portfolio value has been approximated by its Taylor expansion. In Eq. 22.3, the risk factor changes were linear approximate according to Eq. 22.1, and finally, in Eq. 22.4, the average returns have been neglected. The last line in 22.4 is referred to as the *delta-gamma approximation*. Taking the Taylor expansion up to linear order only is called *delta approximation* correspondingly, resulting in an approximation solely consisting of the first of the two sums appearing in the last equation in 22.4.

The abbreviations Δ_i and Γ_{ij}, as usual, denote the *sensitivities* (at time t) of V with respect to the risk factors

$$\Delta_i := \frac{\partial V}{\partial S_i}, \ \Gamma_{ij} := \frac{\partial^2 V}{\partial S_i \partial S_j}, \ i, j = 1, \ldots n .$$

Note here that the mixed partial derivatives arise in the expression for Γ_{ij}. In the literature, the matrix Γ_{ij} is sometimes called the *Hessian matrix*.

We will see below that the sensitivities usually appear in connection with the current levels $S_i(t)$ and $S_j(t)$. The notation $\widetilde{\Delta}_i$ and $\widetilde{\Gamma}_{ij}$ will be used to denote these sensitivities multiplied by the current levels:

$$\widetilde{\Delta}_i := S_i(t)\frac{\partial V}{\partial S_i}, \quad \widetilde{\Gamma}_{ij} := S_i(t)S_j(t)\frac{\partial^2 V}{\partial S_i \partial S_j} \ . \tag{22.5}$$

Using $\widetilde{\Delta}_i$ and $\widetilde{\Gamma}_{ij}$ will prove to substantially simplify the notation. Interpreting the $\widetilde{\Delta}_i$ as components of a vector $\widetilde{\boldsymbol{\Delta}}$, and $\widetilde{\Gamma}_{ij}$ as the elements of a matrix $\widetilde{\boldsymbol{\Gamma}}$, Eq. 22.4 can be written in vector form as

$$\delta V(\mathbf{S}(t)) = \begin{pmatrix} \widetilde{\Delta}_1 \cdots \widetilde{\Delta}_n \end{pmatrix} \begin{pmatrix} \delta Z_1 \\ \vdots \\ \delta Z_n \end{pmatrix}$$

$$+ \frac{1}{2}\begin{pmatrix} \delta Z_1 \cdots \delta Z_n \end{pmatrix} \begin{pmatrix} \widetilde{\Gamma}_{1,1} \cdots \widetilde{\Gamma}_{1,n} \\ \vdots \ \ddots \ \vdots \\ \widetilde{\Gamma}_{n,1} \cdots \widetilde{\Gamma}_{n,n} \end{pmatrix} \begin{pmatrix} \delta Z_1 \\ \vdots \\ \delta Z_n \end{pmatrix}$$

$$= \widetilde{\boldsymbol{\Delta}}^T \delta\mathbf{Z} + \frac{1}{2}\delta\mathbf{Z}^T \widetilde{\boldsymbol{\Gamma}} \delta\mathbf{Z} \ .$$

The approximation of $V(\mathbf{S})$ through its Taylor series expansion up to second order is presented in Fig. 22.1 for a *straddle* (a portfolio made up of a call and a put option) on a risk factor S. The figure has been extracted from the Excel workbook STRADDLE.xlsm available in the download section [50]. We can recognize that the delta-gamma approximation for a simple payoff profile is quite a good approximation. For somewhat more complicated portfolios, however, the delta-gamma approximation fails to be a reasonable representation of the payoff profile. In such cases, we recommend using one of the simulation methods presented in Chap. 23 instead of the Delta-Gamma method. However, the drawback of the simulation approach is of course the significant increase of numerical computations required.
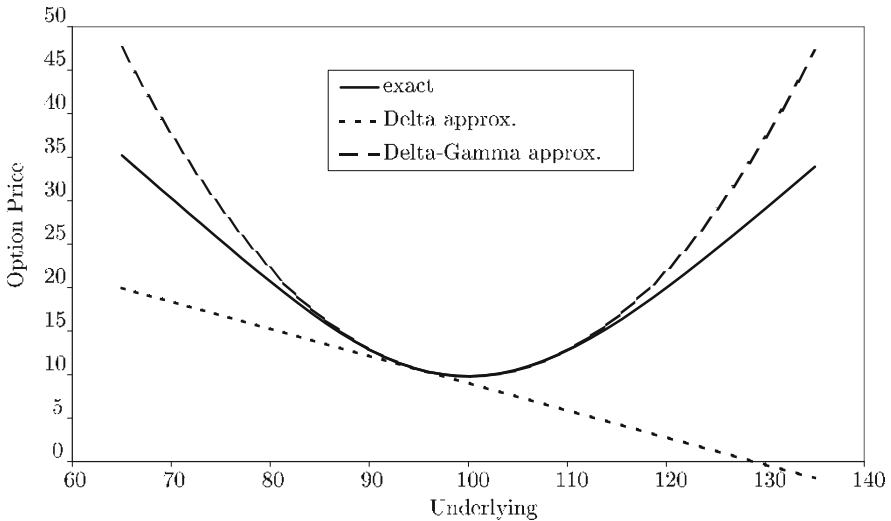
Fig. 22.1 Black-Scholes price of a straddle (strike = 100, time to maturity = 1 year) on an underlying S (volatility 25%, dividend yield 6%, repo rate 3%). The dashed line is the delta-gamma proxy, the dotted line is the simple delta proxy. The Taylor expansion was done about $S = 95$

22.1 Portfolios vs. Financial Instruments

Although we continually refer to *portfolio* sensitivities, the same results hold for individual *financial instruments* as well. In fact, sensitivity of a portfolio composed of financial instruments on the same underlying as described in Sect. 12.4.1 can be obtained by simply adding together the sensitivities of the individual instruments. This is the approach most commonly taken when calculating portfolio sensitivities. For the sake of clarity, we once again present this method explicitly here.

Consider a portfolio with a value V consisting of M different financial instruments with values V_k ,$k = 1, \ldots M$. N_k denotes the number of each instrument with value V_k held in the portfolio. The total value of the portfolio is naturally the sum of the values of each individual position

$$V(t) = \sum_{k=1}^{M} N_k V_k(\mathbf{S}(t)) . \tag{22.6}$$

The change in value $\delta V(t)$ of this portfolio is (approximated up to second order)

$$\delta V(\mathbf{S}(t)) = \sum_{k=1}^{M} N_k \delta V_k(\mathbf{S}(t))$$

$$\approx \sum_{k=1}^{M} N_k \left[\sum_{i}^{n} \frac{\partial V_k}{\partial S_i} \delta S_i(t) + \frac{1}{2} \sum_{i,j}^{n} \delta S_i(t) \frac{\partial^2 V_k}{\partial S_i \partial S_j} \delta S_j(t) \right]$$

$$= \sum_{k=1}^{M} N_k \sum_{i}^{n} \Delta_i^k \delta S_i(t) + \frac{1}{2} \sum_{k=1}^{M} N_k \sum_{i,j}^{n} \delta S_i(t) \Gamma_{ij}^k \delta S_j(t) ,$$

$$(22.7)$$

where the sensitivities of the *financial instruments* have been introduced in the last step. For example, Δ_i^k is the linear sensitivity of the kth financial instrument in the portfolio with respect to the ith risk factor, etc.:

$$\Delta_i^k := \frac{\partial V_k}{\partial S_i}, \ \Gamma_{ij}^k := \frac{\partial^2 V_k}{\partial S_i \partial S_j}, \ i, j = 1, \ldots n; \ k = 1, \ldots M .$$

Simply rearranging the terms makes it clear that the summing over the index k (which denotes the different financial instruments) yields the portfolio sensitivities:

$$\delta V(\mathbf{S}(t)) = \sum_{i}^{n} \delta S_i(t) \sum_{k=1}^{M} N_k \Delta_i^k + \frac{1}{2} \sum_{i,j}^{n} \delta S_i(t) \delta S_j(t) \sum_{k=1}^{M} N_k \Gamma_{ij}^k$$

$$= \sum_{i}^{n} \delta S_i(t) \Delta_i + \frac{1}{2} \sum_{i,j}^{n} \delta S_i(t) \delta S_j(t) \Gamma_{ij} ,$$

Thus, a portfolio sensitivity like Δ_i, for example, contains the sensitivities of all instruments in the portfolio (including all the position sizes N_k) with respect to the considered risk factor. This then yields (as is intuitively clear) the sensitivities of the entire portfolio as sums over the sensitivities of all instruments in the portfolio:

$$\Delta_i = \sum_{k=1}^{M} N_k \Delta_i^k, \ \Gamma_{ij} = \sum_{k=1}^{M} N_k \Gamma_{ij}^k .$$

In practice, this procedure is usually referred to as *position mapping.* Using the approximation in Eq. 22.1 for the change in risk factor, we finally obtain an expression for the portfolio's change in value as

$$\delta V(\mathbf{S}(t)) \approx \sum_i^n \widetilde{\Delta}_i \left[\mu_i \delta t + \delta Z_i\right] + \frac{1}{2} \sum_{i,j}^n \left[\mu_i \delta t + \delta Z_i\right] \widetilde{\Gamma}_{ij} \left[\mu_j \delta t + \delta Z_j\right]$$

$$\approx \sum_i^n \widetilde{\Delta}_i \, \delta Z_i + \frac{1}{2} \sum_{i,j}^n \delta Z_i \widetilde{\Gamma}_{ij} \, \delta Z_j \,,$$

using the modified portfolio sensitivities defined as in Eq. 22.5:

$$\widetilde{\Delta}_i := S_i(t)\Delta_i = S_i(t) \sum_{k=1}^M N_k \Delta_i^k \tag{22.8}$$

$$\widetilde{\Gamma}_{ij} := S_i(t)S_j(t)\Gamma_{ij} = S_i(t)S_j(t) \sum_{k=1}^M N_k \Gamma_{ij}^k \,.$$

Adding the sensitivities of the financial instruments and multiplying by the current levels of the risk factors to obtain these modified portfolio sensitivities is sometimes referred to as *VaR mapping.*

22.2 The Delta-Normal Method

In the *delta-normal method,* the Taylor series 22.4 of the portfolio value is broken off after the linear term.

$$\delta V(\mathbf{S}(t)) \approx \sum_i^n \frac{\partial V}{\partial S_i} \delta S_i(t) \tag{22.9}$$

$$\approx \sum_i^n \widetilde{\Delta}_i \left[\mu_i \delta t + \delta Z_i\right]$$

$$\approx \sum_i^n \widetilde{\Delta}_i \, \delta Z_i = \widetilde{\mathbf{\Delta}}^T \delta \mathbf{Z} \,.$$

22.2.1 The Value at Risk with Respect to a Single Risk Factor

For a single risk factor this means

$$\delta V(S(t)) \approx \Delta \delta S(t) \tag{22.10}$$

with the sensitivity $\Delta := \partial V / \partial S$. The change in the portfolio's value is thus approximated to be a linear function of the change in the underlying risk factor. This corresponds to the situation described in Sect. 21.3. There, the constant of proportionality (the sensitivity) was not Δ but N or $-N$ for a long or short position, respectively. The linear approximation implies intuitively that a portfolio with a linear sensitivity Δ with respect to a risk factor can be interpreted as a portfolio consisting of Δ risk factors. The only subtlety in this argumentation is that in Sect. 21.3, we distinguished between a long and a short position, treating the two cases differently on the basis of whether the proportionality constant N was positive or negative (which leads to the two different VaRs in Eq. 21.16). However, we cannot know a priori whether Δ is greater or less than 0. We do know, however, that V is linear and in consequence, a monotone function of S. Therefore, in the sense of Eq. 21.6, the results following from Eq. 21.16 hold with the correspondence $\Delta \widehat{=} N$ for $\Delta > 0$, and $\Delta \widehat{=} - N$ for $\Delta < 0$. Using this fact allows us to write

$$\mathrm{VaR}_V(c) \approx \max \left\{ \widetilde{\Delta} \left[1 - \exp\left(\mu \delta t + Q_{1-c} \sigma \sqrt{\delta t} \right) \right], \right.$$
$$\left. \widetilde{\Delta} \left[1 - \exp\left(\mu \delta t - Q_{1-c} \sigma \sqrt{\delta t} \right) \right] \right\} \tag{22.11}$$

using the notation $\widetilde{\Delta} := S(t)\Delta = S(t)\partial V / \partial S$. As usual, Q_{1-c} is the $(1-c)$ percentile of the standard normal distribution.[1] The maximum function in Eq. 21.6 effects the correct choice for the VaR. If $\Delta > 0$, the *lower bound* of the confidence interval of the risk factor[2] is relevant and consequently the VaR function as defined above takes on the value corresponding to a *long* position in Eq. 21.16. Likewise for $\Delta < 0$, the *upper bound* of the confidence interval of the risk factor[3] is relevant and the above defined maximum function takes on the value corresponding to the VaR of the short position in Eq. 21.16.

---

[1]For all relevant confidence levels, this percentile is a negative number, see Eq. 21.13.

[2]This corresponds to the percentile Q_{1-c} of the standard normal distribution.

[3]This corresponds to the percentile $-Q_{1-c} = Q_c$ of the standard normal distribution.

In all our deliberations up to this point, only the portfolio value has been approximated with Eq. 22.10. The change in the *risk factor* in Eq. 22.11 is still exact. Approximating this risk factor change with Eq. 22.1, the VaR becomes

$$\text{VaR}_V(c) \approx \max\{\widetilde{\Delta}\left[-\mu\delta t - Q_{1-c}\sigma\sqrt{\delta t}\right],$$
$$\widetilde{\Delta}\left[-\mu\delta t + Q_{1-c}\sigma\sqrt{\delta t}\right]\}$$

This corresponds exactly to Eq. 21.19, since $\widetilde{\Delta}\widehat{=}N$ for $\widetilde{\Delta} > 0$ and $\widetilde{\Delta}\widehat{=}-N$ for $\widetilde{\Delta} < 0$.

The common summand $-\widetilde{\Delta}\mu\delta t$ can now be taken out of the maximum function

$$\text{VaR}_V(c) \approx \max\left\{-\widetilde{\Delta}Q_{1-c}\sigma\sqrt{\delta t}, +\widetilde{\Delta}Q_{1-c}\sigma\sqrt{\delta t}\right\} - \widetilde{\Delta}\mu\delta t$$
$$= \left|\widetilde{\Delta}Q_{1-c}\sigma\sqrt{\delta t}\right| - \widetilde{\Delta}\mu\delta t \qquad (22.12)$$

In this approximation, the maximum function produces precisely the absolute value of the risk which is caused by the volatility of the risk factor. A positive drift μ of the risk factor reduces the portfolio risk when $\widetilde{\Delta} > 0$ (intuitively, the portfolio then represents a long position). If, on the other hand, the portfolio sensitivity is negative, i.e., $\widetilde{\Delta} < 0$, a positive drift *increases* the portfolio risk (intuitively, the portfolio represents a short position). The drift's influence is of course lost if the drift is neglected in the approximation. The value at risk then reduces to

$$\text{VaR}_V(c) \approx \left|\widetilde{\Delta}Q_{1-c}\sigma\sqrt{\delta t}\right| \qquad (22.13)$$

where the absolute value makes it immediately clear that the sign of the portfolio sensitivity no longer plays a role.

22.2.2   The Value at Risk with Respect to Several Risk Factors

In the previous section, linear approximations enabled us to reduce the VaR with respect to a single risk factor to that of a position consisting of Δ instruments representing the risk factor. We were then able, as was done in Sect. 21.3, to deduce information about the unknown distribution of the

portfolio's value V from the known distribution of the *risk factor* S. The extension of these results to the case of several risk factors is not trivial even for the delta-normal approximation. Only by using the roughest approximation in Eq. 22.1 for the change in the risk factors, namely $\delta S_i(t) \approx S_i(t) \delta Z_i$, can we manage to avoid involving the distribution of V in the discussion.

The approximation Eq. 21.26, i.e. $\delta S(t) \approx S(t) \delta Z$, led to the value at risk Eq. 22.13 with respect to a single risk factor. Squaring both sides of this equation yields

$$
\begin{aligned}
\mathrm{VaR}_V^2(c) &\approx \widetilde{\Delta}^2 \, (Q_{1-c'})^2 \, \sigma^2 \delta t \\
&= \Delta^2 \, (Q_{1-c'})^2 \, S(t)^2 \sigma^2 \delta t \\
&= \Delta^2 \, (Q_{1-c'})^2 \, S(t)^2 \, \mathrm{var} \, [\delta Z] \\
&= \Delta^2 \, (Q_{1-c'})^2 \, \mathrm{var} \, [\delta S(t)] \;,
\end{aligned}
$$

since the approximation in Eq. 21.26 allows the approximation of the variance of $\delta S(t)$ with $S(t)^2 \, \mathrm{var} \, [\delta Z]$. On the other hand, the variance of V can be calculated from Eq. 22.10 simply as

$$
\mathrm{var} \, [\delta V] \approx \mathrm{var} \, [\Delta \delta S(t)] = \Delta^2 \, \mathrm{var} \, [\delta S(t)] \;.
$$

This means that this approximation can be used to express the square of the value at risk in terms of a multiple of the *variance of the portfolio*:

$$
\begin{aligned}
\mathrm{VaR}_V^2(c) &\approx (Q_{1-c'})^2 \, \Delta^2 \, \mathrm{var} \, [\delta S(t)] \\
&= (Q_{1-c'})^2 \, \mathrm{var} \, [\delta V] \;.
\end{aligned}
$$

Now, only the variance of the portfolio's value needs to be determined for the computation of the VaR and not its *distribution* or its percentiles.

If several risk factors are involved, Eq. 22.9 can be used to write the portfolio's change in value, δV, in the approximation given by Eq. 21.26, as a linear combination of normally distributed random variables δZ_i (with deterministic coefficients $\widetilde{\Delta}_i$)

$$
\delta V \approx \sum_{i=1}^{n} \widetilde{\Delta}_i \delta Z_i \;.
$$

The variance of a sum of random variables is equal to the sum of the covariances of these random variable as can be seen from Eq. A.17. The

variance of the portfolio value is thus

$$
\begin{aligned}
\mathrm{var}\,[\delta V] &\approx \sum_{i,j=1}^{n} \widetilde{\Delta}_i \widetilde{\Delta}_j \,\mathrm{cov}\left[\delta Z_i, \delta Z_j\right] \\
&= \sum_{i,j=1}^{n} \widetilde{\Delta}_i \delta \Sigma_{ij} \widetilde{\Delta}_j \qquad\qquad (22.14) \\
&= \delta t \sum_{i,j=1}^{n} \widetilde{\Delta}_i \sigma_i \rho_{ij} \sigma_j \widetilde{\Delta}_j \;,
\end{aligned}
$$

where the definition of the covariance matrix in Eq. 21.22 was used in the last step. This means that the value at risk can be approximated as

$$
\begin{aligned}
\mathrm{VaR}_V(c) &\approx |Q_{1-c}| \sqrt{\mathrm{var}\,[\delta V]} \\
&= |Q_{1-c}| \sqrt{\widetilde{\boldsymbol{\Delta}} \delta \boldsymbol{\Sigma} \widetilde{\boldsymbol{\Delta}}} \\
&= |Q_{1-c}| \sqrt{\delta t} \sqrt{\sum_{i,j=1}^{n} \widetilde{\Delta}_i \sigma_i \rho_{ij} \sigma_j \widetilde{\Delta}_j} \;. \qquad (22.15)
\end{aligned}
$$

This is the central equation for the delta-normal method. It summarizes all assumptions, approximations, and computation methods of the delta-normal method.[4]

---

[4]If all portfolio sensitivities are non-negative (which is often the case for instance for a private investor's portfolio containing only long positions), then this equation can be rewritten in an alternative and quite intuitive form. Let $\mathrm{VaR}_i(c)$ denote the value at risk of the portfolio with respect to a particular risk factor S_i. Using Eq. 22.13, we can approximate this by

$$
\mathrm{VaR}_i(c) \approx \left| \widetilde{\Delta}_i Q_{1-c} \sigma_i \sqrt{\delta t} \right| \;.
$$

Thus, for the special case that *none* of the portfolio deltas is negative, the VaR with respect to all risk factors can be obtained by computing the square root of the weighted sum of the products of all the VaRs with respect to the individual risk factors. The weights under consideration are the respective correlations between the risk factors:

$$
\mathrm{VaR}_V(c) \approx \sqrt{\sum_{i,j=1}^{n} \mathrm{VaR}_i(c) \rho_{ij} \mathrm{VaR}_j(c)} \;\; \mathrm{falls}\, \widetilde{\Delta}_i \geq 0 \forall i \in \{1, \ldots, n\} \;. \qquad (22.16)
$$

In the linear approximation, the effect of the drifts can be subsequently introduced into the approximation. The *expected* change in portfolio value is calculated using the deltas and the drifts of the risk factors. Analogously to Eq. 22.12 for a single risk factor, this expected change is then subtracted from the value at risk of the portfolio given in Eq. 22.15:

$$\text{VaR}_V(c) \approx |Q_{1-c}| \sqrt{\delta t} \sqrt{\sum_{i,j=1}^{n} \widetilde{\Delta}_i \sigma_i \rho_{ij} \sigma_j \widetilde{\Delta}_j - \delta t \sum_i \widetilde{\Delta}_i \mu_i} \ . \quad (22.17)$$

The delta-normal approach to the calculation of the value at risk can be summarized as follows:

- Calculate the sensitivities of the portfolio with respect to all risk factors.
- Multiply the covariance matrix with the sensitivities of the portfolio and the current values of the risk factors as in Eq. 22.14 to obtain the variance of the portfolio. The covariance matrix's elements consist of the product of volatilities and correlations of the risk factors as defined in Eq. 21.22.
- Multiply the portfolio variance as in Eq. 22.15 by the liquidation period and the square of the percentile corresponding to the desired confidence interval (for example, -2.326 for 99% confidence).
- The square root of the thus obtained number is the value at risk of the entire portfolio, neglecting the effect of the drifts of the risk factors.
- The effect of the drifts can be taken into account using Eq. 22.17.

For future reference we re-write the final Value at Risk in Eq. 22.17 in terms of the covariances for the logarithmic changes and in terms of the covariances of the returns. According to Eqs. 21.27 and 21.28 the difference is an overall factor δt:

$$\text{VaR}_V(c) \approx |Q_{1-c}| \sqrt{\sum_{i,j=1}^{n} \widetilde{\Delta}_i \widetilde{\Delta}_j \operatorname{cov}\left[\delta \ln S_i, \delta \ln S_j\right] - \sum_i \widetilde{\Delta}_i \operatorname{E}\left[\delta \ln S_i\right]}$$

$$= \delta t \ |Q_{1-c}| \sqrt{\sum_{i,j=1}^{n} \widetilde{\Delta}_i \widetilde{\Delta}_j \operatorname{cov}\left[r_i, r_j\right] - \delta t \sum_i \widetilde{\Delta}_i \operatorname{E}\left[r_i\right]} \ . $$
$$(22.18)$$

Here, the r_i are the historic portfolio returns over holding periods of length δt, each return annualized.

These two forms of the VaR are very important in practice, when one is given historical time series of risk factor prices S_i or *annualized* risk factor returns r_i rather then the volatilities and correlations needed in Eq. 22.17, which otherwise would have to be calculated or bought from some market data vendor.

22.3 The Delta-Gamma Method

The delta-gamma method for calculating the portfolio's VaR makes use of the Taylor series expansion of the value of the portfolio up to and including the second order terms along with the approximation in Eq. 22.1 for the risk factors. The starting point for the delta-gamma method is thus the last line in Eq. 22.4, which when written in vector notation is given by

$$\delta V(\mathbf{S}(t)) = \widetilde{\mathbf{\Delta}}^T \delta\mathbf{Z} + \frac{1}{2}\delta\mathbf{Z}^T \widetilde{\mathbf{\Gamma}}\delta\mathbf{Z} \tag{22.19}$$

where

$$\delta\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \delta\mathbf{\Sigma}) \implies \mathrm{cov}[\delta Z_i, \delta Z_j] = \sigma_i \rho_{ij} \sigma_j \delta t, \ \mathrm{E}[\delta Z_i] = 0 \ .$$

The right-hand side of Eq. 22.19 can *not* be written as the sum of the contributions of each risk factor as was the case for the delta-normal method:

$$\sum_i^n \widetilde{\Delta}_i \, \delta Z_i + \frac{1}{2}\sum_{i,j}^n \delta Z_i \widetilde{\Gamma}_{ij} \, \delta Z_j \neq \sum_i^n (\text{contribution of the } i\text{-th risk factor}) \ .$$

The contributions of the individual risk factors can not be considered separately since they are *coupled* in the above equation by the matrix $\widetilde{\mathbf{\Gamma}}$. Furthermore the random variables δZ_j are not independent. They are *correlated* through the covariance matrix $\delta\mathbf{\Sigma}$. Two essential elements of the method presented here are[5]:

• Using the Cholesky decomposition of the covariance matrix $\delta\mathbf{\Sigma}$ to transform the δZ_j into independent random variables.

---

[5]This goes back to a paper by Rouvinez, see [166].

- Diagonalizing the gamma matrix $\widetilde{\boldsymbol{\Gamma}}$ thereby decoupling the contributions of the individual risk factors.

Also after the Cholesky decompositions, and in contrast to the delta-normal case, it is still not possible in the situation of Eq. 22.19 to reduce the VaR with respect to a risk factor to the VaR of a position consisting of Δ of these risk factors. Thus the (unknown) distribution of the *portfolio value* can no longer be substituted by the (known) distribution of the individual *risk factors*, as could still be done in Sect. 21.3. Instead, the distribution of δV must be determined *directly* in order to calculate the value at risk defined in Eqs. 21.2 or 21.3. A third essential step of the delta-gamma method presented here involves the determination of the distribution of δV.

22.3.1   Decoupling of the Risk Factors

Motivated by Eq. 21.40, we first introduce a matrix $\mathbf{A}$ satisfying the property 21.35. This matrix can be constructed through the Cholesky decomposition of the covariance matrix as described in detail in Sect. 21.5.3. This matrix transforms the correlated δZ_i into uncorrelated random variables. With this goal in mind, we rewrite Eq. 22.19, first introducing identity matrices into the equation and then replacing them with $\mathbf{A}\mathbf{A}^{-1}$ or $(\mathbf{A}^T)^{-1}\mathbf{A}^T$ as follows:

$$
\begin{aligned}
\delta V(\mathbf{S}(t)) &= \widetilde{\boldsymbol{\Delta}}^T\,\mathbf{1}\,\delta\mathbf{Z} + \frac{1}{2}\delta\mathbf{Z}^T\,\mathbf{1}\,\widetilde{\boldsymbol{\Gamma}}\,\mathbf{1}\,\delta\mathbf{Z} \\[2mm]
&= \widetilde{\boldsymbol{\Delta}}^T\,\mathbf{A}\mathbf{A}^{-1}\,\delta\mathbf{Z} + \frac{1}{2}\delta\mathbf{Z}^T\,(\mathbf{A}^T)^{-1}\mathbf{A}^T\,\widetilde{\boldsymbol{\Gamma}}\,\mathbf{A}\mathbf{A}^{-1}\,\delta\mathbf{Z} \\[2mm]
&= \widetilde{\boldsymbol{\Delta}}^T\,\mathbf{A}\mathbf{A}^{-1}\,\delta\mathbf{Z} + \frac{1}{2}\delta\mathbf{Z}^T\,(\mathbf{A}^{-1})^T\mathbf{A}^T\,\widetilde{\boldsymbol{\Gamma}}\,\mathbf{A}\mathbf{A}^{-1}\,\delta\mathbf{Z} \\[2mm]
&= \widetilde{\boldsymbol{\Delta}}^T\,\mathbf{A}\,(\mathbf{A}^{-1}\,\delta\mathbf{Z}) + \frac{1}{2}(\mathbf{A}^{-1}\delta\mathbf{Z})^T\mathbf{A}^T\,\widetilde{\boldsymbol{\Gamma}}\,\mathbf{A}\,(\mathbf{A}^{-1}\delta\mathbf{Z})\ .
\end{aligned}
$$

In the penultimate step, the property 21.39 has been used; recall that this holds for every invertible matrix. In the last step, the parentheses are intended to emphasize the fact that $\delta\mathbf{Z}$ only appears in combination with $\mathbf{A}^{-1}$. We have shown before that the components of the vector $\mathbf{A}^{-1}\delta\mathbf{Z}$ are *iid* random

variables, see Eq. 21.40. We can therefore write

$$\delta V(\mathbf{S}(t)) = \widetilde{\boldsymbol{\Delta}}^T \mathbf{A}\, \delta \mathbf{Y} + \frac{1}{2} \delta \mathbf{Y}^T \mathbf{M}\, \delta \mathbf{Y}$$

$$\text{mit } \delta \mathbf{Y} := \mathbf{A}^{-1} \delta \mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{1}), \ iid \tag{22.20}$$

$$\text{und } \mathbf{M} := \mathbf{A}^T \widetilde{\boldsymbol{\Gamma}} \mathbf{A}\ .$$

Thus, the first goal has been accomplished. The δZ_i have been transformed into *iid* random variables δY_i.

Because $\widetilde{\boldsymbol{\Gamma}}$ is by definition a symmetric matrix, i.e., $\widetilde{\Gamma}_{ij} = \widetilde{\Gamma}_{ji}$, we can show that the newly defined matrix $\mathbf{M}$ is symmetric as well:

$$M_{ij} = \left(\mathbf{A}^T \widetilde{\boldsymbol{\Gamma}} \mathbf{A}\right)_{ij} = \sum_k \sum_m (A^T)_{ik} \widetilde{\Gamma}_{km} A_{mj}$$

$$= \sum_k \sum_m (A^T)_{ik} \widetilde{\Gamma}_{mk} A_{mj} = \sum_k \sum_m A_{mj} \widetilde{\Gamma}_{mk} (A^T)_{ik}$$

$$= \sum_k \sum_m (A^T)_{jm} \widetilde{\Gamma}_{mk} A_{ki} = \left(\mathbf{A}^T \widetilde{\boldsymbol{\Gamma}} \mathbf{A}\right)_{ji} = M_{ji}\ .$$

22.3.2 Diagonalization of the Gamma Matrix

The next step is to decouple the contributions to δV of the individual random variables in Eq. 22.20. This is accomplished by diagonalizing the gamma matrix, or more precisely, the transformed gamma matrix $\mathbf{M}$ introduced in Eq. 22.20. Diagonalizing a matrix is a standard procedure in *linear algebra*. We refer the reader to the relevant literature.[6] We nevertheless take the opportunity to demonstrate the fundamental operations for *diagonalizing a matrix* here since they contain essential elements of the practical value at risk computations to be performed in the delta-gamma method.

The *eigenvectors* $\mathbf{e}^i$ of a matrix $\mathbf{M}$ are the non-zero vectors which are mapped by $\mathbf{M}$ to the same vector multiplied by a number (called a *scalar* in algebra):

$$\mathbf{M}\mathbf{e}^i = \lambda_i \mathbf{e}^i \Leftrightarrow$$

$$(\mathbf{M} - \lambda_i \mathbf{1})\, \mathbf{e}^i = \mathbf{0}\ . \tag{22.21}$$

---

[6] The most important results required for the analysis here receive a clear and concise treatment in [78], for example.

These scalars λ_i are called *eigenvalues* of the matrix. As known from linear algebra, an equation of this kind has a non-trivial solution $\mathbf{e}^i \neq \mathbf{0}$ if and only if the matrix $(\mathbf{M} - \lambda_i \mathbf{1})$ is *singular*. For this to be the case, the *determinant* of this matrix must be zero:

$$\det (\mathbf{M} - \lambda_i \mathbf{1}) = 0 . \tag{22.22}$$

The solutions of Eq. 22.22 are the eigen*values* λ_i. Having determined these values, they can be substituted into Eq. 22.21 to determine the eigen*vectors* $\mathbf{e}^i$ of the matrix. The eigenvectors have as yet only been defined up to a multiplicative scalar since if $\mathbf{e}^i$ solves Eq. 22.21 then $c\mathbf{e}^i$ does as well, for any arbitrary scalar c. The uniqueness of the eigenvalues can be guaranteed by demanding that the eigenvectors have norm 1:

$$(\mathbf{e}^i)^T \mathbf{e}^i = 1 \tag{22.23}$$

As is known from linear algebra, a symmetric, non-singular $n \times n$ matrix has n linearly independent eigenvectors which are *orthogonal*. This means that the *inner product* of each pair of different eigenvectors equals zero (graphically: the angle formed by the two vectors is 90 degrees). Together with the normalization the eigenvectors thus have the following property

$$(\mathbf{e}^i)^T \mathbf{e}^j = \sum_k e^i_k e^j_k = \delta_{ij} \tag{22.24}$$

where δ_{ij} denotes the well-known *Kronecker delta*. A collection of vectors satisfying this property is called *orthonormal*. Since we have shown that the $n \times n$ matrix $\mathbf{M}$ in Eq. 22.20 is symmetric, we can be sure that it indeed has n orthonormal eigenvectors satisfying Eq. 22.24.

To clarify the notation for these eigenvectors: The subscript k and the superscript i of e^i_k identify this value as the k-th component of the i-th eigenvector $\mathbf{e}^i$.

$$\mathbf{e}^j = \begin{pmatrix} e^j_1 \\ e^j_2 \\ \vdots \\ e^j_n \end{pmatrix}, \ (\mathbf{e}^j)^T = \left(e^j_1 \ e^j_2 \ \cdots \ e^j_n \right), \ j = 1, \ldots, n .$$

A matrix $\mathbf{O}$ can now be constructed whose column vectors are composed of the eigenvectors of $\mathbf{M}$:

$$\mathbf{O} = \left(\mathbf{e}^1 \ \mathbf{e}^2 \ \ldots \ \mathbf{e}^n \ \right) = \begin{pmatrix} e_1^1 & e_1^2 & \cdots & e_1^n \\ e_2^1 & e_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ e_n^1 & \cdots & \cdots & e_n^n \end{pmatrix} \Rightarrow O_{ij} = e_i^j \ . \tag{22.25}$$

The jth eigenvector $\mathbf{e}^j$ is in the jth column of the matrix. In the ith row, we find the ith components of all eigenvectors. As we will soon see, this matrix is an indispensable tool for the purpose of the diagonalization. As can be immediately verified[7]

$$\mathbf{O}^T \mathbf{O} = \mathbf{1} \tag{22.26}$$

and therefore also

$$\mathbf{O}^T = \mathbf{O}^{-1} \Rightarrow \mathbf{O}\mathbf{O}^T = \mathbf{1} \ . \tag{22.27}$$

Equation 22.26 characterizes a group of matrices known as *orthonormal transformations.* Applying such a matrix to a vector effects a rotation of the vector.

The eigen*values* of the matrix $\mathbf{M}$ can be used to construct a matrix as well, namely the diagonal matrix

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_n \end{pmatrix} \ . \tag{22.28}$$

From Eq. 22.21, it follows immediately that the relation

$$\mathbf{M}\mathbf{O} = \mathbf{O}\boldsymbol{\lambda} \tag{22.29}$$

---

[7]$(\mathbf{O}^T\mathbf{O})_{ij} = \sum_k (O^T)_{ik} O_{kj} = \sum_k e_k^i e_k^j = \delta_{ij} \ .$

holds for the matrices $\mathbf{M}$, $\mathbf{O}$ and $\boldsymbol{\lambda}$. Such matrix equations can often be verified quite easily by comparing the matrix elements individually:

$$(\mathbf{MO})_{ij} = \sum_k M_{ik} O_{kj} = \sum_k M_{ik} e_k^j = \left(\mathbf{Me}^j\right)_i$$

$$= \left(\lambda_j \mathbf{e}^j\right)_i = \lambda_j e_i^j = O_{ij}\lambda_j = (\mathbf{O}\boldsymbol{\lambda})_{ij} \ .$$

The decisive step in the above proof is the first equality in the second line where the eigenvector equation 22.21 was used. Multiplying both sides of this equation *from the left* by the matrix $\mathbf{O}^T$ and using Eq. 22.26 directly yields the desired diagonalization of $\mathbf{M}$, since $\boldsymbol{\lambda}$ is a diagonal matrix:

$$\mathbf{O}^T\mathbf{MO} = \mathbf{O}^T\mathbf{O}\boldsymbol{\lambda} = \boldsymbol{\lambda} \ . \tag{22.30}$$

Multiplying both sides of Eq. 22.29 *from the right* by the matrix $\mathbf{O}^T$ and using Eq. 22.27, also yields a very useful representation of $\mathbf{M}$, namely the *spectral representation* (also referred to as *eigenvector decomposition*)

$$\mathbf{M} = \mathbf{O}\boldsymbol{\lambda}\mathbf{O}^T = \sum_k \lambda_k \left(\mathbf{e}^k(\mathbf{e}^k)^T\right) \ .$$

We are now in a position to introduce the diagonalized matrix $\mathbf{O}^T\mathbf{MO}$ into Eq. 22.20 by inserting identity matrices in Eq. 22.20 and subsequently replacing them by $\mathbf{OO}^T$. Equation 22.27 ensures that equality is maintained.

$$\delta V(\mathbf{S}(t)) = \widetilde{\boldsymbol{\Delta}}^T \mathbf{A} \mathbf{1} \delta\mathbf{Y} + \frac{1}{2}\delta\mathbf{Y}^T \mathbf{1} \mathbf{M} \mathbf{1} \delta\mathbf{Y}$$

$$= \widetilde{\boldsymbol{\Delta}}^T \mathbf{A}\mathbf{O}\mathbf{O}^T \delta\mathbf{Y} + \frac{1}{2}\delta\mathbf{Y}^T \mathbf{O}\underbrace{\mathbf{O}^T\mathbf{M}\mathbf{O}}_{\boldsymbol{\lambda}}\mathbf{O}^T \delta\mathbf{Y}$$

$$= \widetilde{\boldsymbol{\Delta}}^T \mathbf{A}\mathbf{O}\left(\mathbf{O}^T \delta\mathbf{Y}\right) + \frac{1}{2}\left(\mathbf{O}^T \delta\mathbf{Y}\right)^T \boldsymbol{\lambda}\left(\mathbf{O}^T \delta\mathbf{Y}\right) \ .$$

In the last equality, the parentheses are meant to emphasize that $\delta\mathbf{Y}$ appears only in combination with $\mathbf{O}^T$. In consequence, we can write

$$\delta V(\mathbf{S}(t)) = \widetilde{\boldsymbol{\Delta}}^T \mathbf{A}\mathbf{O} \, \delta\mathbf{X} + \frac{1}{2}\delta\mathbf{X}^T \boldsymbol{\lambda}\delta\mathbf{X}$$

$$\text{mit } \delta\mathbf{X} := \mathbf{O}^T \delta\mathbf{Y} = \mathbf{O}^T\mathbf{A}^{-1}\delta\mathbf{Z} \tag{22.31}$$

$$\text{und } \boldsymbol{\lambda} := \mathbf{O}^T\mathbf{MO} = \mathbf{O}^T\mathbf{A}^T\widetilde{\boldsymbol{\Gamma}}\mathbf{A}\mathbf{O} \ .$$

The δY_i were *iid*, standard normally distributed random variables. This was accomplished in the previous section by the mapping $\mathbf{A}^{-1}$. Now to achieve the diagonalization of the gamma matrix, the random variables must undergo a further transformation under the mapping $\mathbf{O}^T$. The question remains as to whether the accomplishments of the previous section was undone by this new transformation, in other words, whether the transformed random variables have remained independent. We therefore consider the covariance of the new random variables

$$\mathrm{cov}[\delta X_i, \delta X_j] = \mathrm{cov}\left[\sum_k O_{ik}^T \delta Y_k, \sum_m O_{jm}^T Y_m\right]$$

$$= \sum_k \sum_m O_{ik}^T O_{jm}^T \underbrace{\mathrm{cov}[\delta Y_k, \delta Y_m]}_{\delta_{km}}$$

$$= \sum_k O_{ik}^T O_{jk}^T$$

$$= \sum_k O_{ik}^T O_{kj} = \left(\mathbf{O}^T \mathbf{O}\right)_{ij} = \mathbf{1}_{ij} = \delta_{ij} .$$

The covariances have remained invariant under the transformation $\mathbf{O}^T$. Thus, the new random variables are also uncorrelated and all have variance 1. Also, the zero expectation does not change under the transformation $\mathbf{O}^T$:

$$\mathrm{E}[\delta X_i] = \mathrm{E}\left[\sum_k O_{ik}^T \delta Y_k\right] = \sum_k O_{ik}^T \underbrace{\mathrm{E}[\delta Y_k]}_{0} .$$

Since matrix multiplication is a linear transformation (we are operating in the realm of *linear* algebra), the form of the distribution remains the same as well. In summary, the new random variables are uncorrelated and all have the same standard-normal distribution. We have argued after Eq. 21.40 that such variables are indeed *iid* random variables. Summarizing the above deliberations, we can write for the δX_i:

$$\delta \mathbf{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{1}), \ iid .$$

We have only used property 22.26, i.e., the *iid* property of random variables remains invariant under *every* orthonormal transformation.

If we define a "transformed sensitivity vector" as

$$\mathbf{L} := \mathbf{O}^T \mathbf{A}^T \widetilde{\boldsymbol{\Delta}}$$

(which implies for its transposed $\mathbf{L}^T = \widetilde{\boldsymbol{\Delta}}^T \mathbf{A} \mathbf{O}$), the portfolio-change Eq. 22.31 can be brought into the following simple form

$$\delta V(\mathbf{S}(t)) = \mathbf{L}^T \delta \mathbf{X} + \frac{1}{2} \delta \mathbf{X}^T \boldsymbol{\lambda} \delta \mathbf{X} \quad \text{with} \quad \delta \mathbf{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{1}) \qquad (22.32)$$

or, expressed component-wise

$$\delta V(\mathbf{S}(t)) = \sum_i \left[L_i \, \delta X_i + \frac{1}{2} \lambda_i \delta X_i^2 \right] = \sum_i \delta V_i$$

$$\text{mit } \delta V_i = L_i \, \delta X_i + \frac{1}{2} \lambda_i \delta X_i^2, \; i = 1, \ldots, n . \qquad (22.33)$$

The change in the portfolio's value is now the decoupled sum of the individual contributions of *iid* random variables as was our original intention.

At this stage, we collect all transformations involved in mapping the original random variables in Eq. 22.19 into the *iid* random variables in the above expression Eq. 22.32:

$$\delta \mathbf{X} := \mathbf{O}^T \mathbf{A}^{-1} \delta \mathbf{Z}, \; \boldsymbol{\lambda} := \mathbf{O}^T \mathbf{A}^T \widetilde{\boldsymbol{\Gamma}} \mathbf{A} \mathbf{O}, \; \mathbf{L} := \mathbf{O}^T \mathbf{A}^T \widetilde{\boldsymbol{\Delta}} .$$

From Eq. 22.27 we know that $\mathbf{O}^T \mathbf{A}^{-1} = \mathbf{O}^{-1} \mathbf{A}^{-1} = (\mathbf{A} \mathbf{O})^{-1}$. We now recognize that all of these transformations can be represented with a single matrix defined as

$$\mathbf{D} := \mathbf{A} \mathbf{O} \qquad (22.34)$$

With this matrix, the transformations become simply

$$\delta \mathbf{X} := \mathbf{D}^{-1} \delta \mathbf{Z}, \; \boldsymbol{\lambda} := \mathbf{D}^T \widetilde{\boldsymbol{\Gamma}} \mathbf{D}, \; \mathbf{L} := \mathbf{D}^T \widetilde{\boldsymbol{\Delta}} . \qquad (22.35)$$

The matrix $\mathbf{D}$ *directly* diagonalizes the gamma matrix $\widetilde{\boldsymbol{\Gamma}}$ (as opposed to $\mathbf{O}$, which diagonalizes the matrix $\mathbf{M}$). *In addition*, $\mathbf{D}$ is by definition, an orthonormal transformation (a "rotation") of $\mathbf{A}$, the Cholesky decomposition of the covariance matrix. $\mathbf{D}$ is likewise a "square root" of the covariance

matrix, since the square of a matrix remains invariant under orthonormal transformations of the matrix. Explicitly:

$$\mathbf{D}\mathbf{D}^T = \mathbf{A}\mathbf{O}(\mathbf{A}\mathbf{O})^T = \mathbf{A}\mathbf{O}\mathbf{O}^T\mathbf{A}^T = \mathbf{A}\mathbf{1}\mathbf{A}^T = \mathbf{A}\mathbf{A}^T = \delta\boldsymbol{\Sigma} \ . \tag{22.36}$$

Therefore, the matrix $\mathbf{D}$ satisfies *both* tasks, namely the decoupling of the gamma matrix *and* the transformation of the correlated random variables into uncorrelated ones.

As a little consistence check, using the matrix $\mathbf{D}$ we immediately recognize the equivalence of Eqs. 22.32 and 22.19:

$$\delta V(\mathbf{S}(t)) = \left(\mathbf{D}^T\widetilde{\boldsymbol{\Delta}}\right)^T \mathbf{D}^{-1}\delta\mathbf{Z} + \frac{1}{2}\left(\mathbf{D}^{-1}\delta\mathbf{Z}\right)^T \mathbf{D}^T \,\widetilde{\boldsymbol{\Gamma}}\, \mathbf{D}\mathbf{D}^{-1}\delta\mathbf{Z}$$

$$= \widetilde{\boldsymbol{\Delta}}^T \underbrace{\mathbf{D}\mathbf{D}^{-1}}_{\mathbf{1}}\delta\mathbf{Z} + \frac{1}{2}\delta\mathbf{Z}^T \underbrace{\left(\mathbf{D}^T\right)^{-1}\mathbf{D}^T}_{\mathbf{1}} \boldsymbol{\Gamma} \underbrace{\widetilde{\mathbf{D}\mathbf{D}^{-1}}}_{\mathbf{1}}\delta\mathbf{Z} \ ,$$

where Eq. 21.39 was again used in verifying this equivalence.

22.3.3 The Distribution of the Portfolio Value Changes

Having decoupled the individual contributions to δV in Eq. 22.33 into standard normally distributed *iid* random variables, we can now determine the distribution of the sum. δV is nevertheless not simply the sum of normally distributed random variables alone, since the expression also includes the *square* of normally distributed random variables. These additional random variables represent the difference in complexity compared to the delta-normal method. According to Sect. A.4.6, the square of a standard normally distributed random variable is χ^2-distributed with one degree of freedom. We can thus write Eq. 22.33 as

$$\delta V(\mathbf{S}(t)) = \sum_{i=1}^{n} L_i\,\delta X_i + \frac{1}{2}\sum_{i=1}^{n}\lambda_i\widetilde{X}_i \quad \text{with} \quad \delta X_i \sim \mathrm{N}(0,1),\ \widetilde{X}_i \sim \chi^2(1) \ .$$

However, the $\widetilde{X}_i$ are not independent random variables since we obviously have

$$\widetilde{X}_i = (\delta X_i)^2 \ \forall i \ .$$

We need to re-write δV in such a way that every term appearing is statistically independent of every other term. First note that a δX_i is independent of every other term in δV if and only if the corresponding eigenvalue λ_i is zero, since in this case the corresponding $\widetilde{X}_i$ does not appear in the sum. We will emphasize this by introducing the index set J which contains only the indices of non-zero eigenvalues:

$$J = \left\{1, \ldots, n \,\middle|\, \lambda_j \neq 0\right\} \,. \tag{22.37}$$

With this index set we can write[8]

$$\delta V(\mathbf{S}(t)) = \sum_{i \notin J} L_i \, \delta X_i + \sum_{j \in J} L_j \, \delta X_j + \frac{1}{2} \sum_{j \in J} \lambda_j \delta X_j^2 \tag{22.38}$$

$$= \sum_{i \notin J} L_i \, \delta X_i + \sum_{j \in J} \left[L_j \, \delta X_j + \frac{1}{2}\lambda_j \delta X_j^2 \right] \,.$$

The first sum in Eq. 22.38 is actually a sum of normally distributed random variables and as such is again a normally distributed random variable which we denote by u_0. The expectation of this random variable can be calculated as

$$\mathrm{E}\left[u_0\right] = \mathrm{E}\left[\sum_{i \notin J} L_i \, \delta X_i \right] = \sum_{i=1}^{n} L_i \underbrace{\mathrm{E}\left[\delta X_i\right]}_{0} = 0 \,.$$

and the variance is

$$\mathrm{var}\left[u_0\right] = \mathrm{var}\left[\sum_{i \notin J} L_i \, \delta X_i \right] = \sum_{i,j \notin J} \mathrm{cov}\left[L_i \, \delta X_i, \, L_j \, \delta X_j\right]$$

$$= \sum_{i,j \notin J} L_i L_j \underbrace{\mathrm{cov}\left[\delta X_i, \, \delta X_j\right]}_{\delta_{ij}} = \sum_{i \notin J} L_i^2 \,.$$

---

[8]The notation $i \notin J$ denotes all indices i with eigenvalue $\lambda_i = 0$, i.e., the set $\{1, \ldots, n \,|\, \lambda_i = 0\}$.

with the components L_i of the transformed sensitivity vector $\mathbf{L}$ defined in Eq. 22.35. Thus

$$u_0 := \sum_{i \notin J} L_i \, \delta X \sim N\left(0, \sum_{i \notin J} L_i^2\right) .$$

Consider now the sums over $j \in J$ in Eq. 22.38. To combine the *dependent* random numbers δX_j and δX_j^2 in the square brackets of Eq. 22.38 into one single random number, we complete the square for each $j \in J$:

$$\frac{1}{2}\lambda_j \delta X_j^2 + L_j \, \delta X_j = \frac{1}{2}\lambda_j \left(\delta X_j^2 + 2\frac{L_j}{\lambda_j}\delta X_j\right) = \frac{1}{2}\lambda_j \left(\delta X_j + \frac{L_j}{\lambda_j}\right)^2 - \frac{L_j^2}{2\lambda_j} .$$

Since δX_j is a standard normal random variable, we have

$$\delta X_j \sim N(0, 1) \implies \delta X_j + \frac{L_j}{\lambda_j} \sim N\left(\frac{L_j}{\lambda_j}, 1\right) .$$

Therefore, according to Eq. A.94 in Sect. A.4.6 $u_j := \left(\delta X_j + L_j/\lambda_j\right)^2$ has a *non-central* χ^2-distribution with one degree of freedom and non-central parameter L_j^2/λ_j^2:

$$\left(\delta X_j + \frac{L_j}{\lambda_j}\right)^2 =: u_j \sim \chi^2\left(1, \frac{L_j^2}{\lambda_j^2}\right) \quad \forall j \in J .$$

In summary, δV has now become a sum of non-central χ^2-distributed random variables u_j plus a normally distributed random variable u_0 (plus a constant), where all the random variables appearing are independent of each other:

$$\delta V(\mathbf{S}(t)) = u_0 + \frac{1}{2}\sum_{j \in J}\lambda_j u_j - \underbrace{\frac{1}{2}\sum_{j \in J} L_j^2/\lambda_j}_{\text{constant}} \tag{22.39}$$

$$\text{mit } u_0 \sim N(0, \sum_{i \notin J} L_i^2), \quad u_j \sim \chi^2\left(1, (L_j/\lambda_j)^2\right), \; j \in J .$$

The problem now consists in determining the distribution of the sum of independent but *differently* distributed random variables (or at least its percentiles). According to Eq. 21.4, the value at risk at a specified confidence level is then computed with precisely these percentiles or, equivalently, by inverting the cumulative distribution function of δV.

22.3.4   Moments of the Portfolio Value Distribution

We begin by calculating the moments of the random variable δV. The first moments, the expectation, the variance, etc. (see Eqs. A.22 and A.23) have intuitive interpretations and their explicit forms provide an intuitive conception of the distribution. In addition, approximations for the distribution (the Johnson approximation) and for the percentiles (the Cornish-Fisher approximation) will later be presented which can be computed with the moments.

The *Moment Generating Function* is a very useful tool for calculating the moments. The moment generating function (abbreviated as *MGF*) G_x of a random variable x with density function $\mathrm{pdf}(x)$ is defined in Sect. A.3.1 by

$$G_x(s) \equiv \mathrm{E}[e^{sx}] = \int_{-\infty}^{\infty} e^{sx}\,\mathrm{pdf}(x)dx = \sum_{n=0}^{\infty} \frac{s^n}{n!}\,\mathrm{E}[x^n]\,, \qquad (22.40)$$

where in the last step the exponential function e^{sx} has been expanded in its Taylor series. The name *moment generating function* stems from the fact, stated mathematically in Eq. A.27, that the derivatives of the function $G_x(s)$ with respect to s evaluated at $s = 0$ generate the moments of the random variable x

$$\mathrm{E}[x^n] = \left.\frac{\partial^n G_x(s)}{\partial s^n}\right|_{s=0}. \qquad (22.41)$$

As demonstrated in Sect. A.3.1, the MGF can be explicitly computed for many distributions from their integral representation, Eq. A.25. In particular, according to Eq. A.57, the MGF of u_0, i.e., the MGF of a $\mathrm{N}(0, \sum_{i \notin J} L_i^2)$ distributed random variable is given by

$$G_{u_0}(s) = \exp\left(\frac{1}{2}\,s^2 \sum_{i \notin J} L_i^2\right) \qquad (22.42)$$

while Eq. A.95 gives the MGF of u_j, i.e., of a non-central χ^2-distribution with one degree of freedom $\chi^2(1, (L_j/\lambda_j)^2)$

$$G_{u_j}(s) = \frac{1}{\sqrt{1-2s}} \exp\left\{\frac{s}{1-2s} \frac{L_j^2}{\lambda_j^2}\right\}, \quad j \in J \ . \tag{22.43}$$

This function is well-defined for $s < 1/2$, which is sufficient for our needs since as is clear from Eq. 22.41 that we are particularly interested in values of s in a neighborhood of zero.

The usefulness of the MGF in the calculation of the distribution of δV in Eq. 22.39 stems from the fact that according to Eq. A.30 the MGF of the sum of *independent* random variables x, y is simply the product of the MGFs of each of the random variables:

$$G_{x+y}(s) = G_x(s)\, G_y(s) \tag{22.44}$$

and that furthermore, from Eq. A.31

$$G_{ax+b}(s) = e^{bs} G_x(as) \ . \tag{22.45}$$

for all non-stochastic values a, b and random variables x. The MGF of δV can thus be written as the product of the each of the MGFs appearing in the sum:

$$G_{\delta V}(s) = \exp\left\{-s \sum_{j \in J} \frac{L_j^2}{2\lambda_j}\right\} G_{u_0}(s) \prod_{j \in J} G_{u_j}\left(\frac{1}{2}\lambda_j s\right) \ .$$

The MGFs of each of the individual random variables are given explicitly in Eqs. 22.42 and 22.43. Substituting s by $\lambda_j s/2$ in Eq. 22.43 yields the required MGF for the argument $\lambda_j s/2$:

$$G_{u_j}\left(\frac{1}{2}\lambda_j s\right) = \frac{1}{\sqrt{1-\lambda_j s}} \exp\left\{\frac{L_j^2}{2\lambda_j} \frac{s}{1-\lambda_j s}\right\} \ .$$

We thus obtain an *explicit* expression for the moment generating function of the distribution of the portfolio's value changes:

$$
G_{\delta V}(s) = \exp\left\{-s \sum_{j \in J} \frac{L_j^2}{2\lambda_j}\right\} \exp\left(\frac{1}{2} s^2 \sum_{i \notin J} L_i^2\right) \prod_{j \in J} \frac{1}{\sqrt{1 - \lambda_j s}} \exp\left\{\frac{L_j^2}{2\lambda_j} \frac{s}{1 - \lambda_j s}\right\}
$$

$$
= \exp\left(\frac{1}{2} s^2 \sum_{i \notin J} L_i^2\right) \prod_{j \in J} \left[\frac{1}{\sqrt{1 - \lambda_j s}} \exp\left\{\frac{L_j^2}{2\lambda_j} \frac{s}{1 - \lambda_j s}\right\} \exp\left\{-s \frac{L_j^2}{2\lambda_j}\right\}\right]
$$

$$
= \exp\left(\frac{1}{2} s^2 \sum_{i \notin J} L_i^2\right) \prod_{j \in J} \left[\frac{1}{\sqrt{1 - \lambda_j s}} \exp\left\{\frac{L_j^2}{2\lambda_j} s \left(\frac{1}{1 - \lambda_j s} - 1\right)\right\}\right].
$$

Using the same denominator in the second exp-function finally yields

$$
G_{\delta V}(s) = \exp\left(\frac{1}{2} s^2 \sum_{i \notin J} L_i^2\right) \prod_{j \in J} \frac{1}{\sqrt{1 - \lambda_j s}} \exp\left\{\frac{1}{2} L_j^2 \frac{s^2}{1 - \lambda_j s}\right\}.
$$
(22.46)

This can be simplified even further by the following trick: Since $\lambda_i = 0$ for all $i \notin J$, we can re-write the first exp-function in the following way:

$$
\exp\left(\frac{1}{2} s^2 \sum_{i \notin J} L_i^2\right) = \prod_{i \notin J} \exp\left(\frac{1}{2} s^2 L_i^2\right) = \prod_{i \notin J} \frac{1}{\sqrt{1 - \lambda_i s}} \exp\left\{\frac{1}{2} L_i^2 \frac{s^2}{1 - \lambda_i s}\right\}.
$$

Using this form in Eq. 22.46 allows us to write δV very compactly as a product over *all* indexes $j = 1, \ldots, n$

$$
G_{\delta V}(s) = \prod_{j=1}^{n} \frac{1}{\sqrt{1 - \lambda_j s}} \exp\left\{\frac{1}{2} L_j^2 \frac{s^2}{1 - \lambda_j s}\right\}.
$$
(22.47)

This function is well-defined for all $s < \min_{j \in J}\left(\frac{1}{2|\lambda_i|}\right)$, which is sufficient for our needs since because of Eq. 22.41, we are only interested in values of s which are close to zero.

Now, using Eq. 22.41, arbitrary moments of δV can be computed. We start the calculation of the first moment by introducing the abbreviation

$$a_j := \frac{1}{2} L_j^2 \frac{s^2}{1 - \lambda_j s} \ .$$

Application of the well-known *product rule* yields

$$
\begin{aligned}
\mathrm{E}[\delta V] &= \left. \frac{\partial G_{\delta V}(s)}{\partial s} \right|_{s=0} \\
&= \left. \frac{\partial}{\partial s} \prod_{j=1}^{n} \frac{e^{a_j}}{\sqrt{1 - \lambda_j s}} \right|_{s=0} \\
&= \left. \sum_{j=1}^{n} \left(\frac{\partial}{\partial s} \frac{e^{a_j}}{\sqrt{1 - \lambda_j s}} \right) \prod_{k=1, k \neq j}^{n} \frac{e^{a_k}}{\sqrt{1 - \lambda_k s}} \right|_{s=0} \ .
\end{aligned}
$$

The derivative we need to calculate is

$$
\begin{aligned}
\frac{\partial}{\partial s} \frac{e^{a_j}}{\sqrt{1 - \lambda_j s}} &= e^{a_j} \frac{\partial}{\partial s} \frac{1}{\sqrt{1 - \lambda_j s}} + \frac{1}{\sqrt{1 - \lambda_j s}} \frac{\partial}{\partial s} e^{a_j} \\
&= \frac{\frac{1}{2} \lambda_j e^{a_j}}{\left(1 - \lambda_j s\right)^{3/2}} + \frac{\frac{1}{2} L_j^2 e^{a_j}}{\sqrt{1 - \lambda_j s}} \left(\frac{2s}{1 - \lambda_j s} + \frac{\lambda_j s^2}{\left(1 - \lambda_j s\right)^2} \right) \\
&= \frac{1}{2} \frac{e^{a_j}}{\left(1 - \lambda_j s\right)^{3/2}} \left(\lambda_j + 2 L_j^2 s + \lambda_j L_j^2 \frac{s^2}{1 - \lambda_j s} \right) \ .
\end{aligned}
$$

For $s = 0$ almost all terms vanish and we are left with $\lambda_j / 2$. Thus $\mathrm{E}[\delta V]$ becomes simply

$$\mathrm{E}[\delta V] = \left. \sum_{j=1}^{n} \frac{1}{2} \lambda_j \prod_{k \in J, k \neq j} \frac{e^{a_k}}{\sqrt{1 - \lambda_k s}} \right|_{s=0} = \frac{1}{2} \sum_{j=1}^{n} \lambda_j \ .$$

The expectation of the portfolio's value changes in the delta-gamma approximation Eq. 22.4 is thus just half the sum of the eigenvalues of the transformed gamma matrix **M**. This is by definition half the *trace* of the eigenvalue matrix $\boldsymbol{\lambda}$.

With Eqs. 22.35 and 22.36 we arrive at the conclusion that the expectation of δV equals half the trace of the product of the gamma matrix and the covariance matrix[9]:

$$
E[\delta V] = \frac{1}{2} \operatorname{tr}(\boldsymbol{\lambda}) = \frac{1}{2} \operatorname{tr}\left(\mathbf{D}^T \widetilde{\boldsymbol{\Gamma}}\, \mathbf{D}\right) = \frac{1}{2} \operatorname{tr}\left(\widetilde{\boldsymbol{\Gamma}}\, \mathbf{D}\mathbf{D}^T\right) = \frac{1}{2} \operatorname{tr}\left(\widetilde{\boldsymbol{\Gamma}}\, \delta\boldsymbol{\Sigma}\right) .
$$
(22.48)

Note that the drifts of all risk factors have been neglected (see Eqs. 22.19 and 22.1). The *risk factors* are thus all approximated to be *drift-free*. But then, for a portfolio depending only *linearly* on the risk factors (or in the linear approximation of the delta-normal method) the expectation (the drift) of the portfolio value changes also equals zero. In Eq. 22.48, the expectation (the drift) of the portfolio changes is *not* zero because *non-linear* effects were taken into consideration. It is readily seen that the gamma matrix gives rise to the drift of δV in contrast to the linear sensitivities $\widetilde{\boldsymbol{\Delta}}$ which do *not* appear in 22.48.

To find out more about the distribution of δV, we proceed by computing its variance. According to Eq. A.5 the variance is the second central moment which can be calculated via Eq. A.29:

$$
\begin{aligned}
\operatorname{var}[\delta V] &= E[(\delta V - E[\delta V])^2] \\[6pt]
&= \left. \frac{\partial^2}{\partial s^2} \exp\left(-s E[\delta V]\right) G_{\delta V}(s) \right|_{s=0} \\[6pt]
&= \left. \frac{\partial^2}{\partial s^2} \exp\left(-s \frac{1}{2} \sum_{i=1}^{n} \lambda_i \right) \prod_{j=1}^{n} \frac{\exp\left(\frac{1}{2} L_j^2 \frac{s^2}{1-\lambda_j s}\right)}{\sqrt{1-\lambda_j s}} \right|_{s=0} \\[6pt]
&= \left. \frac{\partial^2}{\partial s^2} \prod_{j=1}^{n} \frac{1}{\sqrt{1-\lambda_j s}} \exp\left(\frac{1}{2} L_j^2 \frac{s^2}{1-\lambda_j s} - \frac{1}{2}\lambda_j s\right) \right|_{s=0} \\[6pt]
&= \left. \frac{\partial^2}{\partial s^2} \prod_{j=1}^{n} a_j \right|_{s=0}
\end{aligned}
$$
(22.49)

---

[9] Here the well-known cyclic property of the trace has been used: $\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA})$ for arbitrary matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$.

with the abbreviation

$$a_j := \frac{1}{\sqrt{1 - \lambda_j s}} \exp\left(\frac{1}{2} L_j^2 \frac{s^2}{1 - \lambda_j s} - \frac{1}{2}\lambda_j s\right) .$$

The second derivative of this product is quite involved. We nonetheless present it explicitly here to demonstrate how moments of δV can be determined in practice. Such moments are needed quite often, for instance for the Cornish-Fisher expansion. We start by repeatedly applying the product rule to arrive at

$$\frac{\partial^2}{\partial s^2} \prod_{j=1}^{n} a_j = \frac{\partial}{\partial s}\left(\sum_{j=1}^{n} \frac{\partial a_j}{\partial s} \prod_{k=1,k\neq j}^{n} a_k\right)$$

$$= \sum_{j=1}^{n}\left(\frac{\partial}{\partial s}\frac{\partial a_j}{\partial s}\right)\prod_{k=1,k\neq j}^{n} a_k + \sum_{j=1}^{n} \frac{\partial a_j}{\partial s}\frac{\partial}{\partial s}\prod_{k=1,k\neq j}^{n} a_k$$

$$= \sum_{j=1}^{n} \frac{\partial^2 a_j}{\partial s^2}\prod_{k=1,k\neq j}^{n} a_k + \sum_{j=1}^{n} \frac{\partial a_j}{\partial s}\sum_{k=1,k\neq j}^{n} \frac{\partial a_k}{\partial s}\prod_{\substack{m=1,\\ m\neq k,\, m\neq j}}^{n} a_m .$$

$$(22.50)$$

Thus, we mainly have to differentiate a_j. For ease of notation, we introduce yet another abbreviation, namely

$$b_j := \frac{1}{2}L_j^2\frac{s^2}{1 - \lambda_j s} - \frac{1}{2}\lambda_j s \implies a_j = \frac{e^{b_j}}{\sqrt{1 - \lambda_j s}} .$$

The first derivative with respect to a_j is now calculated as

$$\frac{\partial a_j}{\partial s} = \frac{1}{\sqrt{1 - \lambda_j s}}\frac{\partial e^{b_j}}{\partial s} + e^{b_j}\frac{\partial}{\partial s}\frac{1}{\sqrt{1 - \lambda_j s}}$$

$$= \frac{e^{b_j}}{\sqrt{1 - \lambda_j s}}\left(\frac{s\, L_j^2}{1 - \lambda_j s} + \frac{\frac{1}{2}s^2 L_j^2 \lambda_j}{\left(1 - \lambda_j s\right)^2} - \frac{\lambda_j}{2}\right) + \frac{\frac{1}{2}e^{b_j}\lambda_j}{\left(1 - \lambda_j s\right)^{3/2}}$$

$$= \frac{s L_j^2 e^{b_j}}{\left(1 - \lambda_j s\right)^{3/2}} + \frac{\frac{1}{2}s^2 L_j^2 e^{b_j}\lambda_j}{\left(1 - \lambda_j s\right)^{5/2}} - \frac{\frac{1}{2}e^{b_j}\lambda_j}{\left(1 - \lambda_j s\right)^{1/2}} + \frac{\frac{1}{2}e^{b_j}\lambda_j}{\left(1 - \lambda_j s\right)^{3/2}} .$$

$$(22.51)$$

For $s = 0$ the first two terms in the last line vanish and the last two terms just compensate each other so that $\partial a_j / \partial s$ vanishes completely at $s = 0$:

$$\left. \frac{\partial a_j}{\partial s} \right|_{s=0} = -\frac{1}{2} e^{b_j} \lambda_j + \frac{1}{2} e^{b_j} \lambda_j = 0 \; . \tag{22.52}$$

Therefore only the term involving the second derivative in Eq. 22.50 contributes to Eq. 22.49. Using the result 22.51, this second derivative is explicitly:

$$\frac{\partial^2 a_j}{\partial s^2} = \frac{\partial}{\partial s} \left[\frac{s L_j^2 e^{b_j}}{\left(1 - \lambda_j s\right)^{3/2}} + \frac{\frac{1}{2} s^2 L_j^2 e^{b_j} \lambda_j}{\left(1 - \lambda_j s\right)^{5/2}} - \frac{\frac{1}{2} e^{b_j} \lambda_j}{\left(1 - \lambda_j s\right)^{1/2}} + \frac{\frac{1}{2} e^{b_j} \lambda_j}{\left(1 - \lambda_j s\right)^{3/2}} \right]$$

$$= L_j^2 \frac{\partial e^{b_j}}{\partial s} \frac{s}{\left(1 - \lambda_j s\right)^{3/2}} + L_j^2 e^{b_j} \frac{3}{2} \frac{\lambda_j s}{\left(1 - \lambda_j s\right)^{5/2}} + L_j^2 e^{b_j} \frac{1}{\left(1 - \lambda_j s\right)^{3/2}}$$

$$+ \frac{1}{2} L_j^2 \frac{\partial e^{b_j}}{\partial s} \frac{\lambda_j s^2}{\left(1 - \lambda_j s\right)^{5/2}} + \frac{5}{4} L_j^2 e^{b_j} \frac{\lambda_j^2 s^2}{\left(1 - \lambda_j s\right)^{7/2}} + L_j^2 e^{b_j} \frac{\lambda_j s}{\left(1 - \lambda_j s\right)^{5/2}}$$

$$- \frac{1}{2} \frac{\partial e^{b_j}}{\partial s} \frac{\lambda_j}{\left(1 - \lambda_j s\right)^{1/2}} - \frac{1}{4} e^{b_j} \frac{\lambda_j^2}{\left(1 - \lambda_j s\right)^{3/2}}$$

$$+ \frac{1}{2} \frac{\partial e^{b_j}}{\partial s} \frac{\lambda_j}{\left(1 - \lambda_j s\right)^{3/2}} + \frac{3}{4} e^{b_j} \frac{\lambda_j^2}{\left(1 - \lambda_j s\right)^{3/2}} \; .$$

Most terms above have s as a factor. They all vanish for $s = 0$. The only terms remaining are:

$$\left. \frac{\partial^2 a_j}{\partial s^2} \right|_{s=0} = L_j^2 e^{b_j} - \frac{1}{2} \frac{\partial e^{b_j}}{\partial s} \lambda_j - \frac{1}{4} e^{b_j} \lambda_j^2 + \frac{1}{2} \frac{\partial e^{b_j}}{\partial s} \lambda_j + \frac{3}{4} e^{b_j} \lambda_j^2$$

$$= \underbrace{e^{b_j}}_{1 \text{ for } s=0} \left(L_j^2 + \frac{1}{2} \lambda_j^2 \right) \; . \tag{22.53}$$

Inserting all these results into Eq. 22.49 finally yields

$$\mathrm{var}[\delta V] = \sum_{j=1}^{n} \frac{\partial^2 a_j}{\partial s^2} + \sum_{j=1}^{n} \frac{\partial a_j}{\partial s} \sum_{k=1, k \neq j}^{n} \left. \frac{\partial a_k}{\partial s} \right|_{s=0}$$

$$
= \sum_{j=1}^{n} \frac{\partial^2 a_j}{\partial s^2}\bigg|_{s=0}
$$

$$
= \sum_{j=1}^{n} \left(L_j^2 + \frac{1}{2}\lambda_j^2 \right) ,
$$

In the first step we used Eq. 22.50 and $a_k|_{s=0} = 1$. In the second step the result 22.52 was inserted and in the third step the result 22.53. The sum $\sum L_j^2$ is just the square of the transformed sensitivity *vector* and $\sum \lambda_j^2$ is the trace of the square of the *matrix* of eigenvalues, i.e.,

$$
\mathrm{var}[\delta V] = \mathbf{L}^T \mathbf{L} + \frac{1}{2} \,\mathrm{tr}\left(\boldsymbol{\lambda}^2\right) .
$$

Finally, making use of the transformations in Eq. 22.35 and applying Eq. 22.36, the variance of the portfolio's value change in the framework of the delta-gamma method becomes

$$
\mathrm{var}[\delta V] = \widetilde{\boldsymbol{\Delta}}^T \mathbf{D}\mathbf{D}^T \widetilde{\boldsymbol{\Delta}} + \frac{1}{2} \,\mathrm{tr}\left(\mathbf{D}^T \widetilde{\boldsymbol{\Gamma}} \, \mathbf{D}\mathbf{D}^T \widetilde{\boldsymbol{\Gamma}} \, \mathbf{D}\right)
$$

$$
= \widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma} \widetilde{\boldsymbol{\Delta}} + \frac{1}{2} \,\mathrm{tr}\left(\widetilde{\boldsymbol{\Gamma}} \, \delta\boldsymbol{\Sigma} \widetilde{\boldsymbol{\Gamma}} \, \delta\boldsymbol{\Sigma}\right) . \qquad (22.54)
$$

Note that the first term resulting from the linear portfolio sensitivities $\widetilde{\boldsymbol{\Delta}}$ is identical to the portfolio variance in the delta-normal method (see Eq. 22.14). The non-linear sensitivities $\widetilde{\boldsymbol{\Gamma}}$ effect a correction of the linear portfolio variance which has a form similar to the drift correction from the non-linear term in Eq. 22.48. While in Eq. 22.48, the trace of the product of the gamma matrix and the covariance matrix was relevant, now the trace of the *square* of this product is required for computing the variance.

The variance is the second *central* moment of the random variable. The *central moments* μ_i of a random variable are defined in general terms in Eq. A.23 as the "expectation of powers of the deviation from the expectation":

$$
\mu_i := \mathrm{E}[(\delta V - \mathrm{E}[\delta V])^i], \ i > 1 .
$$

Analogously to the approach for the first two moments demonstrated above, we can continue to calculate the further moments of δV. The first (central)

moments are compiled here:

$$\mu := \mathrm{E}[\delta V] = \frac{1}{2} \operatorname{tr}\left(\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma}\right)$$

$$\mu_2 = \mathrm{E}[(\delta V - \mathrm{E}[\delta V])^2] = \widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma}\,\widetilde{\boldsymbol{\Delta}} + \frac{1}{2}\operatorname{tr}\left((\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma})^2\right)$$

$$\mu_3 = \mathrm{E}[(\delta V - \mathrm{E}[\delta V])^3] = 3\widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma}\,\widetilde{\boldsymbol{\Gamma}}\,\boldsymbol{\delta\Sigma}\,\widetilde{\boldsymbol{\Delta}} + \operatorname{tr}\left((\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma})^3\right) \qquad (22.55)$$

$$\mu_4 = \mathrm{E}[(\delta V - \mathrm{E}[\delta V])^4] = 12\widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma}\,(\widetilde{\boldsymbol{\Gamma}}\,\boldsymbol{\delta\Sigma})^2\,\widetilde{\boldsymbol{\Delta}} + 3\operatorname{tr}\left((\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma})^4\right) + 3\mu_2^2 \ .$$

In this way, a great deal of additional information about the distribution of δV can be generated. For instance *skewness* and *kurtosis* of the distribution of δV are[10]

$$\text{Schiefe } \equiv \frac{\mu_3}{\mu_2^{3/2}} = \frac{3\widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma}\,\widetilde{\boldsymbol{\Gamma}}\,\boldsymbol{\delta\Sigma}\,\widetilde{\boldsymbol{\Delta}} + \operatorname{tr}(\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma})^3}{\left(\widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma}\,\widetilde{\boldsymbol{\Delta}} + \frac{1}{2}\operatorname{tr}(\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma})^2\right)^{3/2}}$$

$$\text{Kurtosis } \equiv \frac{\mu_4}{\mu_2^2} = \frac{12\widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma}\,(\widetilde{\boldsymbol{\Gamma}}\,\boldsymbol{\delta\Sigma})^2\,\widetilde{\boldsymbol{\Delta}} + 3\operatorname{tr}\left((\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma})^4\right) + 3\mu_2^2}{\left(\widetilde{\boldsymbol{\Delta}}^T \delta\boldsymbol{\Sigma}\,\widetilde{\boldsymbol{\Delta}} + \frac{1}{2}\operatorname{tr}(\widetilde{\boldsymbol{\Gamma}}\,\delta\boldsymbol{\Sigma})^2\right)^2} \ .$$

A *percentile*, however, is needed for the computation of the value at risk as given in Eq. 21.4.

Johnson Transformation

Computation of a percentile necessitates knowledge of the distribution of the random variable *directly* and not of its moments. In order to be able to proceed, we could assume a particular functional form of the distribution and then establish a relation between the parameters of this functional form and the moments of the random variable via *moment matching*. Since the moments, as shown above, can be explicitly computed, the parameters of the assumed distribution can thus be determined. For example, if we assume that a random variable is normally or lognormally distributed, we would take Eqs. 22.48 and 22.54 as parameter values. Additional functional forms for

---

[10]Recall that a normal distribution has skewness 0 and kurtosis 3, see Eq. A.59.

approximating the distribution of δV were suggested by Johnson [114]. These *Johnson transformations* have four parameters which can be determined from the first four moments in Eq. 22.55. They represent a substantially better approximation than, for example a lognormal distribution.

Cornish-Fisher Expansion

One possibility of approximating the *percentiles* of a distribution from its *central moments* and the percentiles $Q^{N(0,1)}$ of the standard normal distribution is the *Cornish-Fisher expansion*. Since this expansion makes use of the *standard* normal distribution, we must first transform δV into a centered and normalized random variable $\widetilde{\delta V}$ with expectation 0 and variance 1. This is accomplished by defining

$$\widetilde{\delta V} := \frac{\delta V - \mathrm{E}[\delta V]}{\sqrt{\mathrm{var}[\delta V]}} = \frac{\delta V - \mu}{\sqrt{\mu_2}} \ .$$

The percentile of the distribution of $\widetilde{\delta V}$ can now be approximated with the Cornish-Fisher expansion [41, 196] as follows

$$
\begin{aligned}
Q^{\mathrm{cpf}_{\widetilde{\delta V}}} \approx{} & Q^{N(0,1)} + \frac{1}{6}\left[(Q^{N(0,1)})^2 - 1\right]\frac{\mu_3}{\mu_2^{3/2}} \\
& + \frac{1}{24}\left[(Q^{N(0,1)})^3 - 3Q^{N(0,1)}\right]\left(\frac{\mu_4}{\mu_2^2} - 3\right) \\
& - \frac{1}{36}\left[2\,(Q^{N(0,1)})^3 - 5Q^{N(0,1)}\right]\left(\frac{\mu_3}{\mu_2^{3/2}}\right)^2 ,
\end{aligned}
\tag{22.56}
$$

where the expansion is taken up to the order, which uses only the first four moments from Eq. 22.55. The probability that $\widetilde{\delta V}$ is less than a number a is, naturally, the same as the probability that δV is less than $\mu + \sqrt{\mu_2}\,a$. Thus

$$Q^{\mathrm{cpf}_{\delta V}} = \mu + \sqrt{\mu_2}\,Q^{\mathrm{cpf}_{\widetilde{\delta V}}} \ .$$

holds for the percentiles. From Eq. 21.4, the value at risk is thus

$$\mathrm{VaR}(c) = -Q^{\mathrm{cpf}_{\delta V}}_{1-c} = -\mu - \sqrt{\mu_2}\,Q^{\mathrm{cpf}_{\widetilde{\delta V}}}_{1-c} \ ,$$

where we now can use the approximation 22.56 for $Q_{1-c}^{\widetilde{\delta V}}$ since the percentiles of the standard normal distribution for the confidence level $(1 - c)$ are well known (see for instance Eq. 21.13).

22.3.5 Fourier-Transformation of the Portfolio Value Distribution

While a great deal of information about the distribution function can be gleaned from the computation of the moments with the moment generating function, we are still not able to calculate the distribution itself directly. *Characteristic functions* (CFs), however, generate the distribution directly (this can at least be accomplished numerically). As defined in Sect. A.3.2, the *characteristic function* Φ_x of a random variable x with density function $\mathrm{pdf}(x)$ is[11]

$$\Phi_x(s) \equiv \mathrm{E}[e^{isx}] = \int_{-\infty}^{\infty} e^{isx}\, \mathrm{pdf}(x)dx \ . \tag{22.57}$$

This is precisely the definition of the *Fourier transformation* of the density function. As demonstrated in Sect. A.3.2, the CFs of many random variables with density functions can be computed explicitly. In particular, the CF of u_0, i.e., the CF of a normally distributed random variable $\mathrm{N}(0, \sum_{i \notin J} L_i^2)$ is given by

$$\Phi_{u_0}(s) = \exp\left(-\frac{1}{2}s^2 \sum_{i \notin J} L_i^2\right) \tag{22.58}$$

And according to Eq. A.97, the CF of the u_j , i.e., of a non-central χ^2-distributed random variable with one degree of freedom $\chi^2(1, (L_j/\lambda_j)^2)$ is given by

$$\Phi_{u_j}(s) = \frac{1}{\sqrt{1 - 2is}} \exp\left\{\frac{is}{1 - 2is}\frac{L_j^2}{\lambda_j^2}\right\}, \ j \in J, \ i \equiv \sqrt{-1} \ . \tag{22.59}$$

---

[11]Here, i denotes the *imaginary number* satisfying the property $i^2 = -1$, thus intuitively $i = \sqrt{-1}$.

Similarly to the moment generating function, the usefulness of the CF in computing the distribution of δV in Eq. 22.39 stems from the property A.34. The CF of a sum of independent random variables x, y is simply the product of the CFs of each of these variables:

$$\Phi_{x+y}(s) = \Phi_x(s)\,\Phi_y(s) \tag{22.60}$$

Furthermore, according to Eq. A.35,

$$\Phi_{ax+b}(s) = e^{ibs}\,\Phi_x(as)\,. \tag{22.61}$$

holds for all non-stochastic values a, b and random variables x. Thus the CF of δV can be expressed as the product of the CFs of the random variables appearing in the definition of δV:

$$\Phi_{\delta V}(s) = \exp\left\{-is\sum_{j\in J}\frac{L_j^2}{2\lambda_j}\right\}\Phi_{u_0}(s)\prod_{j\in J}\Phi_{u_j}\left(\frac{1}{2}\lambda_j s\right)\,. \tag{22.62}$$

The CFs of the individual random variables are given explicitly in Eqs. 22.58 and 22.59. We now have all information at our disposal to calculate an *explicit* expression for the characteristic function of the distribution of δV. The result is of course the same as Eq. 22.46 for the MGF with the obvious substitution $s \rightarrow is$:

$$\Phi_{\delta V}(s) = \exp\left(-\frac{1}{2}s^2\sum_{i\notin J}L_i^2\right)\prod_{j\in J}\frac{1}{\sqrt{1-i\lambda_j s}}\exp\left\{-\frac{1}{2}L_j^2\frac{s^2}{1-i\lambda_j s}\right\}\,. \tag{22.63}$$

Since $\lambda_i = 0$ for $i \notin J$, we can—as we did with the MGF—write the first exp-function as

$$\exp\left(-\frac{1}{2}s^2\sum_{i\notin J}L_i^2\right) = \prod_{i\notin J}\frac{1}{\sqrt{1-i\lambda_i s}}\exp\left\{-\frac{1}{2}L_i^2\frac{s^2}{1-i\lambda_i s}\right\}\,.$$

Thus, the characteristic function can be written as a product over *all* indexes j:

$$\Phi_{\delta V}(s) = \prod_{j=1}^{n} \frac{1}{\sqrt{1 - i\lambda_j s}} \exp\left\{ -\frac{1}{2} L_j^2 \frac{s^2}{1 - i\lambda_j s} \right\} \text{ mit } i \equiv \sqrt{-1} \ .$$

(22.64)

In contrast to the moment generating function, there exists an inverse transformation for the characteristic function, namely the *inverse Fourier Transformation* (see Sect. A.3.2). From Eq. 22.64 the density function $\text{pdf}(\delta V)$ can thus be computed (at least numerically).

$$
\begin{aligned}
\text{pdf}_{\delta V}(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx} \Phi_{\delta V}(s) ds \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx} \prod_{j=1}^{n} \frac{\exp\left\{ -\frac{1}{2} L_j^2 \frac{s^2}{1 - i\lambda_j s} \right\}}{\sqrt{1 - i\lambda_j s}} ds \ .
\end{aligned}
$$

(22.65)

The *cumulative* probability function of δV can now be obtained through the (numerical) integration of this probability density

$$
\begin{aligned}
\text{cpf}_{\delta V}(c) &\equiv \int_{-\infty}^{c} \text{pdf}_{\delta V}(x) dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{c} \int_{-\infty}^{\infty} e^{-isx} \prod_{j=1}^{n} \frac{\exp\left\{ -\frac{1}{2} L_j^2 \frac{s^2}{1 - i\lambda_j s} \right\}}{\sqrt{1 - i\lambda_j s}} ds \, dx \ .
\end{aligned}
$$

A method which can likewise be applied in practice does not use the Fourier transformation of the density, but the Fourier transformation of the *cumulative* distribution directly:

$$F_x(s) \equiv \int_{-\infty}^{\infty} e^{isx} \text{cpf}(x) dx \ .$$

This Fourier transformation has the analogous properties to those indicated in Eqs. 22.61 and 22.60. Hence, the Fourier transformation of the cumulative distribution function of the portfolio's change is (up to a constant)

$$F_{\delta V}(s) \sim F_{u_0}(s) \prod_{j \in J} F_{u_j}\left(\frac{1}{2} \lambda_j s \right) \ .$$

(22.66)

analogous to Eq. 22.62. Here, the individual Fourier transformations of the *cumulative* distribution function of a normally distributed random variable (for u_0) and the *cumulative* distribution of the non-central χ^2 random variables with one degree of freedom (for the u_j) appear. These can only be computed numerically. Then, the results as shown in Eq. 22.66 are multiplied in Fourier space. Finally, the function $F_{\delta V}(s)$ thus obtained must be transformed back with the inverse Fourier transformation to obtain the cumulative probability function of δV as

$$\text{cpf}_{\delta V}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx} F_{\delta V}(s) ds \; . \tag{22.67}$$

The Wiener-Chintchine theorem states that the approach just described is equivalent to taking the *convolution* of the cumulative distribution functions. The twice computed Fourier transformation is numerically preferable to computing the convolution. The recommended method for numerically performing Fourier transformations (or inverse Fourier transformations like Eq. 22.67) is the *fast Fourier transformation* (FFT). This method requires significantly fewer computations as compared to other common procedures for numerical integration.[12]

22.3.6 Monte Carlo Simulation of the Portfolio Value Distribution

Calculating the cumulative distribution of δV with characteristic functions involves complicated numerical procedures. Using moment-generating functions to calculate the moments we need additional assumptions and approximations to establish a relation between those moments and the distribution or the percentiles.[13] All methods introduced here therefore offer sufficient scope for error to creep into the calculations. Additionally, significant difficulties are often involved in calculating the gamma and covariance matrices. Hence, it is by all means legitimate to apply a simple Monte Carlo simulation, instead of the often complicated methods described above, to generate the distribution of δV. The statistical error in doing so is often no larger than the errors of the

---

[12]In contrast to other common numerical procedures, the FFT reduces the number of necessary multiplications from order $O(N^2)$ to $O(N \ln(N))$. See, for example [20] or [156].

[13]It should not be forgotten that the Delta-Gamma method itself is only an approximation of the portfolio's value obtained from the second-order Taylor series approximation.

above methods assuming of course that a sufficient number of simulation runs have been carried out.

For a Monte Carlo Simulation we proceed directly from Eq. 22.33. n normally distributed random numbers are generated with which the simulated change in the portfolio's value can be immediately computed with the expression

$$\delta V = \sum_{i=1}^{n} \left[L_i\, \delta X_i + \frac{1}{2} \lambda_i \delta X_i^2 \right] \, .$$

This procedure is repeated N times (as a rule, several thousand times) thus obtaining N simulated changes in the portfolio's value from which the distribution of δV can be approximated. The percentiles of this distribution can be approximated by simply sorting the simulated values of δV in increasing order as described in Sect. 23.1 (a detailed discussion of value at risk computations using the Monte Carlo Method can be found in this section).

Here, in contrast to the method described in Sect. 23.1, we do not simulate each single risk factor separately. Instead, the portfolio change is directly calculated based on Eq. 22.33. Therefore, no time consuming revaluation of the portfolio for the each generated scenarios is required. However, before the simulation can be performed, the eigenvalues of the transformed gamma matrix must be calculated by solving Eq. 22.22, and the transformed sensitivities L_i need to be determined as well. Because of Eqs. 22.35 and 22.34, both the Cholesky decomposition of the covariance matrix as well as the eigenvectors of the gamma matrix must be computed. The eigenvectors are determined by solving Eq. 22.21.

23

Check for
updates

Simulation Methods

23.1 Monte Carlo Simulation

In the calculation of the value at risk by means of Monte Carlo simulations, all of the risk factors influencing a portfolio are simulated over the liquidation period δt as stochastic processes satisfying, for example, Eq. 2.17 or even more general processes of the form 2.19. The value at risk as a function of the risk factors themselves are taken into complete consideration using Eq. 21.16 sometimes neglecting the drift in the simulation if the liquidation period is short:

$$\text{VaR}_{\text{long}}(c) \approx N\,S(t) \left[1 - \exp\left(+Q_{1-c}^{\text{N}(0,1)} \sigma \sqrt{\delta t} \right) \right]$$

$$\text{VaR}_{\text{short}}(c) \approx -N\,S(t) \left[1 - \exp\left(-Q_{1-c}^{\text{N}(0,1)} \sigma \sqrt{\delta t} \right) \right] \;.$$

As explained in Sect. 21.2, the value at risk of a long position in an underlying is only then equal to that of a short position if the drift is neglected *and* the linear approximation has been used. Since the linear approximation is usually not assumed in the Monte Carlo method, the VaR value of a long position will not equal that of a short position on the same underlying.

In carrying out the simulation, it will be taken into consideration that the risk factors are not independent of one another, but are *correlated*. This has already been demonstrated in Sect. 11.3.2 for the case of *two* correlated underlying prices. Processes of the form 2.17 involve Wiener processes whose stochastic components are coupled as given by the covariance matrix 21.22.

In other words, the logarithmic changes in the risk factors are multivariate normally distributed with the covariance matrix 21.22. In this way, market scenarios (combinations of all risk factors) possibly occurring up to the end of the liquidation period (up to time T) are simulated. The *portfolio* values at the conclusion of the liquidation period are then computed on the basis of all these simulated market scenarios. With this information, the distribution of the potential *portfolio* values at the conclusion of the liquidation period can be approximated. The value at risk can then be obtained through the statistical evaluation of this portfolio value distribution.

The advantage of the Monte Carlo simulation compared to other methods (such as the variance-covariance method) is that for the portfolio valuation for each market scenario we can in principle use the same valuation methods as for determining the portfolio's *current* value. No additional approximations for the *valuation* of the financial instruments in the portfolio need to be made (called *full valuation*). In principal, the same valuation methods used for the daily valuation could also be used for the value at risk calculation Nevertheless, it is in practice often not possible to use the same (computationally intensive) routines for both the valuation of a portfolio with respect to, for example, 10,000 scenarios as for (one single) determination of the portfolio's current value, the *mark-to-market*. It is therefore often necessary to use simpler and less precise methods for the revaluation of financial instruments with respect to the Monte Carlo scenarios. The statistical error arising in connection with such simulation methods is also unavoidable since only a finite number of scenarios can be simulated and thus only mean values rather than expectations can be computed (see Sect. 31.2 for more on this subject).

23.1.1 The Risk Factors as Correlated Random Walks

A random number generator usually produces single, uncorrelated random numbers. However, with the methods described in Sect. 21.5.3, via the Cholesky decomposition **A** of the covariance matrix, uncorrelated random numbers can be transformed into correlated ones. This can be exploited when carrying out Monte Carlo simulations:

With the help of the Cholesky decomposition **A** of the covariance matrix standard normally distributed random variables X_j are transformed into components of a multivariate normally distributed random vector Y_j, for $j = 1, 2, \ldots, n$ whose covariances are given in Eq. 21.22:

$$Y_i = \sum_{j=1}^{n} A_{ij} X_j \ .$$

The random walks of the risk factors expressed in terms of the random variables Y_j are

$$d \ln S_j(t + \delta t) = \mu_j \, \delta t + Y_j \,.$$

Similar to Eq. 11.2, the values of the risk factors for a scenario simulated to occur at the end of the time interval δt are

$$\ln S_j(t + \delta t) = \ln S_j(t) + \mu_j \, \delta t + Y_j$$
$$S_j(t + \delta t) = e^{Y_j} e^{\mu_j \, \delta t} S_j(t) \quad j = 1, \, \ldots \, , n \,, \tag{23.1}$$

where here, as has received mention on numerous occasions, the drifts μ_j are often neglected in the analysis.

To generate a complete market scenario for the time $t + \delta t$ a random number Y_j for each risk factor is required. The portfolio is then revaluated at the value date $t + \delta t$ on the basis of this scenario. We thus obtain a simulated portfolio value at the end of the liquidation period. The approach in a Monte Carlo simulation in risk management is summarized below. This type of simulation is sometimes called *structured Monte Carlo*.

23.1.2 Structured Monte Carlo

Simulation

- Generate n standard normally distributed, uncorrelated random numbers X_j, one for each risk factor.
- Generate correlated random numbers Y_j, $j = 1, 2, \ldots, n$ using the equation

$$Y_i = \sum_{j=1}^{n} A_{ij} X_j \,.$$

  The elements of the matrix **A** are given through the Cholesky decomposition of the covariance matrix in accordance with Eq. 21.41.
- Using these Y_j, the risk factors for the simulated scenario at the end of a time interval δt are calculated via Eq. 23.1. If the portfolio contains path-dependent derivatives, it is not possibly to simply jump to the end of the liquidation period in a single step if the liquidation period is longer than

one day. Smaller steps are necessary to simulate the paths of the risk factors up to the conclusion of the liquidation period similar to Eq. 11.1, instead of simulating directly with Eq. 11.2.

- Perform a new valuation of the portfolio with respect to these simulated risk factors. If computationally possible a *full valuation* is preferable.[1]

Thus, one single market scenario is simulated and the portfolio is re-valued with respect to this single scenario. This simulation is now repeated (for example, 10,000 times) in order to generate numerous scenarios and a portfolio value for each of these scenarios. Finally, the statistical evaluation is performed.

Evaluation

The change in the value of the portfolio observed in the i-th simulated scenario will be denoted by δV_i, the vector containing all risk factor values in the i-th simulated scenario by $\mathbf{S}_i$. We let m denote the number of simulations and n the number of relevant risk factors. For every simulated scenario, the induced simulated value change of the portfolio is the difference between the portfolio's value with respect to the simulated scenario and its current value:

$$\delta V_i = V\left(\mathbf{S}_i(t + \delta t)\right) - V\left(\mathbf{S}(t)\right) \quad \text{with} \quad i = 1, \ldots, m . \qquad (23.2)$$

We thus obtain m simulated value changes. The value at risk of the Monte Carlo simulation is the minimum of these δV_i, where a certain number of the least favorable value changes are ignored dependent on the desired confidence level. For 95% confidence, for example, these are 5% of the least favorable value changes. For 10,000 simulated scenarios, for example, the 500 worst scenarios are ignored. We denote by δV_{1-c} the most favorable of the value changes which are ignored at a level of confidence c. The value at risk of the portfolio is now the least favorable result among the set of results remaining after those $(1 - c)\%$ least favorable simulations have been removed from consideration or equivalently, the least favorable value greater than V_{1-c}:

$$\mathrm{VaR}_V(c) = -\min_i \left\{ \delta V_i \mid \delta V_i > \delta V_{1-c} \right\} \quad \text{with} \quad i = 1, \ldots, m .$$
$$(23.3)$$

---

[1]If the portfolio valuation requires for example Monte Carlo *pricing* methods for some (exotic) financial instruments, these additional Monte Carlo simulations (for pricing) have to run *inside* the simulation loop for the VaR calculation. Clearly this may lead to unacceptably large computation times.

With a confidence c, the portfolio will depreciate in value by no more than this value at risk by the end of the liquidation period.

23.2   Historical Simulation

Historical simulations are performed by investigating historical time series with the objective of identifying market changes which have actually occurred in the past and using these changes to compute the value at risk. The covariance matrix in Eq. 21.22 is not necessary for a historical simulation nor is it necessary to assume that the risk factors behave as random walks with constant yields and volatilities or even that they behave as random walks at all! This freedom from model assumptions is the primary advantage of this method.

The independence from model assumptions is at the expense of involved data management. While only three values provide sufficient statistical information about the past behavior of two risk factors (both volatilities and the correlation between the two) for variance-covariance and Monte Carlo methods, entire time series of prices for all risk factors relevant to the portfolio must be kept for a historical simulation to be performed. For example, the closing prices of every underlying for the previous 250 days. For two underlyings, this amounts to 500 values in comparison to just the 3 required for the methods mentioned above. Often, these 3 parameter are estimated based on historical data (if the data is not delivered by an external vendor). In this case, the Variance-Covariance and the Monte Carlo methods require the storage and maintenance of historical data as well. From the historical time series, the value changes $\delta S_j(\delta t_i)$ of all risk factors S_j over time intervals δt_i with the same length as the liquidation period are determined over the entire available history of the risk factors[2]:

$$\delta S_j(\delta t_i) = S_j(t - i\,\delta t + \delta t) - S_j(t - i\,\delta t) \quad \text{with} \quad i = 1, \ldots, m \; ; \; j = 1, \ldots, n \,. \tag{23.4}$$

For example, the time series over 250 days yields 249 daily changes or 240 changes for a liquidation period of 10 days.[3]

The historical risk factor changes applied to the price of the risk factors at time t (today) provide m different scenarios. For reasons of consistency,

---

[2]The number of available liquidation periods obtained from the historical time span for which data is available is denoted by m, the number of relevant risk factors again by n.

[3]In the second case, the liquidation periods overlap resulting in auto-correlations.

the *relative* changes are often be applied to today's price rather than the absolute changes.[4] For each scenario i the thus induced value change δV_i of the portfolio V is computed. This can be accomplished with a full valuation of the portfolio. However, in practice it is often the case that a simple linear (delta valuation) or quadratic (delta-gamma valuation) approximation as in Eq. 22.4 is performed.

In this way, m "historical" value changes $\delta V (t_i)$ are generated from the past time series data. The value at risk of a historical simulation is now the minimum of all δV_i, where—similar to the Monte Carlo simulation—unfavorable changes in the portfolio's value falling outside a previously specified confidence interval are ignored. For a confidence level of 95%, for example, and a history consisting of 250 days, the 12 worst out of the 249 portfolio value changes are not considered when finding the minimum over daily changes.

If δV_{1-c} denotes the most favorable change among the ignored value changes at a confidence level of c, then the value at risk of the portfolio is the least favorable portfolio change greater than V_{1-c}:

$$\text{VaR}_V (c) = - \min_i \{\delta V_i \mid \delta V_i > \delta V_{1-c}\} \quad \text{with} \quad i = 1, \ldots, m .$$

With a confidence c, the portfolio at the end of the liquidation period depreciates by an amount no larger than this value at risk.

At this point we can clearly see the greatest disadvantage of this method: the weak statistical information on the basis of which the probabilistic conclusions such as confidence levels are drawn. Despite the effort in data management of all relevant historical time series, usually only a dozen (in the above example) or so values remain for the final analysis, namely those falling below the lower boundary of the confidence interval. The probabilistic conclusion is drawn on the basis of these few values. In contrast, the statistical basis deriving from 10,000 Monte Carlo simulation runs is approximately 50 times larger (of course, this statistical advantage of the Monte Carlo method is at the expense of assuming that the risk factors are random walks). In addition, the results could be biased because of autocorrelation effects due to the overlapping time intervals.

A further disadvantage of historical simulations is the following effect: For each change of position in a portfolio (after each transaction), the new

---

[4]A historical change for example, a 12 point change in the DAX index which stood at 1200 at the outset of a liquidation period is quite different from a 12 point change when the DAX is at 7000. A relative change of 1% (i.e., 70 points) is therefore more suitable. This is not necessarily the best choice for all possible risk factors, though. For interest rates, absolute shifts are also applied frequently.

portfolio (and its value changes) must be recalculated for all 250 days. In doing so, it may happen that *another* historical risk factor change affects the make-up of the set falling outside the confidence interval for the new portfolio so that suddenly the value at risk based on *another scenario* is relevant. Thus, a trader, after entering into a transaction intended to optimize the VaR according to the *original* scenario, is then informed of a value at risk computed on the basis of *another* scenario. This greatly increases the difficulty of evaluating the success of the transaction.

In general, the historical simulation is carried out by simulating the same risk factors, which are required for the risk-neutral valuation of the portfolio, i.e. volatilities implied from quoted option prices (if the volatilities are not quoted directly anyway) instead of historical volatilities based on historical time series of, e.g. quoted share prices. Therefore, the value at risk is based on the portfolio's simulated *risk-neutral* present values, based on *real-world* historical changes of the underlying risk factors.

23.3 Crash and Stress Testing: Worst Case Scenarios

Each value at risk concept introduced up to this point yields the *potential loss* in the course of a liquidation period and the *probability* with which no bigger loss occurs. The confidence levels most commonly used are 95% or 99%. This means that losses amounting to the value at risk or higher actually occur between 2 and 12 times per year, as they actually should. Otherwise the model on the basis of which these probabilities are derived is incorrect. The value at risk can thus not be considered a worst case scenario, but rather as part of daily business: losses of this magnitude must occur on average once a month at a confidence level of 95%! Accordingly, the value of these losses must be kept below an acceptably small limit.

In order to obtain a measure of a portfolio's risk should a catastrophe occur, a *worst case* or *crash scenario* is constructed by hand through the explicit specification of all risk factors influencing the portfolio. The portfolio is then revaluated on the basis of this market scenario. Such a scenario could, for example, be the financial crisis 2008/2009. The difference between the calculated value and the current portfolio value is the "value at risk" of the portfolio with respect to the crash-scenario. Obviously, this value expresses the potential loss as a result of the crash but no information is available about the *probability* of such an event.

A further method used to get a feeling for the risk of a portfolio in the case of rare and very unfavorable market developments is to use 6 or 8 standard deviations from the expectation as the boundary of the confidence interval rather than the usual 1.65 or 2.36, i.e. to consider selected, extreme scenarios. This approach is sometimes referred to as a *stress test.* In this way, a potential loss is obtained as well as a *theoretical* probability that a loss of this magnitude is incurred. For example, for a standard normal distribution, the probability of a loss of more than six standard deviations in a one-sided confidence interval is approximately one to one billion:

$$1 - \frac{1}{\sqrt{2\pi}} \int_{-6}^{\infty} e^{-x^2/2} dx \approx 9,866 10^{-10} \approx 10^{-9} \ .$$

No great importance should be attached to such probability statements since the random walk assumptions, constant volatilities and correlations, for instance, are in all probability no longer satisfied when such events occur. As a rule, market scenarios of this type change the correlations drastically and the volatilities explode. Thus, de facto, stress tests, like crash tests, provide information on the potential loss involved without specifying the probability of such an event.

23.4 Advantages and Disadvantages of the Commonly Used Value at Risk Methods

In Table 23.1 the advantages and disadvantages of the VaR methods introduced above are summarized. A "+" in the method column indicates that this particular method has the advantage of the property associated with the corresponding row. A "−" means that it has the disadvantage of the corresponding row. No entry indicates that the method does not have the property of the row under consideration. A symbol in parentheses means that the indicated property is usually assumed in the application of the method but that the advantage or disadvantage is not, in *principle*, characteristic of the method.

Table 23.1 The pros and cons of the most common value at risk methods

	Variance-covariance	Monte Carlo	Historical simulation
Models risk factors as random walks	−	−	
Assumes constant vol & correlation	−	(−)	
Requires historical time series			−
Requires vol & correlations	(−)	(−)	
Neglects the mean yield	−		
Linear proxy for risk factors	−		(−)
Linear proxy for prices (delta valuation)	−	−	−
Full valuation		+	+
Specified scenarios		+	+
Specified probabilities	+	+	
Based on large data sets		+	
Valid for long liquidation periods		+	
Takes vega risk into account		(+)	+
Takes theta risk into account		+	+

24

Example of a VaR Computation

In the Excel workbook VALUEATRISK.XLSX available in the download section [50], several of the value at risk concepts already introduced are applied to explicitly compute the VaR of a concrete portfolio within the delta-normal method. The example is quite dense in the sense that many of the concepts introduced above (as well as several concepts to be presented in later chapters, in particular in Sects. 30.3.3 and 30.3.4) are collected in one calculation. However, it is by all means reasonable to present such a summary at this point as it will provide the reader with a complete reference containing all the essential steps for computing a value at risk (at least, the "simple" delta-normal version). We will proceed step by step through the example.

24.1 The Portfolio

The portfolio is presented in Table 24.1. It is composed of:

- A British zero bond (denoted by GBP.R180) with a lifetime of 6 months and a face value of 100,000 British pounds.
- A Japanese zero bond (denoted by JPY.Z06) with a lifetime of 6 years and a face value of 2,000,000 Japanese yen.
- Put options on Japanese 7-year zero bonds (denoted by JPY.Z07) with a face value of 6,000,000 Japanese yen. These puts have a lifetime of 6 months and a strike of 94% (the price of the bonds are also given in percentage

Table 24.1 The portfolio

Financial instrument	Currency	Market prices	Cross rate to EUR	Principal in currency	Market value in EUR
GBP.R180	GBP	97.700%	€1.61720	£10,000	€15,800.68
JPY.Z06	JPY	96.010%	€0.00926	¥2,000,000	€17,777.28
Put on JPY.Z07	JPY	0.389%	€0.00926	¥6,000,000	€216.30
Delta of put		−0.2803		Sum	€33,794.27

points; a value of 100% means that the present value of the bond is equal to its face value). The Black-Scholes price (at a risk-free rate of 3%) of these puts is 0.389% (of the face value), i.e., 23,364 JPY. The delta is −0.2803.

24.2 Market Data

Prices, daily volatilities (both the price as well as the *yield* volatilities; see Sect. 30.3.3) and correlations are available for GPG.R180, JPY.Z07 and JPY.Z05 and are presented in Table 24.2. These three variables are the portfolio's risk factors.

Data for a 6-year zero bond in Japanese yen (JPY.Z06) is not available initially. Its yield and its price volatility were computed from the vertex data for 5 and 7 years linear interpolated. The price of this zero bond was then calculated from its yield in accordance with Table 2.5 for discrete annual compounding and the yield volatility was determined using Eq. 30.18 together with the price volatility. This data is written in italics in Table 24.2 to

Table 24.2 Market data of the portfolio's risk factors given in their original currency

Risk factor	Currency	Price		Yield	
		Market price	Daily price vol	Annual yield	Daily yield vol
USD/EUR	EUR	$0.9083	0.680%		
USD/GBP	USD	$1.4689	0.404%		
USD/JPY	USD	$0.0084	0.731%		
GBP.R180	GBP	97.70%	0.026%	4.70%	1.10%
JPY.Z05	JPY	97.74%	0.093%	0.46%	4.01%
JPY.Z07	JPY	93.89%	0.177%	0.90%	2.99%
JPY.Z06	JPY	*96.01%*	*0.135%*	*0.68%*	*3,33%*

Table 24.3 Volatilities and correlations of the risk factors

	In original currency			In EUR		
	JPY.Z05	JPY.Z07	GBP.R180	JPY.Z05	JPY.Z07	GBP.R180
JPY.Z05	0.093%	0.89410	0.17171	0.705%	0.99006	0.48739
JPY.Z07	0.89410	0.177%	0.00987	0.99006	0.725%	0.49386
GBP.R180	0.17171	0.00987	0.026%	0.48739	0.49386	0.450%

emphasize the fact that the data for this 6-year bond are not original market values, but have been obtained through interpolation and other calculations.[1]

Naturally, the FX rate of the British pound and the Japanese yen with respect to the reference currency in which the value at risk is to be expressed, are also risk factors for the portfolio (these are also listed in Table 24.2). The *reference currency* here is the euro (EUR). As will be shown extensively in Sect. 30.3.4, the *FX risk* is already taken into account by the fact that the given volatilities and correlations with respect to their respective original currencies (here the GBP and JPY) must be transformed in terms of the reference currency. These transformations are determined from the volatilities and correlations of the FX rates concerned. In this way, the FX risk is incorporated into the calculation. If these FX rates are not available in the reference currency, but in yet *another* currency (in our example, US dollar), this new currency (with *its* volatilities and correlations) also comes into play. Therefore, the USD/EUR exchange rate appears in Table 24.2 as well, though there is no explicit risk against USD present here. The transformations of the volatilities and correlations into the reference currency are quite complicated and will receive detailed treatment in Sect. 30.3.4. The result of all these transformations is presented in Table 24.3.

The volatilities (in the diagonal) and the correlations of the three portfolio risk factors with respect to the original currency (as provided by the data provider, for instance) are listed on the left side in Table 24.3. On the right side, the transformed parameters (in accordance with Sect. 30.3.4) are displayed in the form needed for the value at risk computation with respect to the reference currency.

---

[1]The lack of liquidly traded bond options that could be used to imply price volatilities is one reason for using the nodes of the interest rate curve used for discounting the bond cash flows as risk factors.

Table 24.4 Three ways for calculating the value at risk

	10 days	99% confidence	
Risk factor	VaR of each position in EUR	Portfolio delta in EUR w.r.t. each risk factor	Portfolio VaR in EUR w.r.t. each risk factor
GBP.R180	€523.45	€16,171.97	€523.45
JPY.Z05	€426.28	€8,404.73	€426.28
JPY.Z07	€509.86	€−5,383.26	€269.51
Put	€779.37	Sum	€1,219.25
Sum	€2,238.96	Total VaR	€615.66

24.3 Calculation of Risk

The risk of this portfolio in EUR is now calculated with the transformed volatilities and correlations found in Table 24.3. In order to separately illustrate the *hedging effects* (the position in the 7-year bond partially hedges the put on the 7-year bond), the *correlation effects* and the *diversification effects*, we calculate three different risk values in Table 24.4.

Firstly, the value at risk is calculated for each separate *position* with Eq. 22.13. If we were to be extremely conservative and not consider either hedging nor correlation nor diversification effects, we could simply add the four values at risk of the separate positions and arrive at a total (very conservative) risk number.

Alternatively, we could calculate the value at risk taking into account the hedge effects. The value at risk of the entire portfolio with respect to each separate *risk factor* is determined, likewise with Eq. 22.13 as above. To do so, we first need to calculate the portfolio deltas with respect to the three risk factors. For the bond positions, this is equal to the face value (because the discount factors are treated directly as risk factors) and for the bond option, it is the option delta multiplied by the face value of the option. If different positions depend on the same risk factor, their deltas with respect to that risk factor can simply be added to yield the total delta with respect to that risk factor. The hedge effect of the put is in this way taken into consideration. The delta of the put on the 7-year bond is negative and substantially reduces the portfolio sensitivity with respect to the 7-year bond.[2] If we wish to include this *hedge effect* of the put in the risk computation while still neglecting the correlation

---

[2]In fact it more than compensates the positive delta of the 7-year bond position since the resulting portfolio delta with respect to the risk factor JPY.Z07 is negative. This is called *over hedging*. However, this is not as bad as the name suggests. Since JPY.Z07 is strongly correlated to JPY.Z05, this negative portfolio delta significantly reduces the *total* risk of the portfolio, when correlation effects are taken into account.

(respectively, the diversification) effects, we would now add the portfolio VaRs with respect to each risk factor to arrive at a total (and still conservative) risk number. The hedge effect of the put position is considerable. The put reduces the value at risk of the portfolio to almost half of the very conservative figure calculated when neglecting the hedge effect.

Finally, the value at risk is computed taking both hedge and correlation effects into account, in accordance with Eqs. 22.15 using the correlations with respect to the reference currency EUR listed in Table 24.3. The correlation (respectively diversification) effects reduce the value at risk again by roughly one half.[3] Assuming that all model assumptions and approximations made are justified, we can now be 99% confident that the portfolio will lose no more than this VaR over the next ten days.

---

[3]This of course is largely due to the fact that the put "hedges" not only the JPY.Z07 but also the JPY.Z05-risk (i.e. the whole original 6-year bond). This effect can of course only be seen if correlations are taken into account.

25

Check for updates

Backtesting: Checking the Applied Methods

A comparison of the value at risk figures delivered by a risk management system with the actual value changes of a portfolio allows an estimation of the qualitative and quantitative "goodness" of the risk model. Comparisons of realized values with previously calculated values are called *backtesting* procedures.

25.1 Profit and Loss Computations

There are several different *profit & loss* (or *P&L*) methods which can be used for comparison with the value at risk. The differences in these methods reflect the differences in the fundamental "philosophy" behind them.

- The *Dirty* Profit & Loss: The actual P&L of the portfolio, including all changes in position, fees paid and received, commission, etc. over the value at risk period are compiled and compared with the value at risk previously calculated. Position changes arise from continued trading during the value at risk period, the maturing of positions in the portfolio (for example, futures and options), the knock-in or knock-out of barrier options, coupon payments of bonds, etc. The effect of continued trading is not, in general, contained in the value at risk model. The dirty P&L is therefore only suitable for evaluating trading performance and not for the evaluation of model performance.

- The *Cleaned* Profit & Loss: The cleaned P&L is calculated in the same way as the dirty P&L but *without* taking position changes into account which result from continued trading during the value at risk period. Furthermore, the payment and receipt of fees and commissions are also omitted from the calculation. However, the cleaned P&L still contains the position changes resulting from the maturity of instruments occurring during the value at risk period (such as options and futures) or other position effects caused by the market (as opposed to the trader) such as the knock-out or knock-in of barrier options, coupon payments of bonds, etc. The cleaned P&L is therefore suitable to evaluate the model performance of risk models which take account of such maturity effects. The Monte Carlo simulation, for example, allows for such effects, the Variance-Covariance method, on the other hand, does not.
- The *Clean* Profit & Loss: Finally, the clean P&L is calculated in the same ways as the cleaned P&L but with reversing the effects of the maturing of positions during the value at risk period. In calculating the clean P&L, the value of the exact same portfolio as that existing upon initial calculation of the value at risk is re-calculated with the new market data observed at the conclusion of the value at risk period. Of course, a record of the initial portfolio positions at the time of the value at risk computation must have been kept. The clean P&L is thus suitable for evaluating the model performance of risk models such as the Variance-Covariance method which do not account for *aging* effects of the positions.

Independent of the chosen profit & loss method, the profit & loss per backtesting period is recorded for the evaluation of the goodness of the value at risk. Additionally, a record is kept for the calculated value at risk of the portfolio per backtesting-period. The data required for backtesting is thus not very large: neither historical time series nor a history of the portfolio positions must be maintained. Only two values per portfolio must be recorded in order to save the history of the portfolio, namely the value at risk and the associated P&L of the portfolio to be compared with the value at risk. In most cases, the P&L will be more favorable than the value at risk, and in others less favorable. Counting the number of times that the P&L is less favorable than the value at risk enables statistical conclusions about the goodness of the utilized model to be drawn. This is the fundamental idea behind backtesting. The superversing authorities require banks to perform such a backtesting procedure to validate the value at risk calculation (see Sect. 21.1). In the following, we present a standard method for implementing a backtesting procedure.

25.2 The Traffic Light Approach of the Supervising Authorities

25.2.1 Adjusting the Value at Risk (Yellow Zone)

The value at risk is a statistical statement. In general, some of the changes in the portfolio's values will be less favorable than the calculated value at risk. Such changes are referred to as *outliers* in the following discussion. At a confidence level c the probability of such an outlier is $1 - c$. The *expected* number of outliers in n backtesting periods with respect to this level of confidence is thus

$$E[k] = n(1 - c) , \qquad (25.1)$$

where k denotes the number of outliers observed in the n backtesting periods. The value k is not always equal to its expectation; it is a random variable. A deviation of the observed k from the expected number of outliers does not necessarily imply that a model is incorrect. Such an observation may be the result of pure chance. This is particularly true for small deviations from the expected value. In such cases, one speaks of the *yellow zone*, in which the supervising authorities will accept the model, but require that the value at risk is increased in the following manner.

The realization of k outliers actually observed in backtesting allows the definition of a new confidence level c' with respect to which the observed number is equal to the expectation:

$$k = n(1 - c') \quad \Rightarrow \quad c' = 1 - k/n . \qquad (25.2)$$

From the observed outliers, it can be concluded that the value at risk from the model does not correspond to the confidence level c, but to a confidence level c'; or at least that the experimental basis for a confidence c' is greater than for c. The given VaR is then interpreted with respect to a confidence level of c'. A VaR which better corresponds to the *claimed* confidence c is then obtained from the given value of the VaR through multiplication by the ratio of the percentiles (confidence interval bounds) Q_{1-c} and $Q_{1-c'}$ in accordance with Eq. 21.20, respectively Eq. 22.15:

$$\text{VaR}(c, t, T) \approx \frac{Q_{1-c}}{Q_{1-c'}} \ \text{VaR}(\underbrace{1 - \frac{k}{n}}_{c'}, t, T) . \qquad (25.3)$$

In order to use this equation, all assumptions and approximations of the delta-normal method must be made. These are:

- The risk factors are random walks, i.e., they are lognormally distributed implying that $Q_{1-c} = Q_{1-c}^{N(0,1)}$.
- The drifts of the risk factors are neglected in the calculation.
- The exponential time evolutions of the risk factors are linearly approximated.
- The dependence of the portfolio value on the risk factors is linearly approximated. In particular, the portfolio value is also assumed to be lognormally distributed.

Since, for logarithmic changes, all of the variables under consideration are assumed to be normally distributed, the $Q_{1-c'}$ percentile can be calculated as

$$c' = 1 - \frac{k}{n} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Q_{1-c'}} e^{-x^2/2} dx \ .$$

For example, if backtesting over 250 periods is performed resulting in the observation of 6 logarithmic portfolio value changes outside of the claimed confidence interval at a confidence level of 99%, the percentiles Q_{1-c} and $Q_{1-c'}$ are given by

$$0,99 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Q_{1-c}} e^{-x^2/2} dx \Rightarrow Q_{1-c} \approx 2,326$$

$$1 - \frac{6}{250} = 0,976 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Q_{1-c'}} e^{-x^2/2} dx \Rightarrow Q_{1-c'} \approx 1,972 \ .$$

The confidence level of the value at risk calculated by the model is now assumed to be not 99% as claimed but rather 97.6% as calculated on the basis of the actual events. The value at risk must thus be adjusted by a factor of $Q_{1-c}/Q_{1-c'} = 2.326/1.972 = 1.18$. This now larger value at risk is the value which can, based on actual observations, be relied upon with a confidence of 99%.

25.2.2 Criteria for Rejecting a Model (Red Zone)

Adjusting the value at risk as described above may not be applied for *arbitrary* values of k. It is allowed only if the difference between the observed value of k and its expectation with respect to the claimed confidence c can be reasonably explained by fluctuations due to the randomness involved. If the number of outliers is too far removed from the expected value, chance is no longer a plausible explanation and the reasons for the deviation lie in all probability on fundamental errors in the model itself. Here, the notion is used that the deviation of the measured results from the expected results are *significant*. The supervising authorities say that the model is in the *red zone*.

The field of statistics provides *hypothesis tests* which serve to check whether or not the observed deviation from a claimed value can be plausibly explained by random fluctuations. If not, such a deviation is considered significant and the tested *hypothesis* is in all probability not true. But again, absolute statements cannot be made on the basis of statistics. From such a hypothesis test we can only conclude *with a certain probability* that it was correct to reject (or accept) the hypothesis. The possibility remains that the hypothesis is rejected (accepted) although it is true (false). In statistics, these kind of errors are referred to as *type-I error* (rejection of a correct hypothesis) and *type-II error* (acceptance of a wrong hypothesis).

The hypothesis made when backtesting an internal model is that the observed value changes of a portfolio lie within the confidence interval specified by the calculated value at risk with a probability of c, in other words, that with a probability c the observed value changes are more favorable than the computed value at risk. To check this claim, we test whether the observed portfolio's value changes actually lie within the respective confidence interval. This is done for every VaR period over the entire backtesting time span.

The observed results can be categorized into two possible outcomes: the change in the portfolio's value is either more favorable than the respective VaR or not. This corresponds to the *Bernoulli experiment* described in Sect. A.4.2 in full detail. We could associate the event that an actual portfolio change is *more favorable* (or equally favorable) than the VaR with the outcome "tails" when tossing a coin and likewise the event that the portfolio change is *less favorable* than the VaR to the outcome "heads". The probability of observing k "heads" in n trials (less favorable than the VaR) is binomially distributed (see Eq. A.41) where the binomial probability p is the probability of the outcome "heads". In our case here p is then equal to the claimed probability that the portfolio change lies outside of the confidence interval, i.e., $p = 1 - c$. The number of

outliers should therefore be binomially distributed with a density

$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{with} \quad p = 1 - c = 1 - \mathrm{N}(Q_{1-c}) ,$$

(25.4)

where $\mathrm{N}(Q_{1-c})$ denotes the probability that a standard normally distributed variable is $\leq 1 - c$, see Eq. A.53. Assuming that the model is correct, $B_{n,p}(k)$ is the probability that precisely k outliers are observed. The probability that *at most k* outliers are observed is then

$$\sum_{i=0}^{k} B_{n,p}(i) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} .$$

(25.5)

Again, assuming that the model is correct, the probability of observing *more than k* outliers is then

$$1 - \sum_{i=0}^{k} B_{n,p}(i) = \sum_{i=k+1}^{n} \binom{n}{i} p^i (1-p)^{n-i} .$$

(25.6)

This is equal to the probability that the model is *correct* assuming that k or more outliers are observed. Therefore, this is the probability of making a type-I error (the rejection of a correct model) when the hypothesis is rejected if k or more outliers are observed.

The determination of a Type-II error (acceptance of a *false* hypothesis) requires that the *true* probabilities for outcomes of the *false* model have to be known—a luxurious situation which almost never happens in practice. To be more specific: If a hypothesis test accepts a model for up to k outliers, then the probability that a *false* model is accepted equals the *true* probability for the event that this *false* model produces k or fewer outliers. Let's consider a simple example: Assume that a false model claims a is 99% confidence for its calculated VaR-numbers while the *true* confidence for these VaR-number is only 95%. If one accepts that model's VaR-numbers as a 99% VaR as long as only up to 9 out of 250 backtesting periods produce an outlier, one makes a Type-II error (see Eq. 25.5 with $n = 250$, $p = 5\%$) with a probability of

$$\sum_{i=0}^{9} B_{n,p}(i) = \sum_{i=0}^{9} \binom{250}{i} 0,05^i * (0,95)^{250-i} \approx 19,46\% .$$

It should be clear from these considerations that type-II errors only play a minor role in practice since it is very rarely the case that they can be determined in a sensible way.

The supervising authorities make their decision on establishing the limits for the red zone based on the probability of a type-I error (rejection of a correct model). The model is said to be in the red zone for a number of outliers k if the rejection of the model for *more than k* outliers has a probability for a type-I error of less than 0.01%. Using Eqs. 25.6 and 25.5, we find this to be the case when the probability (calculated with the model under consideration) of *at most k* outliers is $\geq$ 99.99% under the assumption that the model under consideration is correct. For $n = 250$ backtesting periods, the probability of at most 9 outliers equals 99.975% (see the Excel Workbook BinomialBacktest.xls from the download section [50]). The probability of at most 10 outliers is equal to 99.995% and is thus larger than 99.99%. The red zone established by the supervising authorities for 250 backtesting periods therefore begins at 10 outliers although the probability of making a type-I error (which is the probability for *more than 9* outliers) is 0.025%, i.e., greater than 0.01%. The probability of a type-I error when deciding to reject the model if more than $k = 10$ (in other words, 11 or more) outliers are observed is 0.005%, smaller than the required 0.01%. Nonetheless, 10 outliers within 250 backtesting periods is already deemed to belong to the red zone.[1]

25.2.3   The Green Zone

The boundary of the red zone is the upper boundary of the yellow zone. Analogously, a lower boundary of the yellow zone has been defined. No add-on is required if the observed number of outliers lies below this boundary. This zone is called the *green zone*. The model is said to be in the yellow zone for a number of outliers k if the rejection of the model for *more than k* outliers has a probability for a type-I error of less than 5.00%. Using Eqs. 25.6 and 25.5, we find this to be the case when the probability of *at most k* outliers is $\geq$ 95% under the assumption that the model is correct. For $n = 250$ backtesting periods for example, the probability of at most 4 outliers is equal to 89.22% (see the Excel workbook BinomialBacktest.xls from the download section [50]). The probability of at most 5 outliers on the other hand is 95.88%. Thus, the supervising authorities establish the boundary for the

---

[1]It may seem inconsistent that the supervising authorities establish k as the boundary for the red zone although this k corresponds to a type-I error of greater than 0.01%. This is the rule, however.

yellow zone as $k = 5$ for 250 backtesting periods, although the probability of a type-I error in this case is 10.78%, i.e., greater than 5%. For the rejection of the model with *more* than 5 outliers, the probability of a type-I error is 4.12%.

These three zones established by the supervising authorities, motivate the name *traffic light approach.*

25.2.4   Multiplication Factor and Add-On

As a rule, the value at risk calculated with the model must be multiplied by a factor of three even when it is found to be in the green zone for the simple reason that it has not been computed in accordance with the standard methods but by means of an internal model (this is just a rule of thumb, though). In the yellow zone, the VaR must *additionally* be multiplied by the ratio of the two percentiles Q_{1-c} and $Q_{1-c'}$ as prescribed in Eq. 25.3 where $Q_{1-c'}$ is the value established from Eq. 25.2. The multiplication factor for the yellow zone is thus $3 Q_{1-c}/Q_{1-c'}$. The amount by which the multiplication factor exceeds the factor 3 is referred to as the *add-on*

$$\text{Add-on} = 3\frac{Q_{1-c}}{Q_{1-c'}} - 3 \approx \frac{6,978}{Q_{1-c'}} - 3 \ ,$$

where in the last step the confidence level $c = 99\%$ (required in SolvV (Germany) resp. CRR) for the standard normal distribution, $N(Q_{1-c}) = 1\%$, and consequently $Q_{1-c} \approx 2.326$ was used.

The concepts described above are illustrated in detail in the Excel workbook BINOMIALBACKTEST.XLS. All of the probabilities mentioned above, the boundaries between the different zones and the add-ons in the yellow zone are computed. The number of backtesting periods as well as the required confidences and the probability thresholds for the boundaries between the zones can be modified and the subsequent effects of these modifications immediately computed. In Fig. 25.1, these values are presented for $n = 250$ backtesting periods and a confidence of $c = 99\%$.

In Table 25.1, the add-ons are again explicitly displayed for the situation in Fig. 25.1 (rounded in increments of 0.05). This table can be found in the SolvV resp. the (CRR). For the red zone, an add-on of one is set and the model is later subjected to a new test.

Despite the multiplication factor, the value at risk computed with internal models is often lower than that calculated in accordance with the standard
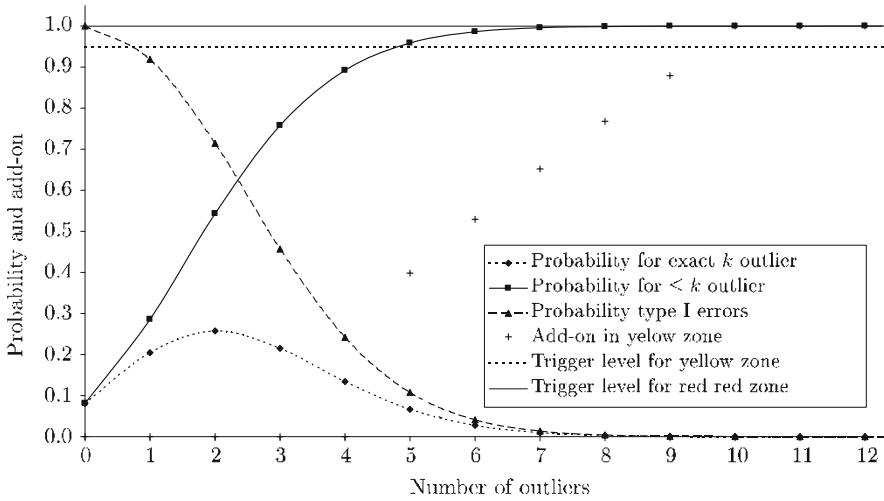
Fig. 25.1 Value at risk backtesting by means of a binomial test for 250 backtesting periods and 99% VaR confidence

Table 25.1 The table of add-ons for 250 backtesting periods as shown in the text of the German law

k	Add-on	Zone
4	0,00	Green
5	0.40	Yellow
6	0.50	Yellow
7	0.65	Yellow
8	0.75	Yellow
9	0.85	Yellow
≥ 10	1.00	Red

methods, since the correlation and compensation effects are more accurately taken into account. In many cases, the VaR of a portfolio computed according to an internal model is, despite the multiplication factor, is significantly lower than that computed with the standard methods.

Part V

Portfolios

26

Check for updates

Classical Portfolio Management

Portfolio management is about maximizing the return of an asset portfolio by minimizing (or keeping control of) risk at the same time. The return considered here is the *expected* return of the portfolio over the *next* holding period. As discussed in great detail in Part IV of this book, *risk* is also an estimator covering the *next* holding (or "liquidation") period, i.e., a period lying in the *future*. Such estimators covering future time spans are sometimes called *ex ante* estimators.

These have to be distinguished from *ex post* estimators, which show how the quantities under consideration (e.g., return, risk, etc.) behaved in the past. The historical properties of a portfolio arise not only as results of market movements but also of trading, i.e., position changes made by the portfolio manager. Therefore, *ex post* estimations are useful for determining past performance (and bonus payments) of portfolio managers. But what one is really interested in are the properties (in particular risk and return) of the *current* portfolio with its *current* holdings over the *next* holding period. Therefore one needs *ex ante* estimates for the current portfolio. For instance, in Chap. 22 we took the *historical* information of the *risk factors* and the *current sensitivities* of the portfolio to produce an *ex ante* estimate of the portfolio risk. We will also show in Chap. 33 below ex ante estimation of risk factor volatilities. Such ex ante estimations for the risk factors together with the *current* portfolio *sensitivities* also yield ex ante estimates of portfolio risk.

Similarly, one needs ex ante estimates of the expected portfolio *return*. Methods for ex ante return estimations are presented for instance in Chap. 31. Other frequently used method for estimating expected returns include fundamental analysis or, at the other end of the spectrum, chart techniques, to name only a few possibilities. In the following, we shall denote *expected* returns by R to distinguish them from the (realized) returns, r, appearing for instance in Eq. 26.1, below.

26.1 From Risk Management to Portfolio Management

26.1.1 Assets and Risk Factors

Consider a portfolio consisting of holdings $N_k, k = 1, \ldots M$ in M financial instruments (also called *assets* in the world of *asset management*) with values V_k dependent on n risk factors $S_i, i = 1, \ldots, n$. The value of this portfolio is as in Eq. 22.6

$$V(t) = \sum_{k=1}^{M} N_k V_k(\mathbf{S}(t)) \, ,$$

with $N_k V_k$ being the value of each single position. We need to get estimates for the risk and expected return of this portfolio with its current holdings N_k (i.e., without trading). The change in the portfolio value is

$$\delta V(\mathbf{S}(t)) = \sum_{k=1}^{M} N_k \delta V_k(\mathbf{S}(t)) \, .$$

The return of the portfolio over a time period δt, defined as the relative price change over that period, is

$$r_V(\mathbf{S}(t))\delta t \equiv \frac{\delta V(\mathbf{S}(t))}{V(\mathbf{S}(t))} = \sum_{k=1}^{M} \frac{N_k V_k(\mathbf{S}(t))}{V(\mathbf{S}(t))} \frac{\delta V_k(\mathbf{S}(t))}{V_k(\mathbf{S}(t))} = \sum_{k=1}^{M} w_k(\mathbf{S}(t)) r_k(\mathbf{S}(t))\delta t \, .$$

$$(26.1)$$

As stressed for instance in connection with Eq. 2.31, these returns, defined as relative price changes, are returns in the *linear* compounding method. Here, we

have introduced the portfolio return r_V, the asset returns r_k, and the position weights w_k:

$$r_V \delta t := \frac{\delta V}{V} \ , \quad r_k \delta t := \frac{\delta V_k}{V_k} \ , \quad w_k := \frac{N_k V_k}{V} \quad \text{with} \quad k = 1, \dots M \ .$$
(26.2)

The weight w_k ist that portion of the total portfolio value, which is invested in asset V_k.[1] Using vector notation, Eq. 26.1 can be written very compactly as[2]

$$r_V = \mathbf{w}^T \mathbf{r} \ .$$

The same holds for the *expected* portfolio return R_V

$$R_V = \mathrm{E}\,[r] = \mathbf{w}^T \mathbf{R} \ ,$$
(26.3)

where $\mathbf{R}$ denotes the vector of *expected* instrument (or asset) returns R_k for $k = 1, \dots, M$.

$$\mathbf{R} = \mathrm{E}\,[\mathbf{r}] \iff R_k = \mathrm{E}\,[r_k] \quad \text{with} \quad k = 1, \dots, M \ .$$

As explained in the context of Eq. 2.32, expected asset returns in linear compounding are equal to the assets' drifts. For instance, if an asset were *per se* a risk factor S modeled by Eq. 2.24 (or equivalently by Eq. 2.17), its expected return would equal its drift $\tilde{\mu} = \mu + \sigma^2/2$.

We will begin by considering *fully invested* portfolios, meaning, that all of the capital is invested in the M financial instruments (or "assets"). No money is left over and neither is money borrowed to invest more than the capital at hand (no *leveraged investments*). For fully invested portfolios, the position

---

[1]From now on, we will suppress the argument $\mathbf{S}(t)$ to simplify the notation. It is understood that all quantities involving instrument values (and their derivatives) are all functions of $\mathbf{S}(t)$ and all quantities are also functions of t.

[2]In the following, we make frequent use of vector notation, i.e.,

$$\mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_M \end{pmatrix}, \ \mathbf{R} = \begin{pmatrix} R_1 \\ \vdots \\ R_M \end{pmatrix}, \ \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \text{etc.}$$

weights fulfill the condition

$$\sum_{k=1}^{M} w_k\left(\mathbf{S}(t)\right) = 1 \tag{26.4}$$

$$\mathbf{w}^T \mathbf{1} = \mathbf{1}^T \mathbf{w} = 1$$

Note that, throughout this book, we distinguish between the *financial instruments* (or *assets*) V_k and the *risk factors* (or *market parameters*) S_i those instruments (and thus the portfolio value) depend on. Though this is not very common in the standard literature on *modern portfolio theory*, this notation has the advantage not to rely on the (unrealistic) assumption that a portfolio contains direct positions of all its risk factors.

As we will see shortly, the covariances of the *instruments* in the portfolio are needed for modern portfolio theory. Usually, however, only the covariances of the *risk factors* are available. Thus, the major problem (and the reason why usually only risk factors instead of instruments are considered) is that the covariance matrix refers to the *risk factors* S_i while the portfolio positions (and their weights) refer to the *financial instruments* V_k. There are several ways to overcome this difficulty. If, for instance the historical time series (and not only moments like covariances and mean returns) of the risk factors are available, then historical time series of the *instruments* can be constructed by re-pricing the instruments according to all the historical risk factor data. From the historical time series of instrument prices thus constructed covariances between these instrument prices can be determined. Then, one proceeds as if each instrument was a risk factor of its own using the covariances between the instruments.

In the following we will take an alternative route which does not require historical time series and which fits directly to the considerations presented in Sect. 22.2. We will work within the delta normal Value at Risk framework using Approximation 22.9 for the portfolio value change, i.e., we will use Eq. 22.7 without the 2nd order term:

$$\delta V \approx \sum_{k=1}^{M} N_k \sum_{i}^{n} \frac{\partial V_k}{\partial S_i} \delta S_i = \sum_{k=1}^{M} N_k \sum_{i}^{n} \Delta_i^k \delta S_i \; .$$

In this approximation the portfolio return over a time period δt can be expressed as a function of the *risk factor* returns[3] $r_i \delta t := \delta S_i / S_i$ as opposed to the *instrument* returns in the general Eq. 26.1:

$$
\begin{aligned}
r_V \, \delta t &= \frac{\delta V}{V} \\
&\approx \sum_{k=1}^{M} N_k \frac{V_k}{V} \sum_{i}^{n} \frac{1}{V_k} \Delta_i^k \delta S_i \\
&= \sum_{k=1}^{M} w_k \sum_{i}^{n} \frac{S_i}{V_k} \Delta_i^k \frac{\delta S_i}{S_i} = \sum_{k=1}^{M} w_k \sum_{i}^{n} \Omega_i^k r(S_i) \delta t \ .
\end{aligned}
$$

Here, the instrument Omegas Ω_i^k as defined in Table 12.2 of Sect. 12.4 appear

$$
\Omega_i^k := \frac{S_i}{V_k} \Delta_i^k = \frac{S_i}{V_k} \frac{\partial V_k}{\partial S_i} \quad \text{with} \quad k = 1, \ldots, M \ . \tag{26.5}
$$

Comparing this with Eq. 26.2 we find that the asset returns r_k are approximately given by the risk factor $r(S_i)$ returns via

$$
r_k \approx \sum_{i}^{n} \Omega_i^k r(S_i) \quad \text{with} \quad r(S_i) \delta t \equiv \delta \ln S_i \approx \frac{\delta S_i}{S_i} \ . \tag{26.6}
$$

26.1.2   Portfolio Risk and Volatility

The goal of modern portfolio theory is to maximize the expected portfolio return (by finding optimal position weights w_k) while keeping the portfolio risk under control. A measure for the risk is of course the *Value at Risk* which to linear order is given by Eq. 22.15. As was emphasized in Sect. 22.2.2, in the delta normal approximation the Value at Risk is simply proportional to the square root of the portfolio variance. Using Eq. 22.8 for $\widetilde{\Delta}_i$, we can write the

---

[3]Recall that all returns are in linear compounding (c.f. comment after Eq. 26.1).

portfolio variance as in Eq. 22.14

$$
\begin{aligned}
\mathrm{Var}\,[\delta V] &= \sum_{i,j=1}^{n} \widetilde{\Delta}_i\, \delta\Sigma_{ij}\, \widetilde{\Delta}_j \\
&= \sum_{i,j=1}^{n} \sum_{k=1}^{M} N_k \Delta_i^k S_i\, \delta\Sigma_{ij} \sum_{l=1}^{M} N_l \Delta_j^l S_j \\
&= \sum_{i,j=1}^{n} \sum_{k=1}^{M} N_k V_k \Delta_i^k \frac{S_i}{V_k}\, \delta\Sigma_{ij} \sum_{l=1}^{M} N_l V_l \Delta_j^l \frac{S_j}{V_l} \\
&= \sum_{k,l=1}^{M} \underbrace{N_k V_k}_{=:\widetilde{N}_k} \sum_{i,j=1}^{n} \Omega_i^k\, \delta\Sigma_{ij}\, \Omega_j^l \underbrace{N_l V_l}_{=:\widetilde{N}_l} \\
&= \sum_{k,l=1}^{M} \widetilde{N}_k\, \delta C_{kl}\, \widetilde{N}_l \;,
\end{aligned}
\tag{26.7}
$$

where in the last line we have defined an $M \times M$ matrix $\delta\mathbf{C}$ with matrix elements[4]

$$
\delta C_{kl} := \sum_{i,j=1}^{n} \Omega_i^k\, \delta\Sigma_{ij}\, \Omega_j^l \approx \mathrm{cov}\,[\delta \ln V_k,\, \delta \ln V_l]\;,\quad k,l = 1, \ldots M\;.
\tag{26.8}
$$

---

[4]The last step follows directly from Eq. 26.6 which in terms of logarithmic (or relative) changes reads

$$
\delta \ln V_k = r_k \delta t \approx \sum_{i}^{n} \Omega_i^k r(S_i)\delta t = \sum_{i}^{n} \Omega_i^k \delta \ln S_i\;,
$$

thus

$$
\begin{aligned}
\mathrm{cov}\,[\delta \ln V_k,\, \delta \ln V_l] &= \mathrm{cov}\left[\sum_i^n \Omega_i^k \delta \ln S_i,\, \sum_j^n \Omega_j^l \delta \ln S_j\right] \\
&= \sum_{i,j}^{n} \Omega_i^k \Omega_j^l \underbrace{\mathrm{cov}\left[\delta \ln S_i,\, \delta \ln S_j\right]}_{\delta\Sigma_{ij}} = \delta C_{kl}\;.
\end{aligned}
$$

In our first-order approximation these matrix elements are just the covariances of the relative *instrument* price changes in much the same way as the $\delta\Sigma_{ij}$ are the covariances of the relative *risk factor* price changes. Likewise, the holdings N_k are the portfolio sensitivities with respect to the *instrument* V_k. And the $\widetilde{N}_k$ are the "instrument sensitivities times current values" in the same sense as $\widetilde{\Delta}_i$ are the "*risk factor* sensitivities times the current risk factor value". The last line in Eq. 26.7 therefore has exactly the same structure as the first line, with the instruments V_k assuming the role of the risk factors S_i. We now replace the holdings with the corresponding weights as in Eq. 26.2 to arrive at

$$\text{Var}\,[\delta V] = V^2 \sum_{k,l=1}^{M} w_k \delta C_{kl} w_l = V^2 \mathbf{w}^T \delta\mathbf{C}\mathbf{w}\,. \tag{26.9}$$

This is where the literature on portfolio theory usually starts. As mentioned earlier, most authors only consider the case of investments in the risk factors themselves (e.g., in stocks) and not in *instruments* (e.g., derivatives) whose prices are *derived* from the risk factor prices. The only difference between this standard case and the delta normal approximation of the more general case presented here is that instead of $\delta\Sigma$ one has to use the matrix $\delta\mathbf{C}$ as the covariance matrix. Apart from that, all equations look exactly the same. Of course, for the special case of the portfolio containing only positions in the risk factors themselves, we have $V_k = S_k$ and[5] $\Delta_i^k = \delta_{ki}$ and therefore $\Omega_i^k = \delta_{ki}$. Thus, in this special case the matrix $\delta\mathbf{C}$ is identical to $\delta\Sigma$.

With Eq. 26.9, the Value at Risk as defined in Eq. 22.15 can now be expressed in terms of the position weights and the matrix $\mathbf{C}$ as

$$\begin{aligned}
\text{VaR}_V(c) &\approx |Q_{1-c}|\sqrt{\text{var}\,[\delta V]} \tag{26.10}\\
&= |Q_{1-c}|\,V\sqrt{\mathbf{w}^T \delta\mathbf{C}\mathbf{w}}\\
&= |Q_{1-c}|\,V\sqrt{\sum_{k,l=1}^{M} w_k \delta C_{kl} w_l}\\
&= |Q_{1-c}|\,V\sqrt{\delta t}\sqrt{\sum_{k,l=1}^{M} w_k w_l \sum_{i,j=1}^{n} \Omega_i^k \sigma_i \sigma_j \rho_{ij} \Omega_j^l}\,.
\end{aligned}$$

---

[5] Here δ_{ki} again denotes the Kronecker delta.

The last line is written down in a way to be as similar to Eq. 22.15 as possible. It explicitly shows that the Value at Risk is proportional to the portfolio value, which is intuitively obvious: if we invest double the money we have double the risk if the composition of the portfolio (i.e., the position weights) is the same. Therefore, this VaR is *not* suitable for optimization purposes. For instance, searching for the portfolio with minimum VaR would trivially yield the solution with $V = 0$ (the least risky thing is of course to not invest at all). A risk measure which is much better suited for optimization is the relative Value at Risk $\mathrm{VaR}_V(c)/V$ which is proportional to the standard deviation of the portfolio's *relative* price changes and thus to the portfolio volatility:

$$\frac{\mathrm{VaR}_V(c)}{V} = |Q_{1-c}| \sqrt{\frac{1}{V^2} \mathrm{var}\,[\delta V]} = |Q_{1-c}| \sqrt{\mathrm{var}\left[\frac{\delta V}{V}\right]}$$

$$\approx |Q_{1-c}| \sqrt{\mathrm{var}\,[\delta \ln V]} = |Q_{1-c}| \sqrt{\sigma_V^2 \delta t}\ .$$

In the second line we have used the fact that, for small changes, relative changes are approximately equal to logarithmic changes.[6] We have also used the relation between the variance of logarithmic changes and the volatility, which in fact is nothing other than the definition of volatility, see the first part of Eq. 21.29. Using the second part of Eq. 21.29, one could also write the VaR in terms of the variances of the portfolio *returns*. In summary, the relative VaR is

$$\frac{\mathrm{VaR}_V(c)}{V} = |Q_{1-c}| \delta t \sqrt{\mathrm{var}\,[r_V]} = |Q_{1-c}| \sqrt{\delta t}\sigma_V\ , \qquad (26.11)$$

where σ_V is the annualized portfolio volatility and the r_V are the historic portfolio returns over holding periods of length δt, each return annualized.

The goal of portfolio management is to optimize *risk adjusted performance measures* (abbreviated "*RAPM*"), i.e., ratios of the kind

$$\frac{\text{(expected) portfolio return}}{\text{portfolio risk}}\ . \qquad (26.12)$$

As we will see later on, it is advantageous when doing portfolio optimization to have such ratios dimensionless. Equation 26.11 shows that the relative Value

---

[6]This approximation is derived in Eq. 30.9, for instance.

at Risk $\mathrm{VaR}_V(c)/V$ is dimensionless: the returns r_V have dimension "percent per year" which cancels the dimension "year" of δt. However, the portfolio return—as any return—in Eq. 26.12 also has the dimension "percent per year". We would therefore be left with a RAPM having dimension "1 over year", would we choose the relative Value at Risk as our risk measure in a RAPM like Eq. 26.12. This motivates the definition of a risk measure *per unit of time*. This is achieved by dividing the relative VaR by δt, in much the same way as we just have achieved a risk measure *per monetary unit* by a division with V. We thus arrive at a risk measure suitable for portfolio optimization, i.e. leading to dimensionless RAPMs:

$$\eta_V \equiv \frac{\mathrm{VaR}_V(c)}{V\,\delta t} \approx |Q_{1-c}|\sqrt{\mathrm{var}\,[r_V]} = q\sigma_V \quad \text{with} \quad q := \frac{|Q_{1-c}|}{\sqrt{\delta t}}\,,$$

(26.13)

where we have introduce the abbreviation q to streamline the notation. This risk measure has the advantage of being independent of portfolio size and holding period. To calculate the Value at Risk (corresponding to confidence c) in EUR of a portfolio (with identical weights in identical assets) having any size over any holding period, simply multiply this η_V by the holding period and the portfolio size.

Using the explicit form of the Value at Risk given in Eq. 26.10 our risk measure can also be written as

$$\eta_V \approx q\frac{1}{\sqrt{\delta t}}\sqrt{\mathbf{w}^T \delta\mathbf{C}\mathbf{w}} = q\sqrt{\sum_{k,l=1}^{M} w_k w_l \sum_{i,j=1}^{n} \Omega_i^k \sigma_i \sigma_j \rho_{ij} \Omega_j^l}\,.$$

Comparing this with Eq. 26.13, we read of the portfolio volatility

$$\sigma_V = \frac{1}{\sqrt{\delta t}}\sqrt{\mathbf{w}^T \delta\mathbf{C}\mathbf{w}} =: \sqrt{\mathbf{w}^T \mathbf{C}\mathbf{w}}$$

(26.14)

with the new matrix $\mathbf{C}$ defined as

$$\mathbf{C} := \frac{1}{\delta t}\delta\mathbf{C}$$

(26.15)

$$C_{kl} = \sum_{i,j=1}^{n} \Omega_i^k \sigma_i \rho_{ij} \sigma_j \Omega_j^l\,, \quad k,l = 1,\ldots M\,.$$

Note that $\mathbf{C}$ is symmetric, i.e., $\mathbf{C}^T = \mathbf{C}$. From Eqs. 26.8 and 21.28, the relation between $\mathbf{C}$ and the covariances of the asset prices V_k respectively asset returns r_k is

$$C_{kl} \approx \frac{1}{\delta t} \operatorname{cov}\left[\delta \ln V_k, \delta \ln V_l\right] = \delta t \operatorname{cov}\left[r_k, r_l\right], \quad k, l = 1, \ldots M .$$

(26.16)

Classical *Markowitz theory*, also called *modern portfolio theory* uses directly the portfolio volatility σ_V as a risk measure. Although this leads to RAPMs (like the Sharpe Ratio, see below) which are *not* dimensionless (a fact usually ignored by market participants) it is possible to do that, since q in Eq. 26.13 would only play the role of an *overall* (and therefore irrelevant) constant in classical Markowitz theory. In fact, one even goes a step further and uses— since the square root is a strictly monotonous function—the portfolio *variance*

$$\sigma_V^2 = \mathbf{w}^T \mathbf{C} \mathbf{w}$$

(26.17)

as the risk measure, since with this measure optimization problems are a lot easier to solve.

To get the factors δt right in the following, we again remind the reader that because of Eq. 21.29, the portfolio volatility σ_V is related to the variance of the portfolio return via

$$\sigma_V^2 \equiv \frac{1}{\delta t} \operatorname{var}\left[\delta \ln V\right] = \delta t \operatorname{var}\left[r_V\right] .$$

(26.18)

26.1.3 Risk Contribution and Attribution

The contribution of each position to the total portfolio risk σ_V can be quantified by asking for each position i separately: "How much will the portfolio risk change when the weight w_i of that position is changed?" This marginal impact of position i can be measured by the sensitivity $\partial \sigma_V / \partial w_i$ of the portfolio risk with respect to the position weight w_i:

$$\frac{\partial \sigma_V}{\partial w_i} = \frac{\partial}{\partial w_i} \sqrt{\sum_{k,l=1}^{M} w_k C_{kl} w_l} = \frac{1}{2\sigma_V} \left(\sum_{l=1}^{M} C_{il} w_l + \sum_{k=1}^{M} w_k C_{ki} \right) = \frac{\sum_{l=1}^{M} C_{il} w_l}{\sigma_V} .$$

where the last step follows since C is symmetric. In vector notation this reads

$$\frac{\partial \sigma_V}{\partial \mathbf{w}} = \frac{\mathbf{C}\mathbf{w}}{\sigma_V} = \frac{\mathbf{C}\mathbf{w}}{\sqrt{\mathbf{w}^T \mathbf{C}\mathbf{w}}} \ . \tag{26.19}$$

This sensitivity is called *Marginal Risk Contribution*, i.e., $\partial \sigma_V / \partial w_i$ is the marginal risk contribution of the i-th asset.

Using the marginal risks $\partial \sigma_V / \partial w_i$ we can accomplish a decomposition of the portfolio risk Eq. 26.14 into a single sum

$$\sigma_V = \frac{\mathbf{w}^T \mathbf{C}\mathbf{w}}{\sqrt{\mathbf{w}^T \mathbf{C}\mathbf{w}}} = \mathbf{w}^T \frac{\partial \sigma_V}{\partial \mathbf{w}} = \sum_{i=1}^{M} w_i \frac{\partial \sigma_V}{\partial w_i} \ . \tag{26.20}$$

In this way we can attribute an amount $w_i \partial \sigma_V / \partial w_i$ of risk to each asset i. This interpretation makes sense because the sensitivity $\partial \sigma_V / \partial w_i$ tells us (to linear order) how much the portfolio risk changes if w_i changes by 1. Therefore w_i times this sensitivity tells us (to linear order) how much the portfolio risk changes if the weight of the i-th position is changed from zero to w_i. In other words, $w_i \partial \sigma_V / \partial w_i$ is indeed the amount of risk stemming from the i-th position.

This risk *attribution* is often done in terms of risk percentages. Dividing Eq. 26.20 by σ_V gives a percentage breakdown of the portfolio risk into the various assets

$$1 = \sum_{i=1}^{M} \frac{w_i}{\sigma_V} \frac{\partial \sigma_V}{\partial w_i} =: \sum_{i=1}^{M} A_i \ . \tag{26.21}$$

Here A_i denotes the percentage of risk attributed to the ith asset:

$$A_i = \frac{w_i}{\sigma_V} \frac{\partial \sigma_V}{\partial w_i} = w_i \frac{\partial \ln \sigma_V}{\partial w_i} \quad \text{for} \quad i = 1, \dots, M \ . \tag{26.22}$$

This risk attribution is widely used in portfolio management.

Another important concept of modern portfolio theory is the beta as defined in Eq. 12.21. There we defined the beta of a single asset with respect to an index. What is also of interest is the *beta* of a single asset with respect to the portfolio

it belongs to. This is given by

$$\beta_{i,V} \equiv \frac{\text{cov}\left[\delta \ln V_i, \delta \ln V\right]}{\text{var}(\delta \ln V)} \tag{26.23}$$

$$= \frac{1}{\sigma_V^2 \, \delta t} \sum_{k=1}^{M} w_k \, \text{cov}\left[\delta \ln V_i, \delta \ln V_k\right]$$

$$\approx \frac{1}{\sigma_V^2} \sum_{k=1}^{M} w_k C_{ik} \; ,$$

where we have used Eqs. 26.18 and 26.16. In vector notation this simplifies with Eq. 26.17 to

$$\boldsymbol{\beta}_V = \frac{\mathbf{C}\mathbf{w}}{\sigma_V^2} = \frac{\mathbf{C}\mathbf{w}}{\mathbf{w}^T \mathbf{C} \mathbf{w}} \; . \tag{26.24}$$

Comparing this beta with Eq. 26.19 we find that the marginal risk contribution of each asset is given simply by its beta with respect to the portfolio times the portfolio risk:

$$\frac{\partial \sigma_V}{\partial \mathbf{w}} = \sigma_V \, \boldsymbol{\beta}_V \quad \Longleftrightarrow \quad \frac{\partial \sigma_V}{\partial w_i} = \sigma_V \, \beta_{i,V} \text{ for } i = 1, \ldots, M \; . \tag{26.25}$$

Likewise, the percentage A_i of risk attributed to the ith asset given by Eq. 26.22 can expressed in terms of the asset beta as

$$A_i = w_i \, \beta_{i,V} \text{ for } i = 1, \ldots, M \; . \tag{26.26}$$

26.2 Portfolio Optimization

26.2.1 The Minimum Risk Portfolio

As a first optimization we minimize the portfolio risk, or equivalently the (square of the) portfolio volatility, Eq. 26.17, by varying the position weights w_k. The w_k can well be negative if short selling is allowed. The only constraint for the weights is Eq. 26.4 which must always hold.

In an optimization problem constraints can be taken into account by the method of *Lagrange multipliers*.[7] This method requires the construction of the *Lagrange function* $\mathcal{L}$ which is given by the function to be optimized (i.e., maximized or minimized, depending on the problem) minus a "zero" multiplied by a *Lagrange multiplier* λ. This zero is written in the form of the constraint to be satisfied, (here, $\sum_{k=1}^{M} w_k - 1 = 0$). The Lagrange function is thus given by

$$\mathcal{L} = \underbrace{\sum_{k,l=1}^{M} w_k C_{kl} w_l}_{\text{to be minimized}} - \lambda \underbrace{\left[\sum_{k=1}^{M} w_k - 1 \right]}_{\text{constraint}}.$$

In order to find the optimal values for w_k subject to this constraint, we differentiate $\mathcal{L}$ with respect to the variation parameters and set the resulting expression equal to zero. In other words, we locate the extremum of the Lagrange function.[8] Since we will be using this method time and again in the following, we demonstrate it here in detail:

$$0 = \frac{\partial \mathcal{L}}{\partial w_i}, \text{ for } i = 1, \ldots, M$$

$$= \sum_{k,l=1}^{M} \frac{\partial}{\partial w_i} (w_k C_{kl} w_l) - \lambda \sum_{k=1}^{M} \underbrace{\frac{\partial w_k}{\partial w_i}}_{\delta_{ki}} - \underbrace{\frac{\partial \lambda}{\partial w_i}}_{0}$$

$$= -\lambda + \sum_{k,l=1}^{M} [\underbrace{\frac{\partial w_k}{\partial w_i}}_{\delta_{ki}} C_{kl} w_l + w_k C_{kl} \underbrace{\frac{\partial w_l}{\partial w_i}}_{\delta_{li}}]$$

$$= -\lambda + \sum_{l=1}^{M} C_{il} w_l + \sum_{k=1}^{M} w_k C_{ki} .$$

---

[7]A introduction to the method of Lagrange multipliers for solving optimization problems subject to constraints can be found in [34], for instance.

[8]Since the Lagrange function differs from the function which we actually wish to optimize by only a zero, the extremum of the Lagrange function coincides with the desired extremum. However, this is only the case if the difference between the two functions is indeed zero, i.e., only if the constraint is satisfied. This is the essential idea of the method of Lagrange multipliers.

Since $\mathbf{C}$ is symmetric, we can write $w_k C_{ki} = C_{ik} w_k$ and therefore the two sums appearing here are identical.

$$0 = \frac{\partial \mathcal{L}}{\partial w_i} = 2 \sum_{l=1}^{M} C_{kl} w_l - \lambda , \quad \text{for} \quad i = 1, \dots, M .$$

This can also be written in vector form as

$$\mathbf{0} \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \mathbf{w}^T} = 2\mathbf{C}\mathbf{w} - \lambda \mathbf{1} . \tag{26.27}$$

The second derivative with respect to the variation parameters is

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^T} = 2\mathbf{C} .$$

Thus, a necessary condition for the existence of a *minimum* risk portfolio is that the matrix $\mathbf{C}$ is positive definite.[9] If $\mathbf{C}$ is positive definite then, as is well known from linear algebra,[10] the inverse $\mathbf{C}^{-1}$ exists and one can easily isolate the weights of the minimum risk portfolio from Eq. 26.27:

$$\mathbf{w} = \frac{\lambda}{2} \mathbf{C}^{-1} \mathbf{1} .$$

This is not yet useful since it still contains the unknown Lagrange multiplier λ. To determine λ we multiply by $\mathbf{1}^T$ from the left and use constraint 26.4

$$1 = \frac{\lambda}{2} \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \Longrightarrow \lambda = \frac{2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} .$$

Thus the weights are explicitly

$$\mathbf{w} = \frac{\mathbf{C}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \iff w_k = \frac{\sum_{l=1}^{M} \left(C^{-1} \right)_{kl}}{\sum_{r,s=1}^{M} \left(C^{-1} \right)_{rs}} , \quad k = 1, \dots M . \tag{26.28}$$

---

[9]If $\mathbf{C}$ were negative definite the extremum of the Lagrange Function would correspond to a *maximum* risk portfolio.

[10]A short and simple overview of *linear algebra* can be found in [78], for example.

Note that these weights depend only on the covariances of the assets in the portfolio but *not* on their returns.

The volatility of this minimum risk portfolio, i.e., the minimum volatility accessible by investing in the assets $V_1, \ldots V_k$ is

$$\sigma_{\min}^2 = \mathbf{w}^T \mathbf{C} \mathbf{w} = \frac{\left(\mathbf{C}^{-1}\mathbf{1}\right)^T \mathbf{C}\mathbf{C}^{-1}\mathbf{1}}{\left(\mathbf{1}^T \mathbf{C}^{-1}\mathbf{1}\right)^2} = \frac{1}{\mathbf{1}^T \mathbf{C}^{-1}\mathbf{1}} = \frac{1}{\sum_{r,s=1}^{M} \left(C^{-1}\right)_{\cdot rs}} \tag{26.29}$$

Here we have frequently made use of the fact that $\mathbf{C}^{-1}$ is symmetric, i.e.,

$$\mathbf{C}^{-1^T} = \mathbf{C}^{-1} , \tag{26.30}$$

since the inverse of any symmetric matrix is again a symmetric matrix.

The *expected* portfolio return of the minimum risk portfolio is

$$R_{\min} = \mathbf{w}^T \mathbf{R} = \frac{\mathbf{1}^T \mathbf{C}^{-1}\mathbf{R}}{\mathbf{1}^T \mathbf{C}^{-1}\mathbf{1}} = \sigma_{\min}^2 \mathbf{1}^T \mathbf{C}^{-1}\mathbf{R} = \sigma_{\min}^2 \sum_{k,l=1}^{M} \left(C^{-1}\right)_{kl} R_l , \tag{26.31}$$

where $\mathbf{R}$ denotes the vector of *expected* asset returns R_k for $k = 1, \ldots, M$.

26.2.2   The Efficient Frontier

Maximizing the Expected Return

While in the section above we *minimized* the portfolio *risk*, we now *maximize* the expected portfolio *return* R_V for a *given*, fixed risk (i.e., for a given portfolio volatility σ) by varying the position weights w_k. Thus we now have the constraint

$$\sigma_V^2 \overset{!}{=} \sigma^2 \Longleftrightarrow \mathbf{w}^T \mathbf{C} \mathbf{w} = \sum_{k,l=1}^{M} w_k C_{kl} w_l \overset{!}{=} \sigma^2 . \tag{26.32}$$

in addition to constraint 26.4. Again, we incorporate these two constraints into the optimization by using Lagrange multipliers. The Lagrange Function is

$$\mathcal{L} = \underbrace{\sum_{k=1}^{M} w_k R_k}_{\text{to be maximized}} - \lambda_1 \underbrace{\left[\sum_{k,l=1}^{M} w_k C_{kl} w_l - \sigma^2 \right]}_{\text{constraint}} - \lambda_2 \underbrace{\left[\sum_{k1}^{M} w_k - 1 \right]}_{\text{constraint}}$$

with two Lagrange multipliers λ_1 and λ_2. At the maximum of this function the first derivative must vanish:

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial w_k} = R_k - 2\lambda_1 \sum_{l=1}^{M} C_{kl} w_l - \lambda_2, \text{ for } k = 1, \ldots, M \ .$$

This can also be written in vector form as

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \mathbf{w}^T} = \mathbf{R} - 2\lambda_1 \mathbf{C} \mathbf{w} - \lambda_2 \mathbf{1} \ . \tag{26.33}$$

The second derivative with respect to the variation parameters is

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^T} = -2\lambda_1 \mathbf{C} \ .$$

A necessary condition for the existence of a *maximum* of the Lagrange function is that this second derivative is negative definite. Since the matrix $\mathbf{C}$ must be positive definite, we must require the Lagrange multiplier λ_1 to be positive.

An investor is of course most interested in the weights which produce this maximal return for a given risk. If $\mathbf{C}^{-1}$ exists[11] one can easily isolate these weights from Eq. 26.33:

$$2\lambda_1 \mathbf{w} = \mathbf{C}^{-1} \mathbf{R} - \lambda_2 \mathbf{C}^{-1} \mathbf{1} \Longleftrightarrow \tag{26.34}$$

$$w_k = \sum_{r=1}^{M} \left(C^{-1} \right)_{kr} \frac{R_r - \lambda_2}{2\lambda_1}, \text{ for } k = 1, \ldots, M \ .$$

---

[11]Which it does, if $\mathbf{C}$ is positive definite.

The Lagrange multipliers can be determined by exploiting the two constraints. First, inserting Eq. 26.34 into the constraint Eq. 26.32 yields

$$
\begin{aligned}
4\lambda_1^2 \sigma^2 &= 2\lambda_1 \mathbf{w}^T \mathbf{C} \mathbf{w} 2\lambda_1 \\
&= \left(\mathbf{C}^{-1}\mathbf{R} - \lambda_2 \mathbf{C}^{-1}\mathbf{1}\right)^T \mathbf{C} \left(\mathbf{C}^{-1}\mathbf{R} - \lambda_2 \mathbf{C}^{-1}\mathbf{1}\right) \\
&= \left(\mathbf{C}^{-1}\mathbf{R} - \lambda_2 \mathbf{C}^{-1}\mathbf{1}\right)^T (\mathbf{R} - \lambda_2 \mathbf{1}) \\
&= \left(\mathbf{R}^T \mathbf{C}^{-1} - \lambda_2 \mathbf{1}^T \mathbf{C}^{-1}\right) (\mathbf{R} - \lambda_2 \mathbf{1}) \\
&= \mathbf{R}^T \mathbf{C}^{-1}\mathbf{R} + \lambda_2^2 \mathbf{1}^T \mathbf{C}^{-1}\mathbf{1} - \lambda_2 \mathbf{1}^T \mathbf{C}^{-1}\mathbf{R} - \lambda_2 \mathbf{R}^T \mathbf{C}^{-1}\mathbf{1} \\
&= \mathbf{R}^T \mathbf{C}^{-1}\mathbf{R} + \frac{\lambda_2^2 - 2\lambda_2 R_{\min}}{\sigma_{\min}^2} \\
&= \mathbf{R}^T \mathbf{C}^{-1}\mathbf{R} + \frac{(\lambda_2 - R_{\min})^2}{\sigma_{\min}^2} - \frac{R_{\min}^2}{\sigma_{\min}^2} ,
\end{aligned}
\tag{26.35}
$$

where we have frequently used the fact that $\mathbf{C}^{-1}$ is symmetric[12] and we have also used Eqs. 26.29 and 26.31.

The second relation between λ_1 and λ_2 follows by left-multiplying Eq. 26.34 with $\mathbf{1}^T$ and using constraint 26.4:

$$
2\lambda_1 = \mathbf{1}^T \mathbf{C}^{-1}\mathbf{R} - \lambda_2 \mathbf{1}^T \mathbf{C}^{-1}\mathbf{1} = \frac{R_{\min} - \lambda_2}{\sigma_{\min}^2} ,
$$

where in the last step Eqs. 26.29 and 26.31 have been used again. Thus,

$$
\lambda_2 = R_{\min} - 2\lambda_1 \sigma_{\min}^2 .
\tag{26.36}
$$

---

[12]Therefore, we also have

$$
\mathbf{1}^T \mathbf{C}^{-1}\mathbf{R} = \sum_{r,s=1}^{M} R_s \left(C^{-1}\right)_{rs} = \sum_{r,s=1}^{M} R_s \left(C^{-1}\right)_{sr} = \mathbf{R}^T \mathbf{C}^{-1}\mathbf{1} .
$$

Inserting this into Eq. 26.35 yields

$$4\lambda_1^2\sigma^2 = \mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} - \frac{R_{\min}^2}{\sigma_{\min}^2} + \frac{\left(R_{\min} - 2\lambda_1\sigma_{\min}^2 - R_{\min}\right)^2}{\sigma_{\min}^2}$$

$$= \mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} - \frac{R_{\min}^2}{\sigma_{\min}^2} + 4\lambda_1^2\sigma_{\min}^2 \ .$$

We now have to solve the quadratic equation

$$4\lambda_1^2 = \frac{\mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2}{\sigma^2 - \sigma_{\min}^2}, \ \text{ wobei } \sigma^2 > \sigma_{\min}^2 \ .$$

Since $\sigma_{\min}$ is the *minimum* volatility, the condition $\sigma^2 > \sigma_{\min}^2$ is obvious. In the case $\sigma = \sigma_{\min}$, all volatilities are equal and the minimum is independent of the weights. Therefore, we do not need to consider this case. However, for λ_1 to be a real number we must also have

$$\mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} \geq R_{\min}^2/\sigma_{\min}^2 \ .$$

To show this, we need the *Cauchy-Schwarz inequality* for a scalar product of two arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$:

$$\left(\mathbf{x}^T\mathbf{y}\right)^2 \leq \left(\mathbf{x}^T\mathbf{x}\right)\left(\mathbf{y}^T\mathbf{y}\right) \ .$$

Equality holds if and only if the two vectors are linearly dependent, i.e., if there is a scalar λ such that $\mathbf{y} = \lambda\mathbf{x}$. If $\mathbf{C}$ is positive definite then $\mathbf{C}^{-1}$ is also positive definite. Therefore, the Cauchy-Schwarz inequality is preserved if "multiplied" by $\mathbf{C}^{-1}$ twice:

$$\left(\mathbf{x}^T\mathbf{C}^{-1}\mathbf{y}\right)^2 \leq \left(\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\right)\left(\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right) \tag{26.37}$$

$$\left(\mathbf{x}^T\mathbf{C}^{-1}\mathbf{y}\right)^2 = \left(\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\right)\left(\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right) \Longleftrightarrow \mathbf{y} = \lambda\mathbf{x} \ .$$

To apply this to our problem we choose $\mathbf{x} = \mathbf{1}$ and $\mathbf{y} = \mathbf{R}$:

$$\left(\mathbf{1}^T\mathbf{C}^{-1}\mathbf{R}\right)^2 \leq \left(\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}\right)\left(\mathbf{R}^T\mathbf{C}^{-1}\mathbf{R}\right)$$

$$\frac{\left(\mathbf{1}^T\mathbf{C}^{-1}\mathbf{R}\right)^2}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}} \leq \mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} \ .$$

Inserting now Eqs. 26.29 and 26.31 on the left hand side yields the desired result. Thus, λ_1 is indeed a real number.

$$\lambda_1 = \pm \frac{1}{2} \sqrt{\frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2}{\sigma^2 - \sigma_{\min}^2}} \ . \tag{26.38}$$

We have already argued above (just below Eq. 26.33) that λ_1 must be *positive* for the portfolio return to be a *maximum*. The negative root yields the portfolio with the *minimum* return for a given σ.

With this result the other Lagrange multiplier 26.36 becomes

$$\lambda_2 = R_{\min} \mp \sigma_{\min}^2 \sqrt{\frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2}{\sigma^2 - \sigma_{\min}^2}} \ . \tag{26.39}$$

Now we are in a position to explicitly formulate the maximum (and also the minimum) portfolio return as a function of the portfolio risk as measured by the portfolio volatility σ. Multiplying Eq. 26.33 by $\mathbf{w}^T$ from the left yields

$$\mathbf{w}^T \mathbf{R} = 2\lambda_1 \mathbf{w}^T \mathbf{C} \mathbf{w} + \lambda_2 \mathbf{w}^T \mathbf{1} \Longleftrightarrow$$

$$R_V = 2\lambda_1 \sigma^2 + \lambda_2 \ .$$

In the second line we made use of both constraints, Eqs. 26.4 and 26.32, and of the Definition of R_V, Eq. 26.3. Inserting now Eqs. 26.38 and 26.39 for the Lagrange multipliers we finally arrive at

$$R_V = R_{\min} + 2\lambda_1 \left(\sigma^2 - \sigma_{\min}^2 \right) \tag{26.40}$$

$$= R_{\min} \pm \sqrt{\sigma^2 - \sigma_{\min}^2} \sqrt{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2} \ .$$

This function with the plus sign in front of the square root yields the maximal portfolio return as a function of the portfolio risk and is called *efficient frontier*. The minus sign yields the minimal portfolio return as a function of the portfolio risk and is sometimes called the *lower branch* of the efficient frontier. Figure 26.1 taken from the Excel-Workbook PORTFOLIOMANAGE-MENT1996.XLSX from the download section [50] shows an example of an efficient frontier. The portfolios and concepts shown there will be discussed in detail in the sections below.
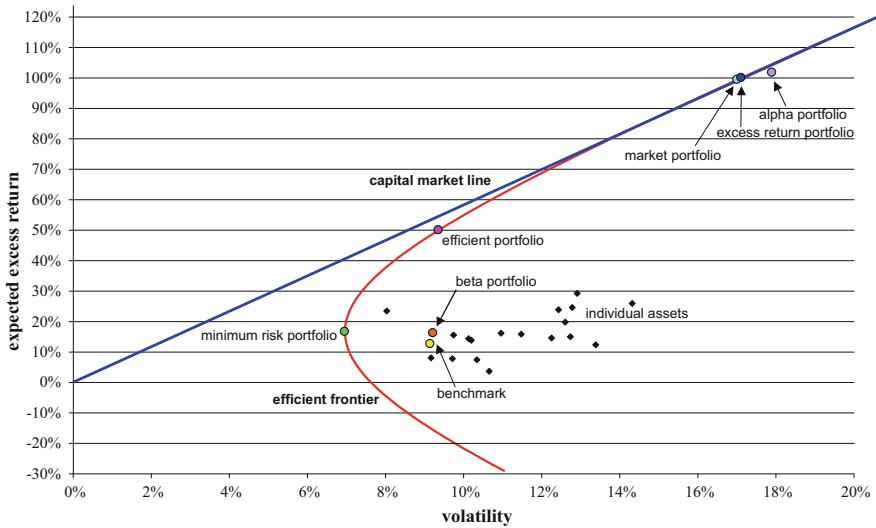
Fig. 26.1 A risk return diagram showing the efficient frontier, the capital market line and some important characteristic portfolios. The "assets" of the investment universe are in fact already indices, namely the DJ STOXX 600 sector indices. The benchmark is the DJ STOXX 600 itself

Most interesting for an investor are of course the position weights associated with the maximal return. These follow directly by inserting Eqs. 26.38 and 26.39 into Eq. 26.34:

$$
\begin{aligned}
\mathbf{w} &= \frac{\mathbf{C}^{-1}\mathbf{R}}{2\lambda_1} - \frac{\mathbf{C}^{-1}\mathbf{1}}{2\lambda_1}\lambda_2 \\
&= \frac{\mathbf{C}^{-1}\mathbf{R}}{2\lambda_1} - \frac{\mathbf{C}^{-1}\mathbf{1}}{2\lambda_1}\left(R_{\min} - 2\sigma_{\min}^2\lambda_1\right) \\
&= \frac{\mathbf{C}^{-1}\mathbf{R} - \mathbf{C}^{-1}\mathbf{1}R_{\min}}{2\lambda_1} + \sigma_{\min}^2\mathbf{C}^{-1}\mathbf{1} \\
&= \sigma_{\min}^2\mathbf{C}^{-1}\mathbf{1} \pm \sqrt{\frac{\sigma^2 - \sigma_{\min}^2}{\mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2}}\,\mathbf{C}^{-1}\left(\mathbf{R} - \mathbf{1}R_{\min}\right)\ , \qquad (26.41)
\end{aligned}
$$

or, written out in components,

$$w_k = \sigma_{\min}^2 \sum_{l=1}^{M} \left(C^{-1}\right)_{kl} \pm \sqrt{\frac{\sigma^2 - \sigma_{\min}^2}{\mathbf{R}^T \mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2}} \sum_{l=1}^{M} \left(C^{-1}\right)_{kl} (R_l - R_{\min})$$

for $k = 1, \ldots, M$.

Again, the plus sign hold for the weights in the upper branch whereas the minus sign holds for the weights in the lower branch of the efficient frontier.

Minimizing the Risk

Instead of *maximizing* the portfolio *return* R_V for a *given* fixed volatility σ, one can just as well *minimize* the portfolio risk (i.e., the variance σ^2) for a *given* fixed *return* R by varying the position weights w_k. This also yields the efficient frontier. This approach has the advantage that the risk along the efficient frontier is a *unique* function of the return, while the return as a function of the risk is not unique, see Fig. 26.1. Therefore it is worthwhile to demonstrate this approach here (from now on we will use vector notation only). We now have the constraint

$$R_V \overset{!}{=} R \Longleftrightarrow \mathbf{w}^T \mathbf{R} \overset{!}{=} R . \tag{26.42}$$

in addition to constraint 26.4. Again, we incorporate these two constraints into the optimization by using Lagrange multipliers. The Lagrange Function is

$$\mathcal{L} = \underbrace{\mathbf{w}^T \mathbf{Cw}}_{\text{to be minimized}} - \lambda_1 \underbrace{\left[\mathbf{w}^T \mathbf{R} - R\right]}_{\text{constraint}} - \lambda_2 \underbrace{\left[\mathbf{w}^T \mathbf{1} - 1\right]}_{\text{constraint}}$$

with two Lagrange multipliers λ_1 and λ_2. At the maximum of this function the first derivative must vanish

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \mathbf{w}^T} = 2\mathbf{Cw} - \lambda_1 \mathbf{R} - \lambda_2 \mathbf{1} . \tag{26.43}$$

The second derivative with respect to the variation parameters is

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^T} = 2\mathbf{C} .$$

A necessary condition for the existence of a *minimum* of the Lagrange function is that this second derivative is positive, i.e., the matrix $\mathbf{C}$ must again be positive definite. Then, $\mathbf{C}^{-1}$ exists and it is possible to isolate the weights:

$$2\mathbf{w} = \lambda_1 \mathbf{C}^{-1}\mathbf{R} + \lambda_2 \mathbf{C}^{-1}\mathbf{1}\} . \tag{26.44}$$

The Lagrange multipliers can be determined by exploiting the two constraints. Left-multiplying Eq. 26.44 by $\mathbf{1}^T$ and using constraint 26.4 yields

$$2 = 2\left(\mathbf{1}^T\mathbf{w}\right) = \lambda_1 \mathbf{1}^T\mathbf{C}^{-1}\mathbf{R} + \lambda_2 \mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}$$

$$\lambda_2 = \frac{2}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}} - \lambda_1 \frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{R}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}} = 2\sigma_{\min}^2 - R_{\min}\lambda_1 , \tag{26.45}$$

where in the last step Eqs. 26.29 and 26.31 have been used. Inserting λ_2 into Eq. 26.44 yields

$$\mathbf{w} = \frac{\lambda_1}{2}\mathbf{C}^{-1}\mathbf{R} + \left(\sigma_{\min}^2 - R_{\min}\frac{\lambda_1}{2}\right)\mathbf{C}^{-1}\mathbf{1} \tag{26.46}$$

$$= \frac{\lambda_1}{2}\left[\mathbf{C}^{-1}\mathbf{R} - R_{\min}\mathbf{C}^{-1}\mathbf{1}\right] + \sigma_{\min}^2\mathbf{C}^{-1}\mathbf{1} .$$

Inserting this into constraint 26.42 and using Eq. 26.31 we get

$$\mathbf{w}^T\mathbf{R} = R$$

$$\frac{\lambda_1}{2}\left[\mathbf{C}^{-1}\mathbf{R} - R_{\min}\mathbf{C}^{-1}\mathbf{1}\right]^T\mathbf{R} + \sigma_{\min}^2\left(\mathbf{C}^{-1}\mathbf{1}\right)^T\mathbf{R} = R$$

$$\frac{\lambda_1}{2}\left[\mathbf{R^T}\mathbf{C}^{-1}\mathbf{R} - R_{\min}\mathbf{1^T}\mathbf{C}^{-1}\mathbf{R}\right] + \sigma_{\min}^2\mathbf{1^T}\mathbf{C}^{-1}\mathbf{R} = R$$

$$\frac{\lambda_1}{2}\left[\mathbf{R^T}\mathbf{C}^{-1}\mathbf{R} - \frac{R_{\min}^2}{\sigma_{\min}^2}\right] + R_{\min} = R . \tag{26.47}$$

Thus with Eq. 26.45

$$\frac{\lambda_1}{2} = \frac{R - R_{\min}}{\mathbf{R^T}\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2} \quad\Longrightarrow$$

$$\frac{\lambda_2}{2} = \sigma_{\min}^2 - \frac{R_{\min}(R - R_{\min})}{\mathbf{R^T}\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2} . \tag{26.48}$$

Having the two Lagrange multipliers we are now able to calculate the minimum portfolio risk (as measured by the portfolio variance σ_V^2) as a function of the required portfolio return R. Multiplying Eq. 26.43 by $\mathbf{w}^T$ from the left and using Eq. 26.45 we obtain

$$\mathbf{w}^T \mathbf{C} \mathbf{w} = \frac{\lambda_1}{2} \mathbf{w}^T \mathbf{R} + \frac{\lambda_2}{2} \mathbf{w}^T \mathbf{1} \qquad \Longleftrightarrow$$

$$\sigma_V^2 = \frac{\lambda_1}{2} R + \frac{\lambda_2}{2} = \frac{\lambda_1}{2} (R - R_{\min}) + \sigma_{\min}^2 \ .$$

Thus

$$\sigma_V^2 = \sigma_{\min}^2 + \frac{(R - R_{\min})^2}{\mathbf{R^T C^{-1} R} - R_{\min}^2 / \sigma_{\min}^2} \ . \tag{26.49}$$

This function, i.e., the minimum portfolio risk as a function of the portfolio return, coincides with the *efficient frontier* introduced above. A quick consistency check: Solving this for R yields

$$R = R_{\min} \pm \sqrt{\left(\sigma_V^2 - \sigma_{\min}^2\right)\left(\mathbf{R^T C^{-1} R} - R_{\min}^2 / \sigma_{\min}^2\right)} \tag{26.50}$$

in complete agreement with Eq. 26.40. Asymptotically for $\sigma_V \longrightarrow \infty$ this approaches two straight lines:

$$R \sim \pm \sigma_V \sqrt{\mathbf{R^T C^{-1} R} - R_{\min}^2 / \sigma_{\min}^2} \ . \tag{26.51}$$

Inserting the Lagrange multipliers 26.48 into Eq. 26.44 we now find the weights to be

$$\mathbf{w} = \frac{R - R_{\min}}{\mathbf{R^T C^{-1} R} - R_{\min}^2 / \sigma_{\min}^2} \mathbf{C}^{-1} \mathbf{R} + \left(\sigma_{\min}^2 - \frac{R_{\min}(R - R_{\min})}{\mathbf{R^T C^{-1} R} - R_{\min}^2 / \sigma_{\min}^2}\right) \mathbf{C}^{-1} \mathbf{1}$$

$$= \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} + \frac{(R - R_{\min})}{\mathbf{R^T C^{-1} R} - R_{\min}^2 / \sigma_{\min}^2} \mathbf{C}^{-1} (\mathbf{R} - \mathbf{1} R_{\min})$$

$$= \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} + \frac{\sigma_V^2 - \sigma_{\min}^2}{R - R_{\min}} \mathbf{C}^{-1} (\mathbf{R} - \mathbf{1} R_{\min}) \ , \tag{26.52}$$

where in the last step we have used Eq. 26.49 to simplify things. In contrast to Eq. 26.41, we now do not have a "plus/minus sign" since for each given return R there is one *unique* portfolio which minimizes the risk. Nonetheless, Eq. 26.52 is in full agreement with Eq. 26.41 which can be verified by inserting the portfolio return R from Eq. 26.50 into the second line of Eq. 26.52:

$$\mathbf{w} = \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} \pm \frac{\sqrt{\left(\sigma_V^2 - \sigma_{\min}^2\right)\left(\mathbf{R}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2\right)}}{\mathbf{R}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2} \mathbf{C}^{-1}\left(\mathbf{R} - \mathbf{1}R_{\min}\right)$$

$$= \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} \pm \sqrt{\frac{\sigma_V^2 - \sigma_{\min}^2}{\mathbf{R}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2}} \mathbf{C}^{-1}\left(\mathbf{R} - \mathbf{1}R_{\min}\right) .$$

Now the weights are again a function of the portfolio risk σ_V^2 (the portfolio return R has disappeared) and therefore not unique.

26.2.3   The Sharpe Ratio and the Optimal Portfolio

We have seen above that, given our set of M assets,[13] for each given risk level there is a unique portfolio build with these M assets, which yields the maximal possible return. Or vice versa: for each required return there is a unique portfolio which minimizes the risk required to achieve this return. All these portfolios are the efficient portfolios making up the efficient frontier in the risk/return diagram. Since there are infinitely many returns (or risk levels), there are infinitely many portfolios along the efficient frontier. However, among all these portfolios there are two very special ones. One of them, the one with the absolute minimum risk attainable, was already presented in Sect. 26.2.1. The other one, the so called *optimal portfolio* is the most interesting portfolio and in fact the only portfolio among all the portfolios along the efficient frontier one should really invest in. It follows from the following considerations.

Let r_f denote the risk free rate. As a reward for taking risk compared to investing without any risk, an investor requires an expected return R with a premium above this risk free rate. Obviously, a rational investor will maximize this expected excess return (above the risk free rate) in relation to the risk taken, i.e., he will maximize the so-called Sharpe Ratio of the investment. The *Sharpe*

---

[13]The set of available (risky) assets is sometimes called *investment universe*.

Ratio of any investment with risk σ and expected return R is defined as

$$\gamma := \frac{R - r_f}{\sigma} \ . \tag{26.53}$$

With Eq. 26.40, the Sharpe Ratio for any portfolio on the efficient frontier is

$$\gamma = \frac{R_{\min} - r_f}{\sigma} \pm \sqrt{1 - \sigma_{\min}^2/\sigma^2} \sqrt{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2} \ .$$

We now have to decide which sign to use. The Cauchy-Schwarz Inequality 26.37 guarantees $\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2 \geq 0$ and $\sigma \geq \sigma_{\min}$ by construction. Thus, all terms after the $\pm$ sign above are non-negative. Since we are interested in finding the *maximal* Sharpe Ratio we use the plus sign here, meaning we use the plus sign in Eq. 26.40.

Moving along the efficient frontier by varying the portfolio risk σ we now want to find the portfolio with maximum γ:

$$0 = \frac{\partial \gamma}{\partial \sigma} = -\frac{R_{\min} - r_f}{\sigma^2} + \frac{\sqrt{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2}}{2\sqrt{1 - \sigma_{\min}^2/\sigma^2}} \frac{2\sigma_{\min}^2}{\sigma^3}$$

$$0 = r_f - R_{\min} + \sigma_{\min}^2 \frac{\sqrt{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2}}{\sqrt{\sigma^2 - \sigma_{\min}^2}} \ . \tag{26.54}$$

Solving this for σ^2 yields the variance of the optimal portfolio[14]

$$\sigma_m^2 = \sigma_{\min}^2 + \sigma_{\min}^4 \frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2}{\left(R_{\min} - r_f\right)^2} \ . \tag{26.55}$$

Inserting $\sigma_m^2 - \sigma_{\min}^2$ into Eq. 26.40 yields[15] the return R_m of the optimal portfolio

$$R_m = R_{\min} + \sigma_{\min}^2 \frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2}{R_{\min} - r_f} \ . \tag{26.56}$$

---

[14]Quantities belonging to the optimal portfolio carry a subscript m in the following.

[15]Remember that we already have committed ourselves to the plus sign in Eq. 26.40, i.e. to the upper branch of the efficient frontier.

We now insert the portfolio return, Eq. 26.56, and the portfolio volatility Eq. 26.55 into Eq. 26.52 to obtain (after some algebra[16]) explicitly the weights of the optimal portfolio:

$$\mathbf{w}_m = \sigma_{\min}^2 \mathbf{C}^{-1} \left(\mathbf{1} + \frac{\mathbf{R} - \mathbf{1}R_{\min}}{R_{\min} - r_f} \right) = \mathbf{w}_{\min} + \sigma_{\min}^2 \mathbf{C}^{-1} \frac{\mathbf{R} - \mathbf{1}R_{\min}}{R_{\min} - r_f} ,$$

(26.57)

where $\mathbf{w}_{\min}$ denotes the weights of the minimum risk portfolio as given by Eq. 26.28.

This can be simplified even further:

$$\mathbf{w}_m = \sigma_{\min}^2 \mathbf{C}^{-1} \left(\frac{\mathbf{1}R_{\min} - \mathbf{1}r_f + (\mathbf{R} - \mathbf{1}R_{\min})}{R_{\min} - r_f} \right)$$

$$= \frac{\sigma_{\min}^2 \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}r_f \right)}{R_{\min} - r_f} = \frac{\sigma_{\min}^2 \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}r_f \right)}{\sigma_{\min}^2 \mathbf{1}^T \mathbf{C}^{-1} \mathbf{R} - r_f}$$

$$= \frac{\mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}r_f \right)}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{R} - r_f / \sigma_{\min}^2} = \frac{\mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}r_f \right)}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{R} - \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} r_f} ,$$

where in the third step we have used Eq. 26.31 in the form $R_{\min} = \sigma_{\min}^2 \mathbf{1}^T \mathbf{C}^{-1} \mathbf{R}$ and in the last step we have used Eq. 26.29. Thus the weights of the optimal portfolio are given explicitly by

$$\mathbf{w}_m = \frac{\mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}r_f \right)}{\mathbf{1}^T \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}r_f \right)} .$$

(26.58)

---

[16]The algebra is:

$$\mathbf{w} = \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} + \frac{\sigma_V^2 - \sigma_{\min}^2}{R - R_{\min}} \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}R_{\min} \right)$$

$$= \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} + \frac{\sigma_V^2 - \sigma_{\min}^2}{\sigma_{\min}^2 \frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2 / \sigma_{\min}^2}{R_{\min} - r_f}} \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}R_{\min} \right)$$

$$= \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} + \frac{\sigma_{\min}^4 \frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2 / \sigma_{\min}^2}{(R_{\min} - r_f)^2}}{\sigma_{\min}^2 \frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2 / \sigma_{\min}^2}{R_{\min} - r_f}} \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}R_{\min} \right)$$

$$= \sigma_{\min}^2 \mathbf{C}^{-1} \mathbf{1} + \frac{\sigma_{\min}^2}{R_{\min} - r_f} \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1}R_{\min} \right) .$$

A Word of Caution

Since $\sigma_m^2 \geq \sigma_{\min}^2$ we must also have $R_m \geq R_{\min}$ since otherwise the Sharpe Ratio of the minimal risk portfolio would be larger than the Sharpe Ratio of the optimal portfolio, which is a contradiction to the requirement that the optimal portfolio has *maximal* Sharpe Ratio. Thus, from Eq. 26.56 we must have

$$\sigma_{\min}^2 \frac{\mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2}{R_{\min} - r_f} \overset{!}{\geq} 0 \ .$$

Since $\sigma_{\min}^2 \geq 0$ (as for any squared number) and since the Cauchy-Schwarz Inequality 26.37 guarantees that $\mathbf{R}^T\mathbf{C}^{-1}\mathbf{R} - R_{\min}^2/\sigma_{\min}^2 \geq 0$, this condition is only fulfilled for

$$R_{\min} > r_f \ . \tag{26.59}$$

Usually, $R_{\min} > r_f$ holds since, after all, the minimum risk portfolio is a portfolio of *risky* assets and therefore the market should require an expected return *above* the risk free rate as a compensation for the risk incurred by that portfolio. In reality, this is not always the case in the real world, which is demonstrated in the Excel workbook PORTFOLIOMANAGEMENT2002.XLSX from the download section [50].

If Eq. 26.59 does not hold, then the expected return of the "optimal" portfolio, Eq. 26.56, is lower than $R_{\min}$. Thus, the "optimal" portfolio cannot have maximum Sharpe Ratio. But Eq. 26.54 still holds. Therefore, the "optimal" portfolio still has extreme Sharpe Ratio. However, the extremum is now the *minimum* and the "optimal" portfolio is in fact the *worst* investment one could possibly make. Or in other words:

> If the return of the fully invested minimum risk portfolio is less than the risk free rate, then there is no *fully invested* portfolio with maximum Sharpe Ratio.

This is can also be understood intuitively: if markets are turning down (as was the case in the years 2000, 2001 and 2002) then any portfolio which is net long cannot possibly be the best choice. Portfolios which are net short can perform much better. By definition, a fully invested portfolio is 100% net long. All portfolios on the efficient frontier are fully invested. Therefore

the whole efficient frontier does *not* contain any good portfolio! There are many portfolios *above* the efficient frontier (namely the ones which are not forced to be 100% net long). This is demonstrated in the Excel workbook PortfolioManagement2002.xls. In Sect. 27.3 below, we will show how to find a portfolio with *maximum* Sharpe Ratio even in such a situation.

26.2.4 The Capital Market Line

Let's for now assume that $R_{\min} > r_f$ holds for our given *investment universe* (i.e., for the given set of M risky assets). Then the *optimal* portfolio is truly optimal, i.e., has indeed maximal Sharpe Ratio. The expected return as well as the risk of the optimal portfolio are uniquely determined by Eqs. 26.56 and 26.55. This is in stark contrast to the *efficient* portfolios discussed in Sect. 26.2.2 where one could find an efficient portfolio for any return requirement or (almost) any risk preference. In particular, the risk of the *optimal* portfolio cannot be chosen by the investor. However, even if the risk σ_m of the optimal portfolio doesn't coincide with the risk preference σ_{required} of the investor, the investor should still invest in the optimal portfolio, although not all of his money. If $\sigma_{\text{required}} < \sigma_m$ the investor should only invest a percentage w of the total capital in the optimal portfolio and the rest of the capital should be invested risk free (in a money market account). On the other hand, if $\sigma_{\text{required}} > \sigma_m$ the investor should borrow money (from the money market) and invest the total sum of his own capital and the loan in the optimal portfolio. This is called a *leveraged investment*. Since the optimal portfolio has the maximum return the investor can expect for the risk taken and since investing or borrowing in the money market doesn't produce any new risk, this strategy gives the best possible Sharpe Ratio for any required risk level.

As long as $R_{\min} > r_f$, one should therefore always invest in a mixture of the money market account and the optimal portfolio. The return R_V of such an investment is

$$R_V = w R_m + (1 - w) \, r_f \, , \tag{26.60}$$

where R_m denotes the return of the (fully invested) optimal portfolio and

$$w := \mathbf{w}^T \mathbf{1} \tag{26.61}$$

denotes the part of the money invested in the optimal portfolio. For $\sigma_{\text{required}} < \sigma_m$ we will have $w < 1$ and for $\sigma_{\text{required}} > \sigma_m$ we will have $w > 1$. The risk of such an investment is simply

$$\sigma_V = w\sigma_m \tag{26.62}$$

since the money market account carries no risk. We can write the return as

$$
\begin{aligned}
R_V &= r_f + (R_m - r_f)w \\
&= r_f + (R_m - r_f)\frac{\sigma_V}{\sigma_m} \\
&= r_f + \gamma_m\sigma_V \quad \text{mit} \quad \gamma_m = \frac{R_m - r_f}{\sigma_m} \ .
\end{aligned}
\tag{26.63}
$$

Thus, the best possible investment return R one can expect as a function of the investment risk σ_V is a straight line with a slope γ_m given by the Sharpe Ratio of the optimal portfolio! This straight line is called the *capital market line*. Every investor should invest in the optimal portfolio, even if he does not prefer the risk of the optimal portfolio. In this case he should distribute his investment between the risk free account and the optimal portfolio such that the mixture as a whole has the desired risk level. The total investment risk in Eq. 26.62 can very easily be controlled by the part w the investor assigns to the optimal portfolio (i.e., to the M risky assets). This is the best possible investment strategy whenever $R_{\text{min}} > r_f$.

It follows directly from Eq. 26.63 that the Sharpe Ratio of this investment equals the Sharpe Ratio of the optimal portfolio, i.e., is maximal:

$$\gamma_V \equiv \frac{R_V - r_f}{\sigma_V} = \frac{(r_f + \gamma_m\sigma_V) - r_f}{\sigma_V} = \gamma_m \ .$$

Equation 26.63 can be read in the following way: For each unit of additional risk σ_V an investor is willing to take, the expected return of the total investment increases by an amount γ_m, i.e., by the Sharpe Ratio of the optimal portfolio. This Sharpe Ratio is therefore the *market price of risk* of the investment universe consisting of the M risky assets (and of course the money market account). This is in complete agreement with the market price of risk in Eq. 13.49, where the investment universe consisted of only one risky asset, namely the risk factor S. As we have shown in Eq. 2.32, the drift $\widetilde{\mu}$ appearing in Eq. 13.49 is the expected return for linear compounding. Since in asset management returns

are usually defined as relative (as opposed to logarithmic) price changes, linear compounding is indeed applicable (see Eq. 2.31). Thus, $\widetilde{\mu}$ in Eq. 13.49 and R_m in Eq. 26.63 exactly correspond to each other and therefore the Sharpe Ratio and the market price of risk are the same thing. We can summarize the above insights in the following theorem which is at the heart of classical portfolio theory:

> The Sharpe Ratio γ_m of the optimal portfolio for an investment universe consisting of M risky assets is the slope of the capital market line describing the expected return of the optimal investment strategy as a function of investment risk. The Sharpe Ratio is therefore the market price of risk for this investment universe.

Remember that all of the above only holds if $R_{\min} > r_f$. We will defer the treatment of the opposite situation until Sect. 27.3.

Being Not Fully Invested

As we have just seen, it is usually not wise to have all of the investor's capital invested in the portfolio of the risky assets. Therefore, from now on, we will drop the constraint Eq. 26.4 of being fully invested and replace it with Eq. 26.61 which is no constraint at all, since w could be any number. Part of an investor's capital could thus be invested risk free or the portfolio could be leveraged, i.e., it could be worth more than the investor's capital, being partly financed by a loan. However, we always assume that all investments in *risky* assets are accounted for in the weights w_k for $k = 1, \ldots M$. Thus, if $\mathbf{w}^T \mathbf{1} \neq 1$ then the rest of the investor's capital can only be invested in or borrowed from the (risk free) money market or bank account. The weight of this risk free investment (or loan if < 0) is simply

$$w_f := 1 - \mathbf{w}^T \mathbf{1} . \tag{26.64}$$

The weights w_k for $k = 1, \ldots M$ (and also w_f in Eq. 26.64) are all expressed as percentages of the *total investment* consisting of the risky part and the money market account. Likewise, if $\mathbf{w}^T \mathbf{1} \neq 1$ the return as defined in Eq. 26.3 is *not* the return of the whole investment but only the return of the *risky part* of the investment, expressed as a percentage of the whole investment. The reader should make sure to fully understand this subtlety which does not arise for fully invested portfolios where the investment value equals the value of the

risky assets.[17] The extension of Eq. 26.3 to the general case of not fully invested portfolios, i.e., the expected total return R_V of the investment V including the risk free part is

$$R_V = \mathbf{w}^T \mathbf{R} + \left[1 - \mathbf{w}^T \mathbf{1}\right] r_f , \qquad (26.65)$$

which of course reduces to Eq. 26.3 for fully invested portfolios, i.e. for $\mathbf{w}^T \mathbf{1} = 1$.

The expected *excess* return (i.e., the expected return above the risk free rate) of the investment is

$$
\begin{aligned}
R_V - r_f &= \mathbf{w}^T \mathbf{R} + \left[1 - \mathbf{w}^T \mathbf{1}\right] r_f - r_f \\
&= \mathbf{w}^T \left(\mathbf{R} - \mathbf{1} r_f\right) \\
&= \sum_{k=1}^{M} w_k \left(R_k - r_f\right) .
\end{aligned}
\qquad (26.66)
$$

This is simply the sum of the expected excess returns of the *risky* assets, which makes sense since the money market position obviously does not have any excess return above the risk free rate.

It is common practice in portfolio theory and asset management to always consider the *excess* return above the risk free rate instead of the "normal" return, since then all equations look less complicated. In the following we denote excess returns by "hats". In this notation Eq. 26.66 simplifies to

$$\widehat{R}_V = \mathbf{w}^T \widehat{\mathbf{R}} \quad \text{mit} \quad \widehat{\mathbf{R}} := \mathbf{R} - \mathbf{1} r_f . \qquad (26.67)$$

Equations 26.65 and 26.67 hold for *every* situation, no matter whether constraint 26.4 holds or not. For instance it holds for the *excess* return of the minimum risk fully invested portfolio: by definition we have $R_{\min} = \mathbf{w}^T \mathbf{R}$ with weights given in Eq. 26.28. Therefore Eq. 26.31 is also true "with hats" above all returns:

$$\widehat{R}_{\min} = \mathbf{w}^T \widehat{\mathbf{R}} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} = \sigma_{\min}^2 \mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} \qquad (26.68)$$

with $\sigma_{\min}^2$ as defined in Eq. 26.29.

---

[17] We will continue to use also the expression "portfolio" and mean by it the *total* investment consisting of the risky part and the money market account.

26.3 Alternative Portfolio Management Approaches

Before we dive into the details of classical portfolio management methods in the following chapters, we present here a few alternative approaches for portfolio management and asset allocation.

26.3.1 Value Investing

The value investor tries to estimate the inner value of an investment by *fundamental analysis* [77]. The fundamental analysis is based on an intensive analysis of the assets' issuer's balance sheet and other accounting figures. Other factors, like management quality, prospect of the business model, continuity of success over the past five or ten years, and further available information of the company are taken into account. Macro economic factors may also be included in the investigations. By nature, such an analysis is quite time intensive and requires a lot of experience. Only those assets which are traded at prices well below there estimated inner value will be bought. The difference between inner value and current market price (the *margin of safety*) serves as a reserve against the estimation error, since this type of analysis is based on many uncertain factors. Simpler versions of this investment approach are based on just a few, publicly available key figures (see, for example, [152]). Under the basic assumption that inner value and market price converge in the long run, value investing can be very profitable.

26.3.2 Behavioral Finance

Behavioral finance studies the psychology and behaviour of investors and analyses specifically the investment decision process of investors [144]. Here, it is especially interesting what causes irrational investment decisions. Behavioural scientists also perform experiments designed to shed light on the causes of investment decisions. The behavioural finance research indicates that investment decisions of portfolio managers, how professional they might be, are far too often not based on rational reasoning. A commonly well known effect is herd mentally: If some market participants start to buy some stock, causing its market prices to go up a bit, this may attract other investors, which start to invest in the same stock, pushing its price further up, until "everybody" starts to buy the same stock in order not to be the only one who misses the potential

profit. Investing strategies based on behavioural finance seek to profit from other investors' irrational decisions.

26.3.3   Chart Analysis

Chart analysis searches for recurring patterns in the graphical representation (i.e. *chart*) of historical market price series. If a pattern has been identified, it is used to estimate the future market price evolution. If the market price process is strictly Markovian, such patterns could not exist (i.e. could not be used to predict future prices). Therefore, chart analysis assumes that the stochastic process is superimposed by some deterministic process. There are indeed some well known recurring patterns, for example the weekly cycle. Another effect is reflected in the commonly known rule "sell in May, go away, but remember to return in September", based on the observation that market price returns are higher during the winter months. Unfortunately, these simple rules are not stable enough to base investment decisions on them. Chart analysis goes far beyond these simple rules. Though chart analysis is neither empirically nor theoretically well founded, its still favoured by many market participants.

27

Attributes and Their Characteristic Portfolios

The financial instruments, or in general the assets, of a portfolio have many *characteristics* or *attributes* such as expected return, market capitalization, beta with respect to an index, membership in a certain economic sector, etc. If we denote a certain attribute by a_i for asset i (with value V_i) then the *exposure* of a portfolio V (with weights w_k for $k = 1, \ldots M$) to this particular attribute is defined as

$$a_V \equiv \sum_{k=1}^{M} w_k a_k = \mathbf{w}^T \mathbf{a} , \qquad (27.1)$$

where we have used the obvious notation $\mathbf{a}^T = (a_1, \ldots a_M)$. If, for instance the characteristics a_i are measures of how strongly the assets $i = 1, \ldots M$ belong to the automotive industry then a_V is the exposure of the portfolio to the automotive industry. Another example: If the characteristics a_i are the asset returns R_i then the exposure a_V is simply the portfolio return R_V. From these examples one can already guess that characteristics (or attributes) are a very general and rather abstract concept with broad applications.[1]

---

[1]We will give specific examples in the sections below.

The *characteristic portfolio*[2] for an attribute **a** is defined as the portfolio with minimum risk and unit exposure, i.e., with

$$\mathbf{w}_a^T \mathbf{a} = 1 \ . \tag{27.2}$$

Other than this, there are no constraints. In particular, there is *no* constraint like Eq. 26.4 to be fully invested. Depending on the attribute, the characteristic portfolio can contain significant leverage.

To find the characteristic portfolio we now simply have to minimize its risk observing the one single constraint 27.2. The Lagrange function is thus given by

$$\mathcal{L} = \underbrace{\mathbf{w}_a^T \mathbf{C} \mathbf{w}_a}_{\text{to be minimized}} - \lambda \underbrace{\left[\mathbf{w}_a^T \mathbf{a} - 1 \right]}_{\text{constraint}} \ .$$

In order to find the optimal values w_k subject to this constraint, we differentiate $\mathcal{L}$ with respect to the variation parameters and set the resulting expression equal to zero.

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \mathbf{w}_a^T} = 2\mathbf{C}\mathbf{w}_a - \lambda \mathbf{a} \ . \tag{27.3}$$

The second derivative of $\mathcal{L}$ with respect to **w** is **C**. Thus, for positive definite covariance matrices Eq. 27.3 is indeed a minimum. Solving for **w** yields

$$\mathbf{w}_a = \frac{\lambda}{2} \mathbf{C}^{-1} \mathbf{a} \ .$$

Inserting the transpose of this into Eq. 27.2 (observing that **C** is symmetric) yields for the Lagrange multiplier

$$\frac{\lambda}{2} = \frac{1}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \ .$$

Thus we find the weights of the characteristic portfolio to be

$$\mathbf{w}_a = \frac{\mathbf{C}^{-1} \mathbf{a}}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \ . \tag{27.4}$$

---

[2]We will use an index a with quantities belonging to the characteristic portfolio of an attribute **a**. For instance its value is denoted by V_a, its weights by $\mathbf{w}_a$, its return by R_a, its volatility by σ_a, etc.

In the remainder of this section we will assemble a few general properties of characteristic portfolios which will be used time and again in the examples below.

27.1 General Properties of Characteristic Portfolios

The extent of investment (the leverage) as defined in Eq. 26.61 of the characteristic portfolio is

$$w \equiv \mathbf{1}^T \mathbf{w}_a = \frac{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{a}}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \ . \tag{27.5}$$

The variance of the characteristic portfolio also follows directly using the weights from Eq. 27.4

$$\sigma_a^2 \equiv \mathbf{w}_a^T \mathbf{C} \mathbf{w}_a = \frac{\left(\mathbf{a}^T \mathbf{C}^{-1}\right) \mathbf{C} \left(\mathbf{C}^{-1} \mathbf{a}\right)}{\left(\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a} \ .\right)^2} \ ,$$

Thus the variance of the characteristic portfolio is

$$\sigma_a^2 = \frac{1}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \tag{27.6}$$

and the excess return of the characteristic portfolio is

$$\widehat{R}_a \equiv \mathbf{w}_a^T \widehat{\mathbf{R}} = \frac{\mathbf{a}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} = \sigma_a^2 \mathbf{a}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} \ . \tag{27.7}$$

These two results yield the Sharpe Ratio, Eq. 26.53, of the characteristic portfolio.

$$\gamma_a \equiv \frac{\widehat{R}_a}{\sigma_a} = \sigma_a \mathbf{a}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} = \frac{\mathbf{a}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\sqrt{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}}} \ . \tag{27.8}$$

The beta as defined in Eq. 26.24 of a single asset with respect to the characteristic portfolio becomes

$$\beta_a = \frac{\mathbf{C} \mathbf{w}_a}{\sigma_a^2} = \left(\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}\right) \mathbf{C} \frac{\mathbf{C}^{-1} \mathbf{a}}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \ .$$

Thus the asset betas with respect to the characteristic portfolio are simply the asset attributes:

$$\boldsymbol{\beta}_a = \frac{\mathbf{C}\mathbf{w}_a}{\sigma_a^2} = \mathbf{a} \ . \tag{27.9}$$

With these asset betas, the marginal risk contribution of each asset, Eq. 26.25, becomes

$$\frac{\partial \sigma_a}{\partial \mathbf{w}_a} = \sigma_a \mathbf{a} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}}} \tag{27.10}$$

and the percentage A_i in Eq. 26.26 of risk attributed to the i-th asset is simply

$$A_i = w_i a_i \ \text{ for } i = 1, \ldots, M \ . \tag{27.11}$$

27.1.1 Relations Involving Several Characteristic Portfolios

Let's now look at relations involving several characteristic portfolios. The covariance between the characteristic portfolio V_a of an attribute $\mathbf{a}$ and the characteristic portfolio V_b of another attribute $\mathbf{b}$ is by definition

$$\text{cov}\left[\delta \ln V_a, \delta \ln V_b\right] = \sum_{i=1}^{M} \sum_{k=1}^{M} w_{a,i} w_{b,k} \, \text{cov}\left[\delta \ln V_i, \delta \ln V_k\right]$$

$$= \delta t \sum_{i=1}^{M} \sum_{k=1}^{M} w_{a,i} C_{ik} w_{b,k} = \delta t \, \mathbf{w}_a^T \mathbf{C} \mathbf{w}_b \ .$$

Inserting the weights of the characteristic portfolio V_b in accordance with Eq. 27.4 yields

$$\text{cov}\left[\delta \ln V_a, \delta \ln V_b\right] = \delta t \, \mathbf{w}_a^T \frac{\mathbf{b}}{\mathbf{b}^T \mathbf{C}^{-1} \mathbf{b}} = \delta t \, b_a \sigma_b^2 \ ,$$

where Eq. 27.6 was used for the variance σ_b^2 of portfolio V_b, and b_a denotes the exposure of portfolio V_a to attribute $\mathbf{b}$:

$$b_a := \mathbf{w}_a^T \mathbf{b} \ .$$

Similarly, one can also write

$$\text{cov}\left[\delta \ln V_a, \delta \ln V_b\right] = \delta t \left(\mathbf{w}_a^T \mathbf{C}\right) \mathbf{w}_b = \frac{\mathbf{a}^T}{\mathbf{a}^T \mathbf{C}^{-1}\mathbf{a}} \mathbf{w}_b = \delta t \, a_b \sigma_a^2 \,,$$

where $a_b := \mathbf{w}_b^T \mathbf{a} = \mathbf{a}^T \mathbf{w}_b$ denotes the exposure of portfolio V_b to attribute $\mathbf{a}$. Comparing the two results yields

$$b_a \sigma_b^2 = a_b \sigma_a^2 \,. \tag{27.12}$$

For any positive number x and any attribute $\mathbf{a}$ the characteristic portfolio of the attribute $x\mathbf{a}$ has the weights

$$\mathbf{w}_{xa} = \frac{\mathbf{C}^{-1}x\mathbf{a}}{x\mathbf{a}^T \mathbf{C}^{-1}x\mathbf{a}} = \frac{1}{x}\mathbf{w}_a \,, \tag{27.13}$$

as follows directly from Eq. 27.4.

If an attribute $\mathbf{a}$ is a linear combination of two attributes $\mathbf{b}$ and $\mathbf{c}$, i.e., if

$$\mathbf{a} = x\mathbf{b} + y\mathbf{c} \,,$$

then the characteristic portfolio for attribute $\mathbf{a}$ is a linear combination of the characteristic portfolios for attributes $\mathbf{b}$ and $\mathbf{c}$ with the following weights

$$\mathbf{w}_a = x\frac{\sigma_a^2}{\sigma_b^2}\mathbf{w}_b + y\frac{\sigma_a^2}{\sigma_c^2}\mathbf{w}_c \,. \tag{27.14}$$

This can be shown by starting from Eq. 27.9

$$\mathbf{a} = \boldsymbol{\beta}_a = \frac{\mathbf{C}\mathbf{w}_a}{\sigma_a^2}$$

$$x\mathbf{b} + y\mathbf{c} = \frac{\mathbf{C}\mathbf{w}_a}{\sigma_a^2}$$

$$\sigma_a^2 \left(x\mathbf{C}^{-1}\mathbf{b} + y\mathbf{C}^{-1}\mathbf{c}\right) = \mathbf{w}_a \,.$$

Using now Eqs. 27.4 and 27.6 we can write $\mathbf{C}^{-1}\mathbf{b} = \mathbf{w}_b/\sigma_b^2$ and analogously $\mathbf{C}^{-1}\mathbf{c} = \mathbf{w}_c/\sigma_c^2$ to arrive at Eq. 27.14. The variance σ_a^2 can be expressed in terms of the variances of the characteristic portfolios for attributes $\mathbf{b}$ and $\mathbf{c}$ by

simply left-multiplying Eq. 27.14 by $\mathbf{a}^T$ and using the defining Eq. 27.2 for the characteristic portfolio V_a:

$$\mathbf{a}^T \mathbf{w}_a = x \frac{\sigma_a^2}{\sigma_b^2} \mathbf{a}^T \mathbf{w}_b + y \frac{\sigma_a^2}{\sigma_c^2} \mathbf{a}^T \mathbf{w}_c$$

$$1 = x \frac{\sigma_a^2}{\sigma_b^2} a_b + y \frac{\sigma_a^2}{\sigma_c^2} a_c ,$$

Thus

$$\frac{1}{\sigma_a^2} = x \frac{a_b}{\sigma_b^2} + y \frac{a_c}{\sigma_c^2} , \qquad (27.15)$$

where again $a_b = \mathbf{w}_b^T \mathbf{a} = \mathbf{a}^T \mathbf{w}_b$ denotes the exposure of portfolio V_b to attribute $\mathbf{a}$ and similarly for a_c. Using now Eq. 27.12 one finds that the exposures of portfolio V_a with respect to the attributes $\mathbf{b}$ and $\mathbf{c}$ fulfill the Equation

$$1 = x b_a + y c_a .$$

27.2 The Leverage

Let's choose the attribute vector $\mathbf{L}$ to be

$$\mathbf{L} = \mathbf{1} \Longleftrightarrow L_k = 1 , \ k = 1, \ldots M . \qquad (27.16)$$

The exposure of *any* portfolio V to this particular attribute is according to Eq. 27.1 simply the *extent of investment* in risky assets which is also called *leverage*

$$L_V = \mathbf{w}^T \mathbf{L} = \mathbf{w}^T \mathbf{1} = \sum_{k=1}^{M} w_k . \qquad (27.17)$$

A fully invested portfolio has $L_V = 1$. From the constraint defining the *characteristic* portfolio, Eq. 27.2, we find that the characteristic portfolio V_L for this particular attribute is just the fully invested portfolio with minimum risk. According to Eqs. 27.4, 27.6, 27.7 and 27.9, this portfolio has the following

weights, variance, excess return and betas

$$\mathbf{w}_L = \frac{\mathbf{C}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}}, \quad \sigma_L^2 = \frac{1}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}}, \quad \boldsymbol{\beta}_L = \mathbf{1} \tag{27.18}$$

$$\widehat{R}_L = \frac{\mathbf{1}^T\mathbf{C}^{-1}\widehat{\mathbf{R}}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}} = \sigma_L^2 \mathbf{1}^T\mathbf{C}^{-1}\widehat{\mathbf{R}},$$

in full accordance with Eqs. 26.28, 26.29 and 26.68 for the fully invested minimum risk portfolio derived in Sect. 26.2.1.

The fact that every asset in the portfolio has a beta of one means that the marginal risk contribution, Eq. 26.25, of each asset is the portfolio risk itself, and that the percentage of risk attributed to each asset, Eq. 26.26, is simply the asset's weight:

$$\frac{\partial \sigma_L}{\partial w_{L,i}} = \sigma_L, \quad A_{L,i} = w_{L,i} \quad \text{for} \quad i = 1, \dots, M.$$

27.3 The Excess Return

Let's now choose the attribute vector to be the assets' *expected excess returns* (i.e., returns above the risk free rate)

$$\mathbf{A} = \widehat{\mathbf{R}} = \mathbf{R} - r_f \mathbf{1} \quad \Longleftrightarrow \quad A_k = R_k - r_f, \quad k = 1, \dots M. \tag{27.19}$$

According to Eq. 27.1, the exposure of *any* portfolio V to this particular attribute is

$$A_V = \mathbf{w}^T \widehat{\mathbf{R}} = \sum_{k=1}^{M} w_k \left(R_k - r_f \right). \tag{27.20}$$

Comparing this with Eq. 26.66 we find the *excess* return of the total investment (including the money market position) to be equal to A_V

$$R_V - r_f = A_V.$$

By definition, the *characteristic* portfolio V_A for this attribute is the *minimum risk* portfolio with unit exposure, i.e., with excess return equal to 1.

$$\widehat{R}_A \equiv R_A - r_f \stackrel{!}{=} 1. \tag{27.21}$$

Thus, for a *given* excess return of 1 we minimize the risk or equivalently: we maximize the ratio $1/\sigma_A = \widehat{R}_A/\sigma_A$. But this ratio is just the investment's *Sharpe Ratio*: The Sharpe Ratio of *any* investment is defined as

$$\gamma_V \equiv \frac{\widehat{R}_V}{\sigma_V} = \frac{\mathbf{w}^T \widehat{\mathbf{R}}}{\sqrt{\mathbf{w}^T \mathbf{C} \mathbf{w}}} \qquad (27.22)$$

with $\widehat{R}_V$ and σ_V as in the general equations 26.14 and 26.67. For the Sharpe Ratio of the characteristic portfolio $\widehat{R}_V = A_V = 1$ and σ is as small as possible. Thus *the characteristic portfolio for attribute A has maximal Sharpe Ratio.*

$$\gamma_{\max} = \gamma_A = \frac{1}{\sigma_A} . \qquad (27.23)$$

This holds for *any* situation, even if Inequality 26.59 does *not* hold.

For any positive number λ, an investment with weights $\lambda \cdot w_k$ for $k = 1, \ldots M$, will have the *same* Sharpe Ratio, as can immediately be seen from Eq. 27.22. All portfolios with maximal Sharpe Ratio (but different leverage via different λ) lie on a straight line with slope $\gamma_{\max}$ called the *Capital Market Line*. If Inequality 26.59 holds, this is the same Capital Market Line as introduced in Sect. 26.2.4. Therefore, when maximizing the Sharpe Ratio there is one degree of freedom left, which could for instance be used to satisfy the constraint Eq. 27.21 for the characteristic portfolio. When Inequality 26.59 holds, the degree of freedom could also be used to satisfy the "fully invested" constraint Eq. 26.4 instead of satisfying Eq. 27.21. In this situation the optimal portfolio introduced in Sect. 26.2.4 has the same (maximal) Sharpe Ratio as the characteristic portfolio to attribute A, i.e., both portfolios lie on the same capital market line. However, when Inequality 26.59 doesn't hold, the "optimal" portfolio introduced in Sect. 26.2.3 is in fact the worst possible choice, i.e., has *minimum* Sharpe Ratio. Therefore, to be on the safe side in all situations, one should always use the characteristic portfolio for the attribute A to construct the Capital Market Line.

Figure 27.1 taken from the Excel Workbook PortfolioManage-ment2002.xlsx from the download section [50] shows an example of such a situation. It shows the same investment universe as in Fig. 26.1 but for the year 2002. In this year, markets were going down consistently. Fully invested portfolios (which are by definition forced to be 100% net long) could not perform well in this year. The optimal portfolio and in fact all portfolios on the efficient frontier are all bad choices. The capital market line constructed
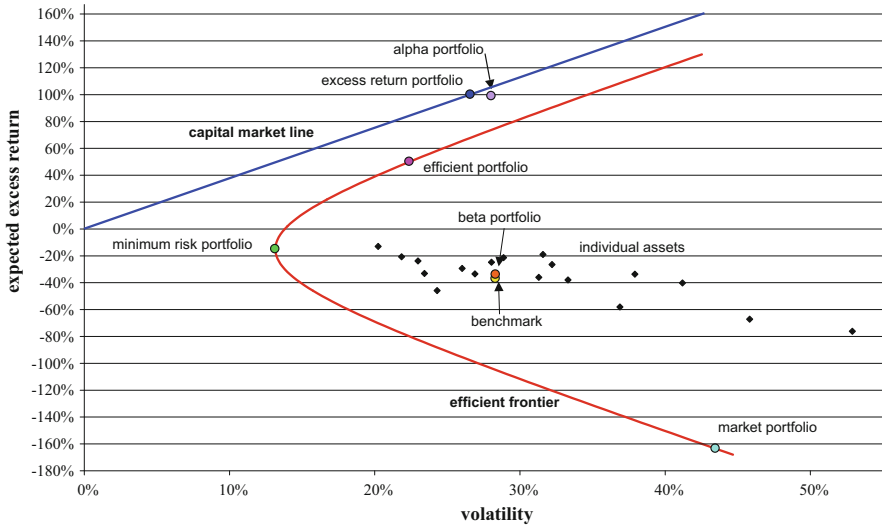
Fig. 27.1 The risk return diagram for the same investment universe as in Fig. 26.1, but for a different time period. In this period, the expected excess return of the minimum risk portfolio is negative. The best portfolios are *not* fully invested, and therefore, the efficient frontier does not reach the capital market line

with the characteristic portfolio for the excess return lies way above the efficient frontier.

According to Eqs. 27.4, 27.6 and 27.9, the characteristic portfolio V_A for the attribute A has the following weights, variance and betas

$$\mathbf{w}_A = \frac{\mathbf{C}^{-1}\widehat{\mathbf{R}}}{\widehat{\mathbf{R}}^{\mathsf{T}}\mathbf{C}^{-1}\widehat{\mathbf{R}}} = \sigma_A^2\, \mathbf{C}^{-1}\widehat{\mathbf{R}}$$

$$\sigma_A^2 = \frac{1}{\widehat{\mathbf{R}}^T\mathbf{C}^{-1}\widehat{\mathbf{R}}}$$

$$\frac{\mathbf{C}\mathbf{w}_A}{\sigma_A^2} = \boldsymbol{\beta}_A = \widehat{\mathbf{R}}\ . \tag{27.24}$$

With these betas, the marginal risk contribution, Eq. 26.25, of each asset and the percentage of risk attributed to each asset, Eq. 26.26, are:

$$\frac{\partial \sigma_A}{\partial w_{A,i}} = \sigma_A \widehat{R}_i\ ,\ \ A_{A,i} = w_{A,i}\,\widehat{R}_i \quad for \quad i = 1, \ldots, M\ .$$

The percentage of risk attributed to an asset is thus simply the contribution of that asset to the expected excess return of the characteristic portfolio V_A.

To get an expression for the excess return of *any* portfolio V, left-multiply the last line in Eq. 27.24 by the weights of portfolio V

$$\mathbf{w}^T \widehat{\mathbf{R}} = \frac{\mathbf{w}^T \mathbf{C} \mathbf{w}_A}{\sigma_A^2} \, .$$

By definition, an investment's excess return divided by the investment's volatility is the *Sharpe Ratio* of that investment:

$$\gamma_V = \frac{\mathbf{w}^T \widehat{\mathbf{R}}}{\sigma_V} = \frac{\mathbf{w}^T \mathbf{C} \mathbf{w}_A}{\sigma_V \sigma_A} \gamma_A \, ,$$

where in the last step we used Eq. 27.23. It is easy to show[3] that the factor in front of the characteristic portfolio's Sharpe Ratio is the correlation between the two portfolios. Therefore, the Sharpe Ratio of the investment in any portfolio V is given by the Sharpe Ratio of the characteristic portfolio V_A times the correlation between those two portfolios:

$$\gamma_V = \rho_{V,A} \gamma_A \text{ for all portfolios } V \, . \tag{27.25}$$

As enforced by constraint 27.21, the characteristic portfolio for attribute A has an excess return of 1, i.e., of 100%. Therefore it usually contains significant leverage. Let's now determine this leverage L_A (i.e., the degree of investment in risky assets) of the excess return's characteristic portfolio V_A. According to Eq. 27.17 this leverage is simply the exposure of portfolio V_A to the attribute $\mathbf{L}$

---

[3]From the definitions of correlation and covariance and from Eq. 26.16 we get

$$\mathbf{w}^T \mathbf{C} \mathbf{w}_A = \sum_{k,i=1}^{M} w_k w_{A,i} \underbrace{\delta t \, \mathrm{cov}(r_k, r_i)}_{C_{ki}} = \delta t \, \mathrm{cov}(r_V, r_A)$$

$$= \delta t \, \mathrm{corr}(r_V, r_A) \sqrt{\mathrm{var}(r_V)} \sqrt{\mathrm{var}(r_A)}$$

$$= \delta t \, \mathrm{corr}(r_V, r_A) \sqrt{\sigma_V / \delta t} \sqrt{\sigma_A / \delta t}$$

$$= \mathrm{corr}(r_V, r_A) \sigma_V \sigma_A \, ,$$

where we have used Eq. 26.18 in the penultimate step.

defined in Eq. 27.16:

$$\sum_{k=1}^{M} w_{A,k} = \mathbf{w}_A^T \mathbf{1} = \mathbf{w}_A^T \mathbf{L} \equiv L_A .$$

This can be determined in a very elegant way using Eq. 27.12 which relates the exposure of one characteristic portfolio to the attribute of *another* characteristic portfolio and vice versa. Using the attributes $\mathbf{L}$ as in Eq. 27.16 and A as in Eq. 27.19 with the corresponding two characteristic portfolios V_L and V_A, Eq. 27.12 reads $A_L \sigma_A^2 = L_A \sigma_L^2$ which directly yields

$$L_A = \frac{\sigma_A^2}{\sigma_L^2} A_L = \frac{\sigma_A^2}{\sigma_L^2} \underbrace{\mathbf{w}_L^T \widehat{\mathbf{R}}}_{\widehat{R}_L} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} , \tag{27.26}$$

where we have used Eq. 27.20 for the exposure A_L of portfolio V_L to attribute $A = \widehat{\mathbf{R}}$ and Eqs. 27.18 and 27.24 for the variances.

27.4 The Optimal Portfolio

Now, that we have calculated the leverage of portfolio V_A, it is very easy to construct a portfolio with leverage equal to one (i.e., a *fully invested* portfolio), by simply dividing all the weights $\mathbf{w}_A$ by L_A. Because of Eq. 27.13, the resulting portfolio is a characteristic portfolio for the attribute vector $\mathbf{m} := L_A \mathbf{A}$:

$$\mathbf{m} = L_A \mathbf{A} = \frac{\sigma_A^2}{\sigma_L^2} \widehat{R}_L \widehat{\mathbf{R}} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \widehat{\mathbf{R}} \tag{27.27}$$

$$\Longleftrightarrow \qquad m_k = \frac{\sigma_A^2}{\sigma_L^2} \left(R_k - r_f \right) \sum_{i=1}^{M} w_{L,i} \left(R_i - r_f \right) , \quad k = 1, \dots M .$$

According to Eq. 27.1, the exposure of *any* portfolio to this particular attribute is

$$m_V = \frac{\sigma_A^2}{\sigma_L^2} \widehat{R}_L \mathbf{w}^T \widehat{\mathbf{R}} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \mathbf{w}^T \widehat{\mathbf{R}} . \tag{27.28}$$

Let's now investigate the *characteristic* portfolio V_m for this attribute. Applying Eq. 27.13 directly yields the weights

$$\mathbf{w}_m = \frac{1}{L_A}\mathbf{w}_A = \frac{\sigma_L^2}{\sigma_A^2 \widehat{R}_L}\mathbf{w}_A \ . \tag{27.29}$$

Observe now that $\widehat{R}_L = \mathbf{w}_L^T\widehat{\mathbf{R}}$ is the excess return of the fully invested minimum risk portfolio (with weights given by Eq. 27.18). If we assume this excess return to be positive, i.e., if we assume Inequality 26.59 to hold,[4] then the holdings $\mathbf{w}_m$ in portfolio V_m are just the holdings $\mathbf{w}_A$, *all* multiplied by the *same positive*[5] constant $A_L\sigma_L^2/\sigma_A^2$. This positive constant is an example of the "positive number λ" mentioned below Eq. 27.23. Therefore, if Inequality 26.59 holds, this portfolio V_m lies on the capital market line and has the same Sharpe Ratio as portfolio V_A, i.e., *maximum* Sharpe Ratio.

We constructed the portfolio to be fully invested and indeed the leverage of this portfolio, i.e., its exposure to the attribute **L** as defined in Eq. 27.16 is:

$$L_m = \mathbf{w}_m^T\mathbf{L} = \frac{1}{L_A}\mathbf{w}_A^T\mathbf{L} = \frac{1}{L_A}L_A = 1 \ .$$

Therefore portfolio V_m with weights given by Eq. 27.29 is the fully invested portfolio with maximum Sharpe Ratio, i.e., the *optimal* portfolio. Indeed, inserting Eqs. 27.24 and 27.29 for $\mathbf{w}_A$ and $\mathbf{w}_L$ and observing Eq. 26.30, we immediately get

$$\mathbf{w}_m = \frac{\sigma_L^2}{\sigma_A^2}\frac{\mathbf{w}_A}{\mathbf{w}_L^T\widehat{\mathbf{R}}} = \frac{\sigma_L^2}{\sigma_A^2}\frac{\sigma_A^2\mathbf{C}^{-1}\widehat{\mathbf{R}}}{\sigma_L^2\mathbf{1}^T\mathbf{C}^{-1}\widehat{\mathbf{R}}} = \frac{\mathbf{C}^{-1}\widehat{\mathbf{R}}}{\mathbf{1}^T\mathbf{C}^{-1}\widehat{\mathbf{R}}} \tag{27.30}$$

in full accordance with Eq. 26.58. We have recovered the optimal portfolio using a cunning combination (given by Eq. 27.29) of two minimum risk portfolios: the minimum risk portfolio with leverage one and the minimum risk portfolio with excess return one!

---

[4]As already mentioned, this is not an unrealistic assumption: the expected return of the fully invested portfolio should by above the risk free rate as a compensation for the risk taken with that portfolio.

[5]The variances appearing in this term are positive numbers anyway.

With these weights, the characteristic portfolio for the attribute **m** has the following excess return, variance and betas

$$\widehat{R}_m = \widehat{\mathbf{R}}^T \mathbf{w}_m = \frac{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}$$

$$\sigma_m^2 = \mathbf{w}_m^T \mathbf{C} \mathbf{w}_m = \frac{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\left(\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}\right)^2} = \frac{\widehat{R}_m}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}$$

$$\boldsymbol{\beta}_m = \frac{\mathbf{C} \mathbf{w}_m}{\sigma_m^2} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \widehat{\mathbf{R}} = \frac{1}{\widehat{R}_m} \widehat{\mathbf{R}} \ . \qquad (27.31)$$

With these betas, the marginal risk contribution, Eq. 26.25, of each asset and the percentage of risk attributed to each asset, Eq. 26.26, are:

$$\frac{\partial \sigma_m}{\partial w_{m,i}} = \sigma_m \frac{\widehat{R}_i}{\widehat{R}_m} \ , \quad A_{m,i} = w_{m,i} \frac{\widehat{R}_i}{\widehat{R}_m} \quad \text{for} \quad i = 1, \ldots, M \ . \qquad (27.32)$$

Finally, using the results in Eq. 27.31, the Sharpe Ratio of the optimal portfolio can be written as

$$\gamma_m \equiv \frac{\widehat{R}_m}{\sigma_m} = \frac{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \frac{\sqrt{\left(\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}\right)^2}}{\sqrt{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}} = \frac{\left|\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}\right|}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \sqrt{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}$$

or

$$\gamma_m \equiv \frac{\widehat{R}_m}{\sigma_m} = \frac{\left|\widehat{R}_L\right|}{\widehat{R}_L} \sqrt{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} = \frac{\left|\widehat{R}_L\right|}{\widehat{R}_L} \frac{\sqrt{\widehat{R}_L \widehat{R}_m}}{\sigma_L} \ , \qquad (27.33)$$

where we have used $\widehat{R}_L$ as in Eq. 27.18. The last step follows from

$$\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} = \mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} \ \widehat{R}_m = \frac{\widehat{R}_L \widehat{R}_m}{\sigma_L^2} \ . \qquad (27.34)$$

We explicitly see here, that the sign of this Sharpe Ratio is the same as the sign of $\widehat{R}_L$.

We conclude this section by showing that Eq. 27.31 is in perfect agreement with the results of Sect. 26.2.3. It takes some algebra to see this, because in Sect. 26.2.3 we worked with "normal" returns instead of excess returns. First

note that

$$\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} = \left(\mathbf{R}^T - \mathbf{1}^T r_f \right) \mathbf{C}^{-1} \left(\mathbf{R} - \mathbf{1} r_f \right)$$
$$= \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - \underbrace{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{1}}_{R_{\min}/\sigma_{\min}^2} r_f - \underbrace{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{R}}_{R_{\min}/\sigma_{\min}^2} r_f + \underbrace{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}_{1/\sigma_{\min}^2} r_f^2 \ ,$$

where Eqs. 26.31 and 26.29 have been used. Thus

$$\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} = \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - 2 \frac{R_{\min}}{\sigma_{\min}^2} r_f + \frac{1}{\sigma_{\min}^2} r_f^2 \ . \tag{27.35}$$

With this "translation" between returns and *excess* returns we can now write the variance in Eq. 26.55 as

$$\sigma^2 = \sigma_{\min}^2 \left(1 + \sigma_{\min}^2 \frac{\mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2/\sigma_{\min}^2}{\left(R_{\min} - r_f \right)^2} \right)$$

$$= \frac{\sigma_{\min}^2}{\left(R_{\min} - r_f \right)^2} \left(\left(R_{\min} - r_f \right)^2 + \sigma_{\min}^2 \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} - R_{\min}^2 \right)$$

$$= \frac{\sigma_{\min}^2}{\widehat{R}_{\min}^2} \left(r_f^2 - 2 R_{\min} r_f + \sigma_{\min}^2 \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} \right)$$

$$= \frac{\sigma_{\min}^2}{\widehat{R}_{\min}^2} \left(r_f^2 - 2 R_{\min} r_f + \sigma_{\min}^2 \left(\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} + 2 \frac{R_{\min}}{\sigma_{\min}^2} r_f - \frac{1}{\sigma_{\min}^2} r_f^2 \right) \right)$$

$$= \frac{\sigma_{\min}^4}{\widehat{R}_{\min}^2} \widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} \ .$$

Observing now Eq. 26.68, we immediately see that this is indeed the same as σ_m^2 in Eq. 27.31.

27.5 The Efficient Frontier Revisited

It is worthwhile to re-do the optimization problem leading to the efficient frontier in the light of the above deliberations. Firstly, let's define efficient portfolios using *excess* return instead of normal returns. The *efficient portfolio* for any *given* excess return is defined as the fully invested minimum risk

portfolio with that excess return, i.e., the minimum risk portfolio satisfying constraint 26.4 and[6]

$$\mathbf{w}_e^T \widehat{\mathbf{R}} = \widehat{R} \Longleftrightarrow \mathbf{w}_e^T \left(\mathbf{R} - \mathbf{1}r_f \right) = R - r_f \qquad (27.36)$$

This constraint is of course exactly the same as Eq. 26.42 since we only added the constant $r_f = \mathbf{w}_e^T \mathbf{1}r_f$ on both sides. Thus it makes no difference whatsoever to define the efficient frontier using excess returns.

All portfolios on the efficient frontier are *fully invested* minimum risk portfolios. They only differ by different additional Constraints 27.36 (i.e., by different required R). We now want to construct those efficient portfolios using the two fully invested *characteristic* portfolios (which are of course also minimum risk portfolios) we introduced above.

Let's start by again solving the optimization problem equivalent to Eq. 26.43:

$$0 \stackrel{!}{=} \frac{\partial \mathcal{L}}{\partial \mathbf{w}_e^T} = \mathbf{C}\mathbf{w}_e - \lambda_1 \widehat{\mathbf{R}} - \lambda_2 \mathbf{1} \ .$$

The weights can easily be isolated:

$$\mathbf{w}_e = \lambda_1 \mathbf{C}^{-1} \widehat{\mathbf{R}} + \lambda_2 \mathbf{C}^{-1} \mathbf{1} \ , \qquad (27.37)$$

which is just Eq. 26.44 with "hats" above the returns.

Left-multiplying this by $\mathbf{1}^T$ and using constraint 26.4 yields

$$\lambda_2 = \frac{1}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} - \lambda_1 \frac{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} = \sigma_{\min}^2 - \lambda_1 \widehat{R}_{\min} \ .$$

Of course, $\widehat{R}_{\min}$ as in Eq. 26.68 and $\sigma_{\min}^2$ as in Eq. 26.29 are just the excess return and the variance of the fully invested minimum risk portfolio, i.e., of the characteristic portfolio V_L to the attribute $\mathbf{L}$, see Eq. 27.18:

$$\widehat{R}_{\min} = \widehat{R}_L \ , \quad \sigma_{\min}^2 = \sigma_L^2 \ . \qquad (27.38)$$

---

[6]To avoid confusion, from now on we will denote portfolios on the efficient frontier by an index e.

Inserting all this into constraint 27.36 yields

$$\mathbf{w}^T \widehat{\mathbf{R}} = \widehat{R}$$

$$\lambda_1 \widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} + \left[\sigma_L^2 - \lambda_1 \widehat{R}_L \right] \mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} = \widehat{R}$$

$$\lambda_1 \widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} + \widehat{R}_L - \lambda_1 \frac{\widehat{R}_L^2}{\sigma_L^2} = \widehat{R} \, ,$$

where in the last step we have used twice Eq. 26.68 in the form $\widehat{R}_L = \sigma_L^2 \mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}$. This is of course no surprise and we have gained nothing compared to Eq. 26.47 yet.

The decisive step now is to use the other fully invested characteristic portfolio at our disposal, namely V_m as defined in Sect. 27.4, to transform the term $\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}$ via Eq. 27.34. Then we solve for λ_1 to find

$$\lambda_1 = \frac{\sigma_L^2}{\widehat{R}_L} \cdot \frac{\widehat{R} - \widehat{R}_L}{\widehat{R}_m - \widehat{R}_L}$$

$$\implies \quad \lambda_2 = \sigma_L^2 \left[1 - \frac{\widehat{R} - \widehat{R}_L}{\widehat{R}_m - \widehat{R}_L} \right] \, .$$

We have now succeeded in expressing the Lagrange multipliers solely in terms of the required excess return $\widehat{R}$ and properties of the two characteristic portfolios V_L and V_m. Let's now also express the rest of Eq. 27.37 in terms of those two characteristic portfolios. The weights of V_L and V_m as given by Eqs. 27.30 and 27.18 can be written as

$$\mathbf{w}_L = \sigma_L^2 \mathbf{C}^{-1} \mathbf{1} \implies \mathbf{C}^{-1} \mathbf{1} = \frac{1}{\sigma_L^2} \mathbf{w}_L$$

$$\mathbf{w}_m = \frac{\sigma_L^2}{\widehat{R}_L} \mathbf{C}^{-1} \widehat{\mathbf{R}} \implies \mathbf{C}^{-1} \widehat{\mathbf{R}} = \frac{\widehat{R}_L}{\sigma_L^2} \mathbf{w}_m \, ,$$

where again Eq. 26.68 has been used. Inserting these expressions along with the above Lagrange multipliers into Eq. 27.37 we obtain

$$\mathbf{w}_e = \frac{\widehat{R} - \widehat{R}_L}{\widehat{R}_m - \widehat{R}_L} \mathbf{w}_m + \frac{\widehat{R}_m - \widehat{R}}{\widehat{R}_{\cdot m} - \widehat{R}_L} \mathbf{w}_L. \tag{27.39}$$

The asset weights $\mathbf{w}$ of any efficient frontier portfolio, regarded as functions of the required excess return $\widehat{R}$, are linear interpolations between the weights of the fully invested minimum risk portfolio and the optimal portfolio!

The exposure of an efficient portfolio to the characteristic $\mathbf{m}$ of the optimal portfolio is given by the ratio of the expected excess returns of the efficient and the optimal portfolio. This follows directly from Eqs. 27.27 and 27.31:

$$
m_e = \mathbf{w}_e^T \mathbf{m} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \underbrace{\mathbf{w}_e^T \widehat{\mathbf{R}}}_{\widehat{R}} = \frac{\widehat{R}}{\widehat{R}_m} \ .
$$

The variance of an efficient portfolio is by definition

$$
\begin{aligned}
\sigma_e^2 &= \mathbf{w}_e^T \mathbf{C} \mathbf{w}_e \\
&= \left(u \mathbf{w}_m^T + v \mathbf{w}_L^T \right) \mathbf{C} \left(u \mathbf{w}_m + v \mathbf{w}_L \right) \\
&= u^2 \sigma_m^2 + v^2 \sigma_L^2 + 2 u v \mathbf{w}_m^T \mathbf{C} \mathbf{w}_L
\end{aligned}
$$

where we have used the abbreviations

$$
u := \frac{\widehat{R} - \widehat{R}_L}{\widehat{R}_m - \widehat{R}_L} \ , \quad v := \frac{\widehat{R}_m - \widehat{R}}{\widehat{R}_m - \widehat{R}_L} \tag{27.40}
$$

to simplify the notation. With Eqs. 27.30 and 27.18 for the weights $\mathbf{w}_m$ and $\mathbf{w}_L$ we arrive at

$$
\begin{aligned}
\sigma_e^2 &= u^2 \sigma_m^2 + v^2 \sigma_L^2 + 2 u v \frac{\sigma_L^4}{\widehat{R}_L} \widehat{\mathbf{R}}^T \mathbf{C}^{-1} \mathbf{1} \\
&= u^2 \sigma_m^2 + v^2 \sigma_L^2 + 2 u v \sigma_L^2 \\
&= u^2 \sigma_m^2 + \left(v^2 + 2 u v + u^2 \right) \sigma_L^2 - u^2 \sigma_L^2 \\
&= u^2 \left(\sigma_m^2 - \sigma_L^2 \right) + \underbrace{(u + v)^2}_{1} \sigma_L^2 \ ,
\end{aligned}
$$

where in the second step Eq. 27.18 has been used for $\widehat{R}_L$. Thus, we arrive at the well known fact that the variance σ_e^2 of an efficient portfolio is a quadratic function of the portfolio return $\widehat{R}_e$, i.e., a parabola:

$$
\sigma_e^2 = \sigma_L^2 + \frac{\left(\sigma_m^2 - \sigma_L^2 \right)}{\left(\widehat{R}_m - \widehat{R}_L \right)^2} \left(\widehat{R} - \widehat{R}_L \right)^2 \ . \tag{27.41}
$$

Eq. 27.41 is exactly the same[7] as Eq. 26.49.

The corresponding expected return of an efficient portfolio in Eq. 26.50 can be written very intuitively. First note that exploiting Eq. 27.35 yields

$$\mathbf{R^T C}^{-1}\mathbf{R} - R_L^2/\sigma_L^2 = \widehat{\mathbf{R}}^T \mathbf{C}^{-1}\widehat{\mathbf{R}} + 2\frac{R_L}{\sigma_L^2}r_f - \frac{1}{\sigma_{\min}^2}r_f^2 - R_L^2/\sigma_L^2$$

$$= \widehat{\mathbf{R}}^T \mathbf{C}^{-1}\widehat{\mathbf{R}} - \frac{\left(R_L - r_f\right)^2}{\sigma_L^2}$$

$$= \gamma_m^2 - \gamma_L^2 \,,$$

where Eq. 27.33 has been used. Thus, the expected return can be written as

$$R_e = R_L \pm \sqrt{\left(\sigma_e^2 - \sigma_L^2\right)\left(\gamma_m^2 - \gamma_L^2\right)} \,.$$

Asymptotically for $\sigma_e \longrightarrow \infty$ this approaches two straight lines.

$$R_e \sim \pm\sigma_e\sqrt{\gamma_m^2 - \gamma_L^2} \,.$$

---

[7]To see this, we have to show that the inverse prefactor of $(\widehat{R} - \widehat{R}_L)^2$ is the same as the denominator in Eq. 26.49. Using Eqs. 27.30 and 27.18 we find

$$\frac{\left(\widehat{R}_m - \widehat{R}_L\right)^2}{\sigma_m^2 - \sigma_L^2} = \frac{\left(\widehat{R}_m - \sigma_L^2 \mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}\right)^2}{\frac{\widehat{R}_m}{\mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}} - \sigma_L^2} = \frac{\left(\widehat{R}_m - \sigma_L^2 \mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}\right)^2}{\left(\widehat{R}_m - \sigma_L^2 \mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}\right) / \left(\mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}\right)}$$

$$= \frac{\widehat{R}_m - \sigma_L^2 \mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}}{\frac{1}{\mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}}} = \widehat{R}_m \left(\mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}\right) - \sigma_L^2 \left(\mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}\right)^2$$

$$= \widehat{\mathbf{R}}^T \mathbf{C}^{-1}\widehat{\mathbf{R}} - \frac{\widehat{R}_L^2}{\sigma_L^2} \,.$$

Now we switch back to "normal" returns instead of excess returns with the help of Eq. 27.35

$$\frac{\left(\widehat{R}_m - \widehat{R}_L\right)^2}{\sigma_m^2 - \sigma_L^2} = \mathbf{R}^T \mathbf{C}^{-1}\mathbf{R} - 2\frac{R_L}{\sigma_L^2}r_f + \frac{1}{\sigma_L^2}r_f^2 - \frac{\widehat{R}_L^2}{\sigma_L^2}$$

$$= \mathbf{R}^T \mathbf{C}^{-1}\mathbf{R} + \frac{1}{\sigma_L^2}\left(-2R_L r_f + r_f^2 - \left(R_L - r_f\right)^2\right)$$

$$= \mathbf{R}^T \mathbf{C}^{-1}\mathbf{R} - \frac{R_L^2}{\sigma_L^2} \,.$$

This is just the denominator in Eq. 26.49 (the index "L" here is the same as the Index "min" in Sect. 26.2.2).

The (absolute value of the) slope of these lines is less then the slope of the capital market line as long as the minimal risk portfolio has an expected return not equal to the risk free rate.

Each efficient frontier portfolio is itself a characteristic portfolio. According to Eq. 27.14 the corresponding attribute $\mathbf{e}$ is a linear combination of the attributes $\mathbf{L}$ and $\mathbf{m}$.

$$\mathbf{e} = x\mathbf{L} + y\mathbf{m} \ . \tag{27.42}$$

Comparing Eq. 27.14 with Eq. 27.39 yields

$$x = \frac{\sigma_L^2}{\sigma_e^2} \frac{\widehat{R}_m - \widehat{R}}{\widehat{R}_m - \widehat{R}_L} = \frac{\sigma_L^2}{\sigma_e^2} v$$

$$y = \frac{\sigma_m^2}{\sigma_e^2} \frac{\widehat{R} - \widehat{R}_L}{\widehat{R}_m - \widehat{R}_L} = \frac{\sigma_m^2}{\sigma_e^2} u \ .$$

Inserting these we can write for the attribute $\mathbf{e}$:

$$\sigma_e^2 \mathbf{e} = v\sigma_L^2 \mathbf{L} + u\sigma_m^2 \mathbf{m}$$

$$= v\sigma_L^2 \mathbf{L} + u\sigma_m^2 \frac{\sigma_A^2}{\sigma_L^2} \widehat{R}_L \widehat{\boldsymbol{R}}$$

$$= v\frac{1}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \mathbf{L} + u\frac{1}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \widehat{\boldsymbol{R}} \ .$$

And finally, using Eq. 27.18, the attribute for which an efficient portfolio is the characteristic portfolio becomes

$$\mathbf{e} = \frac{\sigma_L^2}{\sigma_e^2} \left(\frac{\widehat{R}_m - \widehat{R}}{\widehat{R}_m - \widehat{R}_L} \mathbf{1} + \frac{\widehat{R} - \widehat{R}_L}{\widehat{R}_m - \widehat{R}_L} \frac{\widehat{\boldsymbol{R}}}{\widehat{R}_L} \right) \ . \tag{27.43}$$

Because of Eq. 27.9 the asset betas with respect to the efficient frontier portfolio are given by these attributes. The marginal risk contributions, Eq. 26.25, of the assets to the efficient frontier portfolio are

$$\frac{\partial \sigma_e}{\partial \mathbf{w}} = \sigma_L^2 \left(\frac{\widehat{R}_m - \widehat{R}}{\widehat{R}_m - \widehat{R}_L} \mathbf{1} + \frac{\widehat{R} - \widehat{R}_L}{\widehat{R}_m - \widehat{R}_L} \frac{\widehat{\boldsymbol{R}}}{\widehat{R}_L} \right)$$

and the percentage of risk attributed to each asset, Eq. 26.26, is $A_{e,i} = w_{e,i} \, e_i$ for $i = 1, \ldots, M$.

28

Active Management and Benchmarking

28.1 The Capital Asset Pricing Model

One of the most important results of the previous chapter is the last line in Eq. 27.31. It can be written in the form

$$\widehat{\mathbf{R}} = \boldsymbol{\beta}_m \widehat{R}_m \ . \tag{28.1}$$

Based on this equation, the expected excess returns of all assets are *explained* solely by their betas with respect to the optimal portfolio and the return of the optimal portfolio itself. This is very remarkable. Within the framework of classical Mean-Variance portfolio theory there is not much freedom for the individual assets. The optimal portfolio drives everything. The question now remains, what this optimal portfolio exactly is. The following argument leads to an identification[1]:

If all investors have the same information on all assets (in particular concerning expected returns, variances and covariances) and if all investors behave optimally (in the sense that they invest in such a way that their mean/variance-ratio is maximal), then every investment will lie on the capital market line. If Inequality 26.59 holds, than each investor will invest in a mixture of the optimal portfolio V_m and the risk free money market account. Investments with a variance lower than σ_m are long in the risk free investment while investments with a variance higher than σ_m are leveraged, i.e., they

---

[1]Under the assumption that there are no transaction costs, taxes or other "friction".

borrow from the money market. Adding up all investments of all market participants, the total amount borrowed must of course equal the total amount of lending, i.e., the grand total money market position of all market participants must equal zero. Therefore, the grand total of all investments of all market participants is the optimal portfolio. On the other hand, the grand total of all investments of all market participants is "the market" itself. Thus, **the optimal portfolio is the market itself**, which is why it is also called the *market portfolio*.

One can also look at this from a different angle: you can always do a linear regression for the (historical) asset excess returns $\widehat{r}_i$ on (historical) market excess returns, i.e.

$$\widehat{r}_i = \alpha_{i,m} + \beta_{i,m}\widehat{r}_m + \epsilon_i \text{ for } i = 1, \ldots, M , \qquad (28.2)$$

where the ϵ_i are the error terms which by definition have zero expectation and are uncorrelated with $\widehat{r}_m$. Calculating expected values and comparing this with Eq. 28.1, we immediately see that the CAPM requires the *alpha* of each asset to be zero. In the CAPM, no additional return is expected for any asset above that produced by the asset's beta and the market!

The same holds for any *portfolio* of assets, since in the regression everything is linear.

$$\underbrace{\sum_i w_i\widehat{r}_i}_{\widehat{r}_V} = \underbrace{\sum_i w_i\alpha_{i,m}}_{\alpha_{V,m}} + \underbrace{\sum_i w_i\beta_{i,m}\widehat{r}_m}_{\beta_{V,m}} + \underbrace{\sum_i w_i\epsilon_i}_{\epsilon_V} ,$$

$$\widehat{r}_V = \alpha_{V,m} + \beta_{V,m}\widehat{r}_m + \epsilon_V . \qquad (28.3)$$

Since the $\alpha_{i,m}$ should all be zero, so must $\alpha_{V,m}$. Thus, **in the CAPM, the alpha of any portfolio is zero**. No portfolio can beat the market in the long run. This is not a new insight really, but in fact trivial: We *constructed* the optimal portfolio to have the best possible mean/variance ratio. Then we argued that the optimal portfolio should have the same position weights as the market. Thus, the market has the best possible mean/variance-ratio. Therefore, no other portfolio has a better mean/variance ratio because if it did we would have taken this portfolio as the optimal portfolio.

The "market" contains everything one could possibly invest in: equity, debt, commodities, real estate, etc., even art. Most often, this is not a feasible concept for practical situations, so instead, "the market" is divided into several (sub-)markets like the German Bond Market, the US Mid Cap Market or

the Japanese Real Estate Market, to name a few of many hundreds or even thousands. But even these smaller markets are usually too large to handle, so *Indices* are defined as representatives of those markets. Examples of such Indices are the S&P 500, the German DAX, the Japanese Nikkei, etc. They contain a limited number of assets (with weights defined and calculated by the respective index provider) and serve as a proxy of the market they are supposed to represent. Indices are often used as *benchmarks* to measure the performance of a portfolio. In the following we will use the subscript B for the benchmark. The benchmark portfolio V_B is supposed to be a proxy for the market portfolio V_m (of whatever market we are interested in), but it is usually *not* identical to the market portfolio. In the following, we will carefully distinguish between benchmark and market portfolio properties using subscripts B and m, respectively.

28.2 Theory of Efficient Markets

At the heart of *active portfolio management* lies the hope of a portfolio manager that he can outperform the market, despite the fact that the CAPM says this is impossible. That does not necessarily mean, however, that the portfolio manager does not believe in the CAPM. He rather doubts the assumptions of the CAPM, in particular the one that everybody has the same information. The information edge the manager may believe to have on some assets causes him to cast asset return expectations that differ from the "consensus returns" of Eq. 28.1. This then results in different position weights, even if the manager uses the same mean/variance optimization as in the CAPM. If his return estimates turn out right he will indeed outperform the market. However, efficient market theory (at least in its strong form) states that this is still not possible in the long run. *Efficient market theory* comes in three varieties:

- The *weak* efficient market theory states: Investors cannot outperform the market using historical price and volume data, only.
- The *semi strong* efficient market theory states: Investors cannot outperform the market using publicly available information, only (i.e., historical price and volume data, fundamental data, recommendations published by analysts, etc.)
- The *strong* efficient market theory states: Investors can never outperform the market. Market prices contain all relevant information.

A couple of observations support the view that at most the weak efficient market theory is in fact valid in the real world. For example, the long-term success of some value investors apparently excludes the semi-strong and strong theory. Also, not all market participants have equal access to all available informations, otherwise insider trading and manipulation of market indices (like in the Libor scandal) wouldn't be possible. In addition, we know from behavioural finance that even professional asset managers oft act quite irrational. Finally, institutional investors are bound to obey regulatory requirements, which sometimes stands in the way of rational decisions.

28.3 Benchmarking Against an Index

Nonetheless the main goal of so-called *active* asset management is "to beat the market", i.e., to gain a positive alpha against a benchmark (in most cases an index) which represents the market to be beaten. To assess how well an asset has been doing compared to the *benchmark* one can do a regression of the historical asset returns against the historical benchmark returns as in Eq. 28.2.

$$\widehat{r}_i = \alpha_{i,B} + \beta_{i,B}\widehat{r}_B + \epsilon_i \text{ for } i = 1, \dots, M \ .$$

Taking the expectation (e.g., as a historical average) we get the relation between asset and benchmark returns that serves as the definition of $\boldsymbol{\alpha}_B$, the vector of the assets alphas with respect to the benchmark:

$$\widehat{\mathbf{R}} = \boldsymbol{\alpha}_B + \boldsymbol{\beta}_B \widehat{R}_B \ . \tag{28.4}$$

If the benchmark perfectly represented the market and if the CAPM were perfectly true, this would collapse to Eq. 28.1, i.e., the assets alphas would all vanish.

The asset betas with respect to the Benchmark are given by the matrix $\mathbf{C}$ and the position weights $\mathbf{w}_B$ of the benchmark portfolio as in Eq. 26.23.

$$\boldsymbol{\beta}_B = \frac{1}{\sigma_B^2}\mathbf{C}\mathbf{w}_B = \frac{\mathbf{C}\mathbf{w}_B}{\mathbf{w}_B^T\mathbf{C}\mathbf{w}_B} \ . \tag{28.5}$$

Equations 28.4 and 28.5 directly imply[2] that the following combination of $\boldsymbol{\alpha}_B$, $\boldsymbol{\beta}_B$ and $\mathbf{C}$ is be zero:

$$\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B = 0 \ . \tag{28.6}$$

The same considerations that lead us to Eq. 28.4 also work for the *portfolio*'s expected excess return

$$\widehat{R}_V = \mathbf{w}^T \widehat{\mathbf{R}} = \underbrace{\mathbf{w}^T \boldsymbol{\alpha}_B}_{\alpha_{V,B}} + \underbrace{\mathbf{w}^T \boldsymbol{\beta}_B}_{\beta_{V,B}} \widehat{R}_B \ . \tag{28.7}$$

Here, the portfolio beta[3] $\beta_{V,B}$ with respect to the benchmark follows from the regression analysis in the usual way[4] given by Eq. A.20:

$$\beta_{V,B} \equiv \frac{\mathrm{cov}(r_V, r_B)}{\mathrm{Var}(r_B)} = \frac{\mathrm{cov}(\widehat{r}_V, \widehat{r}_B)}{\mathrm{Var}(\widehat{r}_B)} \ . \tag{28.8}$$

The portfolio alpha $\alpha_{V,B}$ can also be determined by such an regression of the historical portfolio excess returns $\widehat{r}_V$ against the historical benchmark excess returns $\widehat{r}_B$. In this way one gets an *ex post* estimation of how the portfolio did in the past. The historical portfolio returns are of course not only a result of market movements but also of trading, i.e., position changes done by the portfolio manager. Therefore, for the estimation of the impact of portfolio

---

[2]Right-multiplying the transpose of Eq. 28.4 by $\mathbf{C}^{-1}$ yields

$$\widehat{\mathbf{R}}^T \mathbf{C}^{-1} = \boldsymbol{\alpha}_B^T \mathbf{C}^{-1} + \widehat{R}_B \boldsymbol{\beta}_B^T \mathbf{C}^{-1} \ .$$

Right-multiplying this by $\boldsymbol{\beta}_B$ and then inserting Eq. 28.5 results in

$$\begin{aligned}
\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B &= \widehat{\mathbf{R}}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B - \widehat{R}_B \boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B \\
&= \frac{1}{\sigma_B^2} \widehat{\mathbf{R}}^T \mathbf{C}^{-1} \mathbf{C} \mathbf{w}_B - \widehat{R}_B \frac{1}{\sigma_B^2} \mathbf{w}_B^T \mathbf{C} \mathbf{C}^{-1} \frac{1}{\sigma_B^2} \mathbf{C} \mathbf{w}_B \\
&= \frac{1}{\sigma_B^2} \underbrace{\mathbf{w}_B^T \widehat{\mathbf{R}}}_{\widehat{R}_B} - \frac{1}{\sigma_B^4} \widehat{R}_B \underbrace{\mathbf{w}_B^T \mathbf{C} \mathbf{w}_B}_{\sigma_B^2} = 0 \ .
\end{aligned}$$

[3]The Beta appearing here is in terms of returns, not in terms of relative price changes as for instance in Eq. 26.23. For Beta this does not make any difference, since the factors δt appearing due to Eq. 21.27 are the same in numerator and denominator and therefore cancel out.
[4]Covariances and variances do not change upon adding a constant r_f. Therefore, working with returns or with *excess* returns doesn't make any difference.

changes, it is important to get *ex ante* estimates for alpha and beta over the *next* holding period of the *current* portfolio with its *current* holdings. To get these, we take the *historical* information of the *assets* and the *current* position weights in the *portfolio* to produce ex ante estimates of portfolio properties, like for instance alpha and beta. This can be done by using the *historical* alphas and betas of the *assets* with respect to the benchmark together with the *current weights* of those assets in the portfolio:

$$\alpha_{V,B} = \sum_{i=1}^{M} w_i \alpha_{i,B} = \mathbf{w}^T \boldsymbol{\alpha}_B \ , \quad \beta_{V,B} = \sum_{i=1}^{M} w_i \beta_{i,B} = \mathbf{w}^T \boldsymbol{\beta}_B \ . \quad (28.9)$$

The historical alphas and betas of the assets are determined via the regressions given in Eq. 28.2. One might also use other estimates for the asset alphas and betas in this ex ante estimation of the portfolio alpha and beta (e.g., an asset manager's personal opinion).

In the context of active portfolio management, not only alpha and beta but usually *all* quantities (return, risk, etc.) are defined *relative to the benchmark*: if the portfolio loses money but the benchmark performs even worse, the portfolio manager is still happy in this relative framework (this is not necessarily true for the investor). This is in stark contrast to the previous sections, where we considered the excess return and the risk of the portfolio *itself*, an approach sometimes called *Total Return Management*. Still, all of the above considerations in the Total Return framework can easily be recovered within the benchmark-relative framework by simply defining the benchmark to be the money market account.

There are two different ways commonly used by market participants to define everything relative to the benchmark: using *active* quantities and *residual* quantities. Let's start with the active quantities.

28.3.1   Active Portfolio Properties

The *active position* is defined as the deviation of the portfolio weights from the benchmark weights

$$\overline{\mathbf{w}} = \mathbf{w} - \mathbf{w}_B \ . \quad (28.10)$$

Similarly, the difference between the cash positions of the portfolio and the benchmark is called the *active cash.* From Eq. 26.64, the active cash is simply

$$\text{Active Cash} = \left(1 - \mathbf{w}^T \mathbf{1}\right) - \left(1 - \mathbf{w}_B^T \mathbf{1}\right) = -\overline{\mathbf{w}}^T \mathbf{1}$$

The *active return* is defined as the difference between the portfolio return, Eq. 28.3, and the benchmark return.

$$\overline{r}_V := \overline{\mathbf{w}}^T \widehat{\mathbf{r}} = \widehat{r}_V - \widehat{r}_B = \alpha_{V,B} + (\beta_{V,B} - 1)\widehat{r}_B + \epsilon_V \ .$$

The *expected* active return is defined as the expected return of the portfolio with weights given by the active position

$$
\begin{aligned}
\overline{R}_V = \mathrm{E}\,[\overline{r}_V] &= \overline{\mathbf{w}}^T \widehat{\mathbf{R}} \\
&= \mathbf{w}^T \boldsymbol{\alpha}_B + (\mathbf{w}^T \boldsymbol{\beta}_B - 1)\widehat{R}_B \\
&= \alpha_{V,B} + (\beta_{V,B} - 1)\widehat{R}_B
\end{aligned}
\tag{28.11}
$$

and the *active risk* is defined in terms of the variance of active returns:

$$\overline{\sigma}_V^2 := \delta t \ \mathrm{var}\,[\overline{r}_V] = \delta t \ \mathrm{var}\,[\epsilon_V] + (\beta_{V,B} - 1)^2 \sigma_B^2 \ . \tag{28.12}$$

The last equation holds since the errors ϵ_V are by construction uncorrelated all other parameters. The term $(\beta_{V,B} - 1)$ appearing here is called *active beta.* The active risk is also called *Tracking Error* since it is a measure of how well the portfolio tracks the benchmark. We can of course also write the tracking error in the form of Eq. 26.17 as the risk of a portfolio with position weights $\overline{\mathbf{w}}$:

$$\overline{\sigma}_V^2 = \delta t \ \mathrm{var}\,[\overline{r}_V] = \delta t \ \mathrm{var}\left[\overline{\mathbf{w}}^T \widehat{\mathbf{r}}\right] = \overline{\mathbf{w}}^T \mathbf{C} \overline{\mathbf{w}} \ .$$

The same holds for the expected active return in Eq. 28.11 when compared to the expected portfolio return in Eq. 26.3. Indeed, many concepts introduced in the previous sections carry over to the benchmark-relative framework by simply replacing the weights $\mathbf{w}$ by the active weights $\overline{\mathbf{w}}$. For instance, in analogy to Eq. 26.19 we can define the *marginal active risk* as

$$\frac{\partial \overline{\sigma}_V}{\partial \overline{\mathbf{w}}} = \frac{\mathbf{C}\overline{\mathbf{w}}}{\sqrt{\overline{\mathbf{w}}^T \mathbf{C}\overline{\mathbf{w}}}} = \frac{\mathbf{C}\overline{\mathbf{w}}}{\sqrt{\delta t \ \mathrm{var}\,[\epsilon_V] + (\beta_{V,B} - 1)^2 \sigma_B^2}}$$

and attribute the active risk to the individual positions as in Eqs. 26.20 or 26.21.

$$\overline{\sigma}_V = \overline{\mathbf{w}}^T \frac{\partial \overline{\sigma}_V}{\partial \overline{\mathbf{w}}} = \sum_{i=1}^{M} \overline{w}_i \frac{\partial \overline{\sigma}_V}{\partial \overline{w}_i} \ .$$

The amount of active risk attributed to the ith asset is thus

$$\overline{w}_i \frac{\partial \overline{\sigma}_V}{\partial \overline{w}_i} = \frac{\overline{w}_i^2 \sum_{k=1}^{M} C_{ki}}{\sqrt{\delta t^2 \, \mathrm{var}\,[\epsilon_V] + (\beta_{V,B} - 1)^2 \sigma_B^2}} \quad \text{for } i = 1, \dots, M \ .$$

The percentage of active risk attributed to the ith asset is this number divided by $\overline{\sigma}_V$.

28.3.2   Residual Portfolio Properties

While the above *active* quantities focus on the differences between the portfolio and the benchmark itself, the *residual* quantities focus on difference between portfolio properties and the properties implied by the CAPM. To be specific, the *residual (excess) return* is defined as the difference between the portfolio excess return and its excess return implied by the CAPM via Eq. 28.1.

$$\widetilde{r}_V := \widehat{r}_V - \beta_{V,B}\widehat{r}_B = \alpha_{V,B} + \epsilon_V \ . \qquad (28.13)$$

Correspondingly, the *residual risk,* is defined as the volatility of the residual return:

$$\widetilde{\sigma}_V^2 := \delta t \, \mathrm{var}\,[\widetilde{r}_V] = \delta t \, \mathrm{var}\,[\epsilon_V] \ . \qquad (28.14)$$

According to Eq. 28.4, the *expectation* of the residual return is equal to the portfolio's *alpha*.

$$\widetilde{R}_V = \mathrm{E}\,[\widetilde{r}_V] = \mathbf{w}^T \boldsymbol{\alpha}_B = \alpha_{V,B} \ . \qquad (28.15)$$

The *Information Ratio* is defined as the expected residual (excess) return per residual risk and is therefore the benchmark-relative analogue to the Sharpe

Ratio, Eq. 27.22.

$$\widetilde{\gamma}_V \equiv \frac{\widetilde{R}_V}{\widetilde{\sigma}_V} = \frac{\alpha_{V,B}}{\sqrt{\delta t \operatorname{var}[\epsilon_V]}} \; . \tag{28.16}$$

The residual position weights producing the above residual return and residual risk are

$$\widetilde{\mathbf{w}} = \mathbf{w} - \beta_{V,B} \mathbf{w}_B \; . \tag{28.17}$$

With these weights, we can write

$$\widetilde{R}_V = \widetilde{\mathbf{w}}^T \widehat{\mathbf{R}} \quad \text{and} \quad \widetilde{\sigma}_V^2 = \widetilde{\mathbf{w}}^T \mathbf{C} \widetilde{\mathbf{w}} \; . \tag{28.18}$$

and define the *marginal* residual risk in analogy to Eq. 26.19 as

$$\frac{\partial \widetilde{\sigma}_V}{\partial \widetilde{\mathbf{w}}} = \frac{\mathbf{C} \widetilde{\mathbf{w}}}{\sqrt{\widetilde{\mathbf{w}}^T \mathbf{C} \widetilde{\mathbf{w}}}} = \frac{\mathbf{C} \widetilde{\mathbf{w}}}{\sqrt{\delta t \operatorname{var}[\epsilon_V]}} \; . \tag{28.19}$$

In analogy to Eq. 26.20, the residual risk attributed to the ith position is

$$\widetilde{w}_i \frac{\partial \widetilde{\sigma}_V}{\partial \widetilde{w}_i} = \frac{\widetilde{w}_i^2 \sum_{k=1}^M C_{ki}}{\sqrt{\delta t \operatorname{var}[\epsilon_V]}} \quad \text{for} \quad i = 1, \dots, M \; . \tag{28.20}$$

The percentage $\widetilde{A}_i$ of residual risk attributed to the ith asset is this number divided by $\widetilde{\sigma}_V$.

28.4 Benchmark and Characteristic Portfolios

We have shown that attributes and characteristic portfolios prove to be very useful tools in the "total return" framework. Motivated by this we will now define some attributes and their characteristic portfolios in the benchmark-relative framework. In Chap. 27 we analyzed the properties (i.e., leverage, return, variance, etc.) of the characteristic portfolio for any arbitrary vector **a** of asset attributes. We will now add to this list the characteristic portfolio's *residual* properties as well as its alpha and beta with respect to the benchmark.

The ex ante estimators for alpha and beta are given by Eq. 28.9 with the vectors of *asset* alphas and betas as in Eq. 28.4. With the weights of the characteristic portfolio given by Eq. 27.4, the characteristic portfolio's alpha and beta with respect to the benchmark are

$$\alpha_{a,B} \equiv \mathbf{w}_a^T \boldsymbol{\alpha}_B = \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}}$$

$$\beta_{a,B} \equiv \mathbf{w}_a^T \boldsymbol{\beta}_B = \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \ . \tag{28.21}$$

Let's look now at the residual properties of the characteristic portfolio. With Eq. 27.4, the residual weights defined in Eq. 28.17 are

$$\begin{aligned}
\widetilde{\mathbf{w}}_a &= \mathbf{w}_a - \beta_{a,B} \mathbf{w}_B \\
&= \frac{\mathbf{C}^{-1} \mathbf{a}}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} - \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \mathbf{w}_B \\
&= \frac{\mathbf{C}^{-1} \mathbf{a}}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} - \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}} \frac{\mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} ,
\end{aligned} \tag{28.22}$$

where in the last step we have used the implied benchmark weights which will be introduced below in Eq. 28.28. According to Eq. 28.18, residual return and residual risk are then

$$\widetilde{R}_a \equiv \widetilde{\mathbf{w}}^T \widehat{\mathbf{R}} = \widehat{R}_a - \sigma_a^2 \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} \boldsymbol{\beta}_B^T \mathbf{C}^{-1} \widehat{\mathbf{R}},$$

$$\widetilde{\sigma}_a^2 \equiv \widetilde{\mathbf{w}}^T \mathbf{C} \widetilde{\mathbf{w}} = \sigma_a^2 \left[1 - \frac{\left(\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B \right)^2}{\left(\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B \right) \left(\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a} \right)} \right] , \tag{28.23}$$

where we have used Eqs. 27.6 and 27.7. Note that the Cauchy-Schwarz inequality, Eq. 26.37, implies that $\widetilde{\sigma}_a^2$ is always larger than zero except for those attributes which are multiples[5] of $\boldsymbol{\beta}_B$.

With the above results, Information Ratio and marginal residual risk contributions can easily be calculated from their respective Definitions 28.16

---

[5]For such attributes we have $\widetilde{\sigma}_a^2 = 0$. Thus, the characteristic portfolios of such attributes have no residual risk.

and 28.19 for any characteristic portfolio.

$$\widetilde{\gamma}_a = \frac{\widetilde{R}_a}{\widetilde{\sigma}_a} = \frac{\sigma_a^2}{\widetilde{\sigma}_a} \mathbf{a}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} - \frac{\sigma_a^2}{\widetilde{\sigma}_a} \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} \boldsymbol{\beta}_B \mathbf{C}^{-1} \widehat{\mathbf{R}}$$

$$\frac{\partial \widetilde{\sigma}_a}{\partial \widetilde{\mathbf{w}}} = \frac{\mathbf{C}\widetilde{\mathbf{w}}}{\widetilde{\sigma}_a} = \frac{\sigma_a^2}{\widetilde{\sigma}_a} \mathbf{a} - \frac{\sigma_a^2}{\widetilde{\sigma}_a} \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} \boldsymbol{\beta}_B \ . \tag{28.24}$$

The residual risk attributions follow directly from definition 28.20.

28.4.1   The Fully Invested Minimal Risk Portfolio

As a first example, we will apply the above results to the characteristic portfolio for the attribute $\mathbf{L} = \mathbf{1}$ introduced in Sect. 27.2, i.e., for the fully invested minimal risk portfolio. Its alpha and beta with respect to the benchmark are

$$\alpha_{L,B} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \ , \qquad \beta_{L,B} \frac{\mathbf{1}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \ .$$

Its residual weights are

$$\widetilde{\mathbf{w}}_L = \frac{\mathbf{C}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} - \frac{\mathbf{1}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \frac{\mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} \ .$$

and the portfolio's residual properties are

$$\widetilde{R}_a = \widehat{R}_L - \sigma_L^2 \frac{\mathbf{a}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} \boldsymbol{\beta}_B^T \mathbf{C}^{-1} \widehat{\mathbf{R}}$$

$$\widetilde{\sigma}_L^2 = \sigma_L^2 \left[1 - \frac{\left(\mathbf{1}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B \right)^2}{\left(\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B \right) \left(\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \right)} \right] \ .$$

Information Ratio and marginal residual risk contributions are

$$\widetilde{\gamma}_L = \frac{\sigma_L^2}{\widetilde{\sigma}_L} \mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}} - \frac{\sigma_L^2}{\widetilde{\sigma}_L} \frac{\mathbf{1}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} \boldsymbol{\beta}_B \mathbf{C}^{-1} \widehat{\mathbf{R}}$$

$$\frac{\partial \widetilde{\sigma}_a}{\partial \widetilde{\mathbf{w}}} = \frac{\sigma_L^2}{\widetilde{\sigma}_L} \mathbf{1} - \frac{\sigma_L^2}{\widetilde{\sigma}_L} \frac{\mathbf{1}^T \mathbf{C}^{-1} \boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \boldsymbol{\beta}_B} \boldsymbol{\beta}_B \ .$$

As usual, the residual risk attributions now follow directly from definition 28.20.

We will postpone the discussion of the other two important characteristic portfolios already introduced above, namely the portfolio for the attribute $\mathbf{A} = \widehat{\mathbf{R}}$ of Sect. 27.3 and the market portfolio of Sect. 27.4 until we have introduced two new portfolios, namely the characteristic portfolios for alpha and beta.

28.4.2   The Characteristic Portfolio for Beta

Let the attribute vector $\mathbf{b}$ be the beta of the assets with respect to the benchmark as in Eq. 28.4

$$
b_i := \beta_{i,B} \equiv \frac{\mathrm{cov}(\widehat{r}_i, \widehat{r}_B)}{\mathrm{var}(\widehat{r}_B)} = \frac{\delta t}{\sigma_B^2} \sum_{k=1}^{M} w_{B,k} \underbrace{\mathrm{cov}(\widehat{r}_i, \widehat{r}_k)}_{C_{ik}/\delta t}
$$

or in vector notation

$$
\mathbf{b} = \boldsymbol{\beta}_B = \frac{\mathbf{C}\mathbf{w}_B}{\mathbf{w}_B^T \mathbf{C}\mathbf{w}_B} \tag{28.25}
$$

According to Eq. 27.1, the exposure of *any* portfolio V to this particular attribute is

$$
b_V = \mathbf{w}^T \mathbf{b} = \frac{\mathbf{w}^T \mathbf{C}\mathbf{w}_B}{\mathbf{w}_B^T \mathbf{C}\mathbf{w}_B} \; .
$$

Comparing this with the *portfolio* beta $\beta_{V,B}$ with respect to the benchmark, i.e., with

$$
\beta_{V,B} \equiv \frac{\mathrm{cov}(\widehat{r}_V, \widehat{r}_B)}{\mathrm{var}(\widehat{r}_B)} = \frac{\delta t}{\sigma_B^2} \sum_{k=1}^{M} w_{B,k} \, \mathrm{cov}(\widehat{r}_V, \widehat{r}_k)
$$

$$
= \frac{\delta t}{\sigma_B^2} \sum_{k=1}^{M} w_{B,k} \sum_{i=1}^{M} w_i \underbrace{\mathrm{cov}(\widehat{r}_i, \widehat{r}_k)}_{C_{ik}/\delta t} = \frac{\mathbf{w}^T \mathbf{C}\mathbf{w}_B}{\mathbf{w}_B^T \mathbf{C}\mathbf{w}_B} \tag{28.26}
$$

we find that the exposure of any portfolio V to the attribute $\mathbf{b}$ is equal to the portfolio beta with respect to the benchmark:

$$b_V = \mathbf{w}^T \mathbf{b} = \beta_{V,B} \ .$$

Let's investigate now the *characteristic* portfolio V_b for this particular attribute, i.e., the minimum risk portfolio with unit exposure $b_V = 1$. According to Eqs. 27.4 and 28.25, this portfolio has the following weights

$$\mathbf{w}_b = \frac{\mathbf{C}^{-1}\mathbf{b}}{\mathbf{b}^T \mathbf{C}^{-1}\mathbf{b}} = \left(\mathbf{w}_B^T \mathbf{C} \mathbf{w}_B \right) \frac{\mathbf{C}^{-1} (\mathbf{C}\mathbf{w}_B)}{\left(\mathbf{w}_B^T \mathbf{C} \right) \mathbf{C}^{-1} (\mathbf{C}\mathbf{w}_B)} = \mathbf{w}_B \ .$$

Thus, the characteristic portfolio for the attribute "beta to a benchmark" is the benchmark itself:

$$\mathbf{w}_b = \mathbf{w}_B \ . \tag{28.27}$$

Therefore, among all portfolios with $\beta_{V,B} = 1$, the benchmark itself has minimum risk.

Equation 28.27 shows a way how to imply the weights of the individual assets in the benchmark, if those weights are not known to the investors: From historical time series determine the asset betas with respect to the benchmark and the asset's covariances. Then the *implied benchmark weights* are given by

$$\mathbf{w}_B = \frac{\mathbf{C}^{-1}\boldsymbol{\beta}_B}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1}\boldsymbol{\beta}_B} \tag{28.28}$$

The portfolio risk and return and the asset betas follow from Eqs. 27.6, 27.7 and 27.9 as usual:

$$\sigma_b^2 = \frac{1}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1}\boldsymbol{\beta}_B} \ , \quad \widehat{R}_b = \frac{\boldsymbol{\beta}_B^T \mathbf{C}^{-1}\widehat{\mathbf{R}}}{\boldsymbol{\beta}_B^T \mathbf{C}^{-1}\boldsymbol{\beta}_B} \ , \quad \boldsymbol{\beta}_b = \boldsymbol{\beta}_B \ . \tag{28.29}$$

Finally, the marginal risk contribution, Eq. 27.10, of each asset and the percentage of risk attributed to each asset, Eq. 27.11, are:

$$\frac{\partial \sigma_b}{\partial w_{b,i}} = \sigma_b \beta_{i,B} \ , \quad A_{b,i} = w_{b,i}\beta_{i,B} \quad \text{for } i = 1, \ldots, M \ .$$

According to Eqs. 28.21 and 28.6, the portfolio's alpha and beta with respect to the benchmark are simply zero and one

$$\alpha_{b,B} = 0 , \quad \beta_{b,B} = 1 . \tag{28.30a}$$

Since the portfolio is the benchmark itself, all above properties also describe the benchmark portfolio V_B. Moreover, the residual weights all vanish. This can also be verified explicitly by setting $\mathbf{a} = \boldsymbol{\beta}_B$ in Eq. 28.22. Therefore residual risk and return are both zero, and the Information Ratio is undefined.

28.4.3 The Characteristic Portfolio for Alpha

Let's now choose an attribute vector $\mathbf{a}$ to be the alpha of the assets with respect to the benchmark, i.e., according to Eq. 28.4

$$\mathbf{a} := \boldsymbol{\alpha}_B = \widehat{\mathbf{R}} - \boldsymbol{\beta}_B \widehat{R}_B . \tag{28.31}$$

According to Eqs. 27.1 and 28.9, the exposure of *any* portfolio V to this particular attribute is the (ex ante estimate for the) portfolio alpha:

$$a_V = \mathbf{w}^T \mathbf{a} = \sum_{i=1}^{M} w_i \alpha_{i,B} = \alpha_{V,B} .$$

The *characteristic* portfolio V_a for this particular attribute is the minimum risk portfolio with unit exposure $a_V = 1$, i.e., with a portfolio alpha of 1. According to Eqs. 27.4, 27.6 and 27.9, the *characteristic* portfolio for the attribute $\mathbf{a}$ has the following weights, variance, return and betas:

$$\mathbf{w}_a = \frac{\mathbf{C}^{-1} \boldsymbol{\alpha}_B}{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B} = \sigma_a^2 \mathbf{C}^{-1} \boldsymbol{\alpha}_B$$

$$\sigma_a^2 = \frac{1}{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B}$$

$$\widehat{R}_a = \sigma_a^2 \boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \widehat{\mathbf{R}} = \frac{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B}$$

$$\frac{\mathbf{C} \mathbf{w}_a}{\sigma_a^2} = \boldsymbol{\beta}_a = \boldsymbol{\alpha}_B . \tag{28.32}$$

Thus, the asset betas with respect to this *characteristic* portfolio V_a are equal to the asset alphas with respect to the benchmark portfolio V_B. By contrast, the *portfolio* alpha and beta of V_a with respect to the benchmark are simply one and zero[6] as follows directly[7] from Eqs. 28.21 and 28.6

$$\alpha_{a,B}=1 \,, \quad \beta_{a,B}=0 \,. \tag{28.33}$$

Since $\beta_{a,B} = \rho_{a,B}\sigma_a/\sigma_B$ this implies that portfolio V_a is totally uncorrelated with the benchmark:

$$\rho_{a,B} = 0 \,.$$

With the asset betas as in Eq. 28.32, the marginal risk contribution, Eq. 26.25, of each asset and the percentage of risk attributed to each asset, Eq. 26.26, are:

$$\frac{\partial \sigma_a}{\partial w_{a,i}} = \sigma_a \alpha_{i,B} \,, \quad A_{a,i} = w_{a,i} \alpha_{i,B} \quad \text{for} \quad i = 1, \ldots, M \,. \tag{28.34}$$

Let's look now at the residual quantities of this portfolio. Since $\beta_{a,B} = 0$ we have

$$\widetilde{\mathbf{w}}_a = \mathbf{w}_a - \beta_{a,B}\mathbf{w}_B = \mathbf{w}_a$$

and therefore

$$\widetilde{R}_a = \widehat{R}_a \text{ und } \widetilde{\sigma}_a^2 = \sigma_a^2 \,. \tag{28.35}$$

---

[6] V_a usually contains long and short positions.

[7] The fact that the portfolio beta is zero can also be derived very elegantly by applying Eq. 27.12 to the characteristic portfolios V_a and V_b:

$$b_a \sigma_b^2 = a_b \sigma_a^2 \,.$$

Here $a_b := \mathbf{w}_b^T \mathbf{a} = \mathbf{w}_b^T \boldsymbol{\alpha}_B = \alpha_{b,B}$ is the portfolio alpha of portfolio V_b with respect to the benchmark. However, as shown in Eq. 28.27, portfolio V_b is the benchmark itself and has of course zero alpha w.r.t. to itself, since a regression like Eq. 28.3 of any variable onto itself always yields $\alpha = 0$ and $\beta = 1$. Thus b_a must vanish as well, since both variances σ_a^2 and σ_b^2 are positive. But b_a, being the exposure of V_a to attribute b, is the portfolio Beta of V_a with respect to the benchmark, $b_a := \mathbf{w}_a^T \mathbf{b} = \mathbf{w}_a^T \boldsymbol{\beta}_B = \beta_{a,B}$.

Moreover, the marginal residual risk contribution of each asset and the percentage of residual risk attributed to each asset are the same as in Eq. 28.34.

$$\frac{\partial \widetilde{\sigma}_a}{\partial \widetilde{w}_{a,i}} = \sigma_a \alpha_{i,B}, \quad \widetilde{A}_{a,i} = w_{a,i} \alpha_{i,B} \quad \text{for} \quad i = 1, \dots, M .$$

Note that because of Eq. 28.15 we have

$$\widetilde{R}_a = \alpha_{a,B} = \mathbf{w}_a^T \boldsymbol{\alpha}_B = 1 \tag{28.36}$$

where the last step follows because V_a is the characteristic portfolio for $\boldsymbol{\alpha}_B$. Thus, portfolio V_a not only has $\alpha_{a,B} = 1$ but it also has expected excess return $\widehat{R}_a = 1$. It therefore has the same expected excess return as the characteristic portfolio V_A for attribute $\widehat{\mathbf{R}}$ discussed in Sect. 27.3. However, both portfolios are not necessarily the same: among all portfolios with $\mathbf{w}^T \widehat{\mathbf{R}} = 1$ there are some which also fulfill $\mathbf{w}^T \boldsymbol{\alpha}_B = 1$. However, among the other portfolios with $\mathbf{w}^T \widehat{\mathbf{R}} = 1$ there could well be some with less risk than portfolio V_a. Thus, portfolio V_a is not necessarily the minimum risk portfolio with excess return equal one, i.e., not necessarily equal to V_A.

Because of Eqs. 28.36, 28.35, and 28.32, the information ratio of the characteristic portfolio V_A is

$$\widetilde{\gamma}_a = \frac{\widetilde{R}_a}{\widetilde{\sigma}_a} = \frac{1}{\widetilde{\sigma}_a} = \frac{1}{\sigma_a} = \sqrt{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B} . \tag{28.37}$$

This can also be read in the following way: the total risk as well as the residual risk of this portfolio are both equal to the reciprocal of the Information Ratio.

Portfolio V_a has minimum risk, and because of Eq. 28.35, it also has minimum *residual* risk among all portfolios V with $\alpha_{V,B} = 1$. Thus, portfolio V_a maximizes the Information ratio Eq. 28.37 among those portfolios. To show that portfolio V_a has maximal Information Ratio among *all* portfolios (and not only among those with $\alpha_{V,B} = 1$), we will now show that for any arbitrary portfolio V with weights $\mathbf{w}$, there is a portfolio with the *same* information ratio and a portfolio alpha equal to one. We find this portfolio by constructing its weights: the residual position weights of portfolio V are given by Eq. 28.17. Because of Eq. 28.18, the Information Ratio (defined in Eq. 28.16) for any portfolio V_λ with residual weights $\lambda \widetilde{\mathbf{w}}$ is the same for any $\lambda > 0$. The weights

$\mathbf{w}_\lambda$ of this portfolio follow directly from the requirement $\widetilde{\mathbf{w}}_\lambda = \lambda \widetilde{\mathbf{w}}$:

$$\widetilde{\mathbf{w}}_\lambda = \lambda \widetilde{\mathbf{w}}$$
$$\mathbf{w}_\lambda - \beta_{\lambda,B}\mathbf{w}_B = \lambda \left(\mathbf{w} - \beta_{V,B}\mathbf{w}_B\right)$$
$$\mathbf{w}_\lambda = \lambda\mathbf{w} + (\beta_{\lambda,B} - \lambda\beta_{V,B})\mathbf{w}_B \ . \tag{28.38}$$

We can adjust λ such that the alpha of portfolio V_λ becomes one:

$$1 \overset{!}{=} \alpha_{\lambda,B} = \widetilde{R}_\lambda = \lambda\widetilde{R} = \lambda\alpha_{V,B} \implies \lambda = \frac{1}{\alpha_{V,B}} \ .$$

Thus, for any given portfolio V with weights $\mathbf{w}$ there is indeed a portfolio V_λ with the *same* information ratio and a portfolio alpha equal to one. The weights of this portfolio can be obtained by solving the equation[8]

$$\mathbf{w}_\lambda = \frac{1}{\alpha_{V,B}}\mathbf{w} + (\beta_{\lambda,B} - \frac{\beta_{V,B}}{\alpha_{V,B}})\mathbf{w}_B \ .$$

This proves that portfolio V_a, having maximal Information Ratio among all portfolios V with $\alpha_{V,B} = 1$, has indeed maximal Information Ratio among *all* portfolio.

$$\widetilde{\gamma}_{\max} = \widetilde{\gamma}_a = \frac{1}{\widetilde{\sigma}_a} = \frac{1}{\sigma_a} \ . \tag{28.39}$$

This Equation is the benchmark-relative analogue to Eq. 27.23.

Choosing the arbitrary portfolio V in Eq. 28.38 to be the characteristic portfolio V_a itself (and observing Eq. 28.33) we get the most general form a portfolio with the same (i.e., maximal) Information Ratio as V_a may have:

$$\mathbf{w}_\lambda = \lambda\mathbf{w}_a + \beta_{\lambda,B}\mathbf{w}_B \quad \text{mit } \lambda > 0 \tag{28.40}$$
$$\implies \widetilde{\gamma}_\lambda = \widetilde{\gamma}_a \ .$$

---

[8]According to Eq. 28.26, $\beta_{\lambda,B}$ also contains the weights $\mathbf{w}_\lambda$:

$$\beta_{\lambda,B} = \frac{\mathbf{w}_\lambda^T \mathbf{C}\mathbf{w}_B}{\mathbf{w}_B^T \mathbf{C}\mathbf{w}_B} \ .$$

Therefore solving for $\mathbf{w}_\lambda$ can only be done numerically.

Equation 28.40 provides a handy method for *constructing* portfolios with maximum Information ratio using the characteristic portfolios V_a and $V_b = V_B$. Furthermore, it provides a powerful tool to *check* whether a given portfolio has maximum IR. We will come back to this point later.

To complete this section, we will now show that a relation similar to Eq. 27.25 for Sharpe Ratios also holds in the benchmark-relative framework for Information Ratios. With Eq. 28.32 and the general Eq. 28.9, the alpha of any portfolio with weights $\mathbf{w}$ can be expressed in terms of covariances with the characteristic portfolio V_a as

$$\alpha_{V,B} = \mathbf{w}^T \boldsymbol{\alpha}_B = \frac{\mathbf{w}^T \mathbf{C} \mathbf{w}_a}{\sigma_a^2} = \frac{\delta t \, \mathrm{cov}(r_V, r_a)}{\sigma_a^2} = \frac{\delta t \, \mathrm{cov}(\widehat{r}_V, \widehat{r}_a)}{\sigma_a^2} = \frac{\delta t \, \mathrm{cov}(\widetilde{r}_V, \widetilde{r}_a)}{\sigma_a^2} \, ,$$

where the second-to-last step is trivial, since subtracting a constant r_f doesn't change the covariances. The last step, however, is not trivial and only holds because portfolio V_a has zero beta with respect to the benchmark.[9] With this alpha, the IR of portfolio V can be written as[10]

$$\widetilde{\gamma}_V = \frac{\widetilde{R}_V}{\widetilde{\sigma}_V} = \frac{\alpha_{V,B}}{\widetilde{\sigma}_V} = \frac{\delta t \, \mathrm{cov}(\widetilde{r}_V, \widetilde{r}_a)}{\widetilde{\sigma}_V \sigma_a^2} = \underbrace{\frac{1}{\widetilde{\sigma}_a}}_{\widetilde{\gamma}_a} \underbrace{\frac{\delta t \, \mathrm{cov}(\widetilde{r}_V, \widetilde{r}_a)}{\widetilde{\sigma}_V \widetilde{\sigma}_a}}_{\mathrm{corr}(\widetilde{r}_V, \widetilde{r}_a)} \, ,$$

where in the second step we have used Eq. 28.15 and in the last step we have used Eqs. 28.35 and 28.37. Therefore, the IR of any portfolio V is given by the (maximum) IR of the characteristic portfolio V_a and the correlation between the *residual* returns of those two portfolios:

$$\widetilde{\gamma}_V = \widetilde{\gamma}_a \, \mathrm{corr}(\widetilde{r}_V, \widetilde{r}_a) \quad \text{for all portfolios } V \, . \tag{28.41}$$

This equation is the benchmark-relative analogue to Eq. 27.25.

---

[9]This can be derived from first principles with the help of Eq. 28.33:

$$\mathrm{cov}(\widetilde{r}_V, \widetilde{r}_a) = \mathrm{cov}(\widehat{r}_V - \beta_{V,B}\widehat{r}_B, \widehat{r}_a - \underbrace{\beta_{a,B}\widehat{r}_B}_{0})$$

$$= \mathrm{cov}(\widehat{r}_V - \beta_{V,B}\widehat{r}_B, \widehat{r}_a)$$

$$= \mathrm{cov}(\widehat{r}_V, \widehat{r}_a) - \beta_{V,B} \, \mathrm{cov}(\widehat{r}_B, \widehat{r}_a)$$

$$= \mathrm{cov}(\widehat{r}_V, \widehat{r}_a) - \beta_{V,B} \, \sigma_b^2 \underbrace{\beta_{a,B}}_{0} \, .$$

[10]To get the factors δt right, observe Eqs. 26.16, 26.18 and 21.31.

28.5 Relations Between Sharpe Ratio and Information Ratio

To find out how the Sharpe Ratio used in Total Return Management is related to the Information Ratio used in the benchmark-relative framework, we will analyze the two characteristic portfolios with maximum Sharpe Ratio, i.e., portfolios V_A of Sect. 27.3 and portfolio V_m of Sect. 27.4, in the benchmark framework.

28.5.1 The Market Portfolio

We will begin by applying Eq. 28.40 to check if the market portfolio of Sect. 27.4 has maximum IR. To bring its weights $\mathbf{w}_m$ into a form similar to Eq. 28.40, we write the asset returns in Eq. 28.4 with Eqs. 28.32 for $\boldsymbol{\alpha}_B$ and 28.29 for $\boldsymbol{\beta}_B$:

$$\widehat{\mathbf{R}} = \boldsymbol{\alpha}_B + \boldsymbol{\beta}_B \widehat{R}_B = \frac{\mathbf{C}\mathbf{w}_a}{\sigma_a^2} + \frac{\mathbf{C}\mathbf{w}_B}{\sigma_B^2} \widehat{R}_B \; . \tag{28.42}$$

Expressing now the asset returns on the left hand side in terms of the market portfolio as in Eq. 27.31, we get

$$\frac{\mathbf{C}\mathbf{w}_m}{\sigma_m^2} \widehat{R}_m = \frac{\mathbf{C}\mathbf{w}_a}{\sigma_a^2} + \frac{\mathbf{C}\mathbf{w}_B}{\sigma_B^2} \widehat{R}_B \; .$$

Solving for $\mathbf{w}_m$ shows that the market portfolio can be written as a combination of the benchmark portfolio V_B and the characteristic portfolio V_a.

$$
\begin{aligned}
\mathbf{w}_m &= \frac{\sigma_m^2}{\sigma_a^2} \frac{1}{\widehat{R}_m} \mathbf{w}_a + \frac{\sigma_m^2}{\sigma_B^2} \frac{\widehat{R}_B}{\widehat{R}_m} \mathbf{w}_B \\
&= \frac{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \, \mathbf{w}_a + \frac{\boldsymbol{\beta}_B^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \, \mathbf{w}_B \\
&= \underbrace{\frac{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B}{\mathbf{1}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}}}_{\lambda} \mathbf{w}_a + \underbrace{\left(\boldsymbol{\beta}_B^T \mathbf{w}_m\right)}_{\beta_{m,B}} \mathbf{w}_B \; ,
\end{aligned}
\tag{28.43}
$$

Here we have used Eqs. 27.31, 28.29 and 28.32 in the second step and Eq. 27.30 in the last step. Although it looks very similar, λ is *not* the portfolio alpha of the market portfolio.[11] Nonetheless, Eq. 28.43 has indeed the form of Eq. 28.40 with $\lambda > 0$ as long as $\widehat{R}_m > 0$. Therefore, as is always the case when the market portfolio is involved, everything only works if Inequality 26.59 holds.[12]

If Inequality 26.59 holds, then the market portfolio has maximal Information Ratio (in addition to having maximal Sharpe Ratio).

$$\widetilde{\gamma}_m = \widetilde{\gamma}_a = \widetilde{\gamma}_{\max} \ . \tag{28.44}$$

Because of Eq. 28.41, this means that the residual returns of the market portfolio are fully correlated with the residual returns of portfolio V_a.

$$\mathrm{corr}(\widetilde{r}_m, \widetilde{r}_a) = 1 \ .$$

From the above expression for $\mathbf{w}_m$, we find that the *residual* weights of the market portfolio can be expressed in terms of the weights of portfolio V_a as

$$\widetilde{\mathbf{w}}_m \equiv \mathbf{w}_m - \beta_{m,B}\mathbf{w}_B = \frac{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1}\boldsymbol{\alpha}_B}{\mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}}\mathbf{w}_a = \frac{\sigma_m^2}{\sigma_a^2}\frac{1}{\widehat{R}_m}\mathbf{w}_a \ . \tag{28.45}$$

With these weights, the residual risk of the market portfolio becomes

$$\widetilde{\sigma}_m = \sqrt{\widetilde{\mathbf{w}}_m^T \mathbf{C}\widetilde{\mathbf{w}}_m} = \frac{\sigma_m^2}{\sigma_a^2}\frac{1}{\widehat{R}_m}\sqrt{\mathbf{w}_a^T \mathbf{C}\mathbf{w}_a} = \frac{\sigma_m^2}{\sigma_a \widehat{R}_m} \ . \tag{28.46}$$

Similarly, the residual return of the market portfolio is

$$\widetilde{R}_m = \widetilde{\mathbf{w}}_m^T \widehat{\mathbf{R}} \equiv \frac{\sigma_m^2}{\sigma_a^2}\frac{1}{\widehat{R}_m}\mathbf{w}_a^T \widehat{\mathbf{R}} = \frac{\sigma_m^2}{\sigma_a^2}\frac{\widehat{R}_a}{\widehat{R}_m} \ . \tag{28.47}$$

---

[11] The Alpha of the market portfolio is given by

$$\alpha_{m,B} = \boldsymbol{\alpha}_B^T \mathbf{w}_m = \frac{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1}\widehat{\mathbf{R}}}{\mathbf{1}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}} \ .$$

[12] This is usually (but unfortunately not always) the case, since the expected return of the market portfolio should be above the risk free rate as a compensation for the risk of the market portfolio.

Inserting these results into Eq. 28.16 we find the *Information* Ratio of the market portfolio to be the same as the *Sharpe* Ratio of portfolio V_a.

$$\widetilde{\gamma}_m = \frac{\widetilde{R}_m}{\widetilde{\sigma}_m} = \frac{\widehat{R}_a}{\sigma_a} = \gamma_a$$

or, written with $\widetilde{\gamma}_{\max}$ from Eq. 28.44:

$$\widetilde{\gamma}_{\max} = \widetilde{\gamma}_m = \gamma_a \ . \tag{28.48}$$

where in the last step we have used Eq. 28.44. Therefore, the (maximum) Information Ratio of portfolio V_a is equal to its Sharpe Ratio. However, this is not the *maximum* Sharpe Ratio. The relation between the maximum Sharpe Ratio γ_m (see Eq. 27.33) and the *maximum* Information Ratio $\widetilde{\gamma}_m = \widetilde{\gamma}_a$ (see Eq. 28.37) is

$$\frac{\gamma_{\max}}{\widetilde{\gamma}_{\max}} = \frac{\gamma_m}{\widetilde{\gamma}_a} = \frac{\sigma_a \widehat{R}_m}{\sigma_m} = \frac{\sigma_m}{\widetilde{\sigma}_m} \ ,$$

where we have used Eq. 28.46 in the last step. **The maximum Sharpe Ratio is as much larger than the maximum Information Ratio as the total risk of the market portfolio is larger than its residual risk**.

$$\gamma_{\max} = \frac{\sigma_m}{\widetilde{\sigma}_m} \widetilde{\gamma}_{\max} \ . \tag{28.49}$$

With $\gamma_{\max} = \gamma_m = \widehat{R}_m / \sigma_m$ and $\widetilde{\gamma}_{\max} = \widetilde{\gamma}_m = \widetilde{R}_m / \widetilde{\sigma}_m$, this means that the ratio of excess and residual return of the market portfolio equals the ratio of the *squares* of its risk and residual risk:

$$\frac{\widehat{R}_m}{\widetilde{R}_m} = \frac{\sigma_m^2}{\widetilde{\sigma}_m^2} = \frac{\mathrm{var}\,[r_m]}{\mathrm{var}\,[\epsilon_m]} \ , \tag{28.50}$$

where in the last step we have used the original definition 28.14 of the residual risk.

The *marginal* contributions of the assets to the residual risk of the market portfolio are by definition

$$\frac{\partial \widetilde{\sigma}_m}{\widetilde{\mathbf{w}}_m} = \frac{\mathbf{C}\widetilde{\mathbf{w}}}{\sqrt{\widetilde{\mathbf{w}}^T \mathbf{C}\widetilde{\mathbf{w}}}} = \frac{\mathbf{C}\widetilde{\mathbf{w}}}{\widetilde{\sigma}_m} \ .$$

With the above results 28.45 and 28.46, this can be written as

$$\frac{\partial \widetilde{\sigma}_m}{\widetilde{\mathbf{w}}_m} = \frac{\sigma_m^2}{\sigma_a^2} \frac{1}{\widehat{R}_m} \frac{\mathbf{C}\mathbf{w}_a}{\widetilde{\sigma}_m} = \frac{\mathbf{C}\mathbf{w}_a}{\sigma_a} \ .$$

We can now apply Eqs. 28.32 to write the marginal residual risk contributions to the market portfolio in several useful forms.

$$\frac{\partial \widetilde{\sigma}_m}{\widetilde{\mathbf{w}}_m} = \sigma_a \boldsymbol{\beta}_a = \sigma_a \boldsymbol{\alpha}_B \ . \tag{28.51}$$

With Eq. 28.37 for σ_a we get a particularly insightful result:

$$\boldsymbol{\alpha}_B = \widetilde{\gamma}_{\max} \frac{\partial \widetilde{\sigma}_m}{\widetilde{\mathbf{w}}_m} \ . \tag{28.52}$$

The asset alphas are proportional to the assets' marginal residual risk contributions to the market portfolio, with the maximum Information Ratio being the constant of proportionality.

28.5.2 The Characteristic Portfolio of the Excess Return

Let's now analyze the excess return's characteristic portfolio V_A with its properties given by Eq. 27.24. Since V_A is closely related to the market portfolio via Eq. 27.29, we can directly use Eq. 28.43 to express V_A in terms of the characteristic portfolios for alpha and beta as

$$\mathbf{w}_A = \frac{\sigma_A^2}{\sigma_L^2} \widehat{R}_L \mathbf{w}_m = \frac{\sigma_m^2}{\sigma_L^2} \frac{\widehat{R}_L}{\widehat{R}_m} \left[\frac{\sigma_A^2}{\sigma_a^2} \mathbf{w}_a + \frac{\sigma_A^2}{\sigma_B^2} \widehat{R}_B \mathbf{w}_B \right]$$

$$= \frac{\sigma_A^2}{\sigma_a^2} \mathbf{w}_a + \frac{\sigma_A^2}{\sigma_B^2} \widehat{R}_B \mathbf{w}_B = \frac{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1} \boldsymbol{\alpha}_B}{\widehat{\mathbf{R}}^T \mathbf{C}^{-1} \widehat{\mathbf{R}}} \ \mathbf{w}_a + \underbrace{\left(\boldsymbol{\beta}_B^T \mathbf{w}_A \right)}_{\beta_{A,B}} \mathbf{w}_B \ ,$$

where Eqs. 27.18, 27.24, 28.29 and 28.32 have been used. This again has the form of Eq. 28.40 with $\lambda > 0$ for any positive definite Matrix $\mathbf{C}$. Thus, **the excess return's characteristic portfolio has maximal Information Ratio (in addition to having maximal Sharpe Ratio).**

$$\widetilde{\gamma}_A = \widetilde{\gamma}_a = \widetilde{\gamma}_{\max} \ .$$

This holds in *every* situation, even if Inequality 26.59 does not hold. The *residual* weights of portfolio V_A can now be expressed in terms of the weights of portfolio V_a as

$$\widetilde{\mathbf{w}}_A \equiv \mathbf{w}_a - \beta_{A,B}\mathbf{w}_B = \frac{\sigma_A^2}{\sigma_a^2}\mathbf{w}_a \ .$$

With these weights, the residual variance of the portfolio becomes

$$\widetilde{\sigma}_m = \sqrt{\widetilde{\mathbf{w}}_A^T \mathbf{C} \widetilde{\mathbf{w}}_A} = \frac{\sigma_A^2}{\sigma_a^2}\sqrt{\mathbf{w}_a^T \mathbf{C} \mathbf{w}_a} = \frac{\sigma_A^2}{\sigma_a^2}\sigma_a \ .$$

Similarly, the residual return of the portfolio is

$$\widetilde{R}_A = \widetilde{\mathbf{w}}_A^T \widehat{\mathbf{R}} = \frac{\sigma_A^2}{\sigma_a^2}\mathbf{w}_a^T \mathbf{R} = \frac{\sigma_A^2}{\sigma_a^2}\widehat{R}_a \ .$$

Inserting these results into Eq. 28.16, we find the *Information* Ratio of the portfolio to be the same as the *Sharpe Ratio* of portfolio V_a.

$$\widetilde{\gamma}_A = \frac{\widetilde{R}_A}{\widetilde{\sigma}_A} = \frac{\widehat{R}_a}{\sigma_a} = \gamma_a \ .$$

This is no surprise since we already established $\widetilde{\gamma}_A = \widetilde{\gamma}_a$ and we know from Eq. 28.48 that $\gamma_a = \widetilde{\gamma}_A$.

The relation between the maximum Sharpe Ratio γ_A (see Eq. 27.23) and the maximum Information Ratio $\widetilde{\gamma}_A = \widetilde{\gamma}_a$ (see Eq. 28.37) is

$$\frac{\gamma_A}{\widetilde{\gamma}_a} = \sqrt{\frac{\widehat{\mathbf{R}}^T \mathbf{C}^{-1}\widehat{\mathbf{R}}}{\boldsymbol{\alpha}_B^T \mathbf{C}^{-1}\boldsymbol{\alpha}_B}} = \frac{\sigma_a}{\sigma_A} \ .$$

Thus the relation between the maximum Sharpe Ratio and the maximum Information Ratio can also be expressed in terms of the two volatilities σ_a and σ_A. Together Eq. 28.49, the relationships between maximum Sharp Ratio and maximum Information Ratio are in summary:

$$\frac{\sigma_m}{\widetilde{\sigma}_m} = \frac{\gamma_{\max}}{\widetilde{\gamma}_{\max}} = \frac{\sigma_a}{\sigma_A} \ . \tag{28.53}$$

The first equality, involving the market portfolio, is only true if Inequality 26.59 holds, while the second equality holds in *every* situation.

Part VI

Market Data

29

Check for updates

Construction of the Yield Curve Universe

In the previous chapters we made intensive use of *yield curves*. With the help of yield curves discount factors, i.e., zero-bond prices, are determined and thus the (risk-neutral) present value of future cash flows. Yield curves are therefore of fundamental importance for the valuation of all financial products. The interest rates, also called *zero interest rates*, underlying the discount factors or zero bond prices depend on the maturities of the respective zero-bonds. Strictly speaking, a yield curve describes an entire continuum of different market or risk factors, where each risk factor corresponds to a zero rate with a certain maturity (as seen from today). To emphasize this time-dependency, this is often called the *interest rate term structure* or *term structure* for short.

Since these different zero rates are not independent of each other, but on the contrary very strongly correlated (see Chap. 34), and since only certain maturities of interest rate instruments are traded anyway, it is perfectly sufficient to describe a yield curve by a few constant parameters. For example, these can be the zero rates or discount factors for a few maturities (the grid points of the yield curve). Alternatively, a yield curve can also be modelled as a parametric function using a finite set of constant parameters. In both cases, the parameters are determined in such a way that the prices of a specified basket of benchmark instruments (or benchmark quotes) are replicated as exactly as possible. If the number of parameters is equal to the number of benchmark instruments and no two benchmark instruments have the same maturity, it is in principle possible to hit the benchmark prices exactly. The construction of a yield curve can therefore also be understood as a methodology for the

interpolation and extrapolation of zero rates or discount factors for such maturities that are not present among the benchmark instruments.

The discount curve or yield curve described above is the *risk-free* yield curve. It is independent of any risk other than pure interest rate risk. However, even the most liquidly traded instruments contain other risks in addition to interest rate risk, such as basis risk, liquidity risk and default risk. Therefore, numerous other *risky* yield curves are needed in addition to the risk-free curve. This set of curves is also called the *yield curve universe*. For the calculation of forward-rates, for instance, *basis spread yield curves* (also called *forward yield curves*) are needed, which take into account that forward interest rates are different from risk-free rates by a basis depending on the maturity as well as the start time. For the valuation of bonds, yield curves are required that incorporate the specific default risk of the issuer. Occasionally, liquidity risk is considered as separate from default risk. The valuation of instruments denominated in foreign currencies, transformed into the home currency via a cross currency swap, requires yield curves that reflect the cross currency basis (see Sect. 16.4).

29.1 Required Features of Discount Curves

We initially limit our considerations to (risk-free) discount curves. At any given time t, the yield curve is a function $f(T)$ of the maturity T. This function specifies the discount factor for discounting from T to t for each T and is simply the present value at time t of a zero bond which pays one currency unit at the later time T:

$$f(T) = B(t, T) \ .$$

Thus, we simply write $B(t, T)$ with t fixed. Alternatively, we could also view the yield curve as a function of the zero rates R, i.e., $f(T) = R(T)$. Then the specification of an interest rate convention would also be necessary to interpret the yield curve unambiguously. However, nothing is gained this way because discount factors can be transformed into zero rates within any interest rate convention, see Table 2.5.[1] In the following, we therefore always view the yield curve as discount factors as a function of the their maturity.

---

[1]Strictly speaking, this only holds true for interest rates greater than $-m$, where m is the number of discrete compoundings per year, since otherwise the formula for discrete compounding would fail. The authors firmly believe that this case can be ruled out safely in the real world.

For $t =$ today this is today's yield curve. Just as well, however, we can also specify a future yield curve with $t >$ today, which might be simulated in the context of a term structure model; or, for $t <$ today, a historical yield curve, which is calculated using historical benchmark prices. Instead of time points/dates, the time can also be expressed as a duration relative to t. This, however, necessitates the additional specification of a day count method (see Sect. 2.1.1). For numerical calculations e.g. act/365 is a suitable choice since this method provides a simple and fast conversion of a date into a time period. Today (or now) is then simply $t = 0$.

In the market, there exists no established standard method for determining $B(t, T)$ if T does not coincide with the remaining maturity T_i of a benchmark instrument. Nevertheless, we can impose some requirements on the characteristics of the yield curve that limit the choice of possible interpolation and extrapolation procedures:

1. The prices of benchmark instruments should be replicated as accurately as possible. This in particular ensures that the various procedures all yielding (almost) the same values for the grid points.
2. The function $f(T)$ should be continuous. It's not realistic for a discount factor or interest rate to jump strongly from one day to the next.
3. The extrapolation, especially at the long end of the yield curve, i.e. for large T, should not become arbitrarily large.
4. Oscillations of the yield curve should be avoided. Such oscillations can occur with some interpolation methods such as Cubic Spline, for example, when the rates between closely spaced sampling points differ greatly.
5. Depending on how the yield curve is used, the function $f(T)$ should be differentiable once or twice. This is important, for example, for some short rate models whose drift terms depend on the first and possibly also the second derivative of the yield curve with respect to time.

These requirements are not strict boundary conditions, which must be fulfilled in any case. Note, for example, that in a past version this list contained the requirement to avoid negative interest rates. Reality has proven that this requirement was unfounded. Nevertheless, these requirements provide us with a kind of guideline to qualitatively assess different inter- and extrapolation methods.

29.2 Modeling the Yield Curve

Since only a finite number n of benchmark instruments is available to generate a yield curve, that curve is actually only available at certain points in time, the so-called *grid points* or *Vertices* $T_1, T_2, \ldots, T_n$. A benchmark instrument implicitly contains information up to and including the maturity of the last interest rate on which the benchmark instrument depends. For a deposit or a swap, this is the maturity of the instrument, for a future or FRA it is the maturity of the underlying forward rate, even if the last payment was actually made earlier. In this way one can associate exactly one vertex T_i to the i-th benchmark instrument. Then, a discount factor $B_i = B(t, T_i)$ is assigned to each vertex T_i. The Sect. 29.6 deals with the question how these B_i are determined. Here and for the time being, however, we take the n pairs (T_i, B_i) with $i = 1, 2, \ldots, n$ as given and only investigate the question how to use these data to interpolate discount factors for arbitrary T with $T_1 < T < T_n$ and extrapolate for $T < T_1$ or $T > T_n$.

The interpolation does not necessarily have to take place for the discount factors. Instead, you can, for example, interpolate and extrapolate the corresponding interest rates. These can be calculated from the B_i using both the correct compounding method (annual, semi-annual, continuous, simple, etc.) and the correct day count method applicable to the market/instrument in question. Another alternative is to interpolate the logarithms of the discount factors, i.e. $\ln(B_i)$ instead of B_i. This is called *log interpolation*. Therefore, it is essential for any interpolation procedure to specify exactly which variables (discount factors, zero rates, forward rates, etc.) are interpolated.

Ultimately, there is always freedom of choice with regard to the interpolation method. One method or another may seem more appropriate for a particular purpose, but there is no "right" or "wrong", as the market data does not provide information on the "right" discount factor between two vertices Negatively formulated, this can also be considered as a systematic model risk: Valuation differences resulting from different (but plausible) interpolation methods cannot be resolved.

29.2.1 Interpolation Methods

An *interpolation method* is generally a function $f(t)$, that for a given set of value pairs $\{(t_1, F_1), (t_2, F_2), \ldots, (t_n, F_n)\}$, has the following properties:

$$f(t_i) = F_i \quad \text{for all } i = 1, 2, \ldots, n,$$

i.e. at the grid points the function matches exactly with the F_i and otherwise there are no further conditions. Depending on the intended use, however, further requirements may be added, such as e.g. differentiability (also at the grid points).

Linear Interpolation

For linear interpolation, the function $f(t)$ is defined piecewise for adjacent t_i. To do this, the smaller and larger vertex t_m and t_{m+1} closest to t must first be determined with $t_m \leq t < t_{m+1}$. On the interval $[t_m, t_{m+1}]$ the interpolated $f(t)$ is then defined as

$$f(t) = F_m + \frac{F_{m+1} - F_m}{t_{m+1} - t_m}(t - t_m) \tag{29.1}$$

$$= \frac{F_m(t_{m+1} - t) + F_{m+1}(t - t_m)}{t_{m+1} - t_i} .$$

This function is continuous everywhere. The function has kinks at the grid points, so that it cannot be differentiated there. The second derivative is always zero (except at the grid points where it is not defined). Linear interpolation is a fast, simple method which is commonly used, if e.g. differentiability is not a necessity. Often, the interpolation is done for the logarithms of the discount factors. This is called *log linear* interpolation.

Constant Interpolation

The constant interpolation sets a constant value F for $f(t)$ on each interval $[t_m, t_{m+1}]$:

$$f(t) = F \quad t \in [t_m, t_{m+1}] .$$

Since the constant value F can be different in each interval, the method is also called *piecewise constant* interpolation. For example, you can choose $F = F_m$, $F = F_{m+1}$ or $F = (F_{m+1} + F_m)/2$, where F is different for each interval. It follows directly that the function has jumps at the vertices and is therefore neither differentiable nor continuous there. A constant interpolation of discount factors or zero rates is therefore out of the question. However, as an interpolation method for forward rates, piecewise constant interpolation

is a sensible alternative. In particular, constant interpolation of the forward rates in continuous compounding is identical to a log linear interpolation of the discount factors. To show this, we consider the interpolation of a discount factor in log-linear interpolation:

$$\ln B(t, T) = \ln B(t, T_m) + \frac{\ln B(t, T_{m+1}) - \ln B(t, T_m)}{T_{m+1} - T_m}(T - T_m)$$

$$= \ln B(t, T_m) + \frac{\ln \left(\frac{B(t, T_{m+1})}{B(t, T_m)} \right)}{T_{m+1} - T_m}(T - T_m)$$

$$= \ln B(t, T_m) + \frac{\ln B(T_m, T_{m+1}|t)}{T_{m+1} - T_m}(T - T_m) . \qquad (29.2)$$

The last line follows from 2.7. If $B(T_m, T_{m+1}|t)$ is expressed by a constant forward rate $F_{m,m+1}$ for the period from T_m to T_{m+1}, $B(T_m, T_{m+1}|t) = \exp[-F_{m,m+1}(T_{m+1} - T_m)]$ (see Table 2.5), then it follows:

$$\ln B(t, T) = \ln B(t, T_m) - \frac{F_{m,m+1}(T_{m+1} - T_m)}{T_{m+1} - T_m}(T - T_m)$$

$$= \ln B(t, T_m) - F_{m,m+1}(T - T_m)$$

$$= \ln B(t, T_m) + \ln B(T_m, T|t) ,$$

where the last line again is valid because of Eq. 2.7 if the natural logarithm is calculated for both sides. Then $F_{m,m+1}$ is the forward rate interpolated for the period T_m to T_{m+1}, which corresponds to constant interpolation on the interval $[T_m, T_{m+1}]$ with $F = F_{m,m+1}$.

Other Interpolation Methods

In addition to the interpolation methods presented here, many others are also used in practice. Cubic splines approximate the sections between the grid points by 3rd degree polynomials and have the property of being differentiable twice with respect to time (compare this to linear interpolation which has kinks at the grid points and therefore cannot even be differentiated once). However, unwanted oscillations may occur within cubic spline interpolations, i.e. the interpolated curve exhibits significantly greater fluctuations than the values at the vertices suggest. Constraint splines are differentiable only once,

but avoid the problem of oscillations, since it is ensured that any interpolated value always lies between the values of its adjacent vertices [126]. Another smooth interpolation method, which, however, also leads to oscillations (albeit a bit less severe than cubic splines) is the Hermite interpolation. The Excel workbook INTERPOLATION.XLSX from the download area [50] demonstrates the differences between different interpolation methods.

29.2.2   Extrapolation Methods

Sometimes discount factors are also needed beyond the range covered by the available benchmark instruments. The calculation of values outside the range of known quantities is called *extrapolation*. An extrapolation is inherently more uncertain than an interpolation (at least if one assumes that the "true" function underlying the observed data is "good", i.e. smooth and without strong deflections in one direction or another). At the *long end* of the yield curve, i.e. for long maturities, it is not clear from the outset whether interest rates should rise, fall or remain constant. It is often assumed that interest rates for very long maturities fluctuate around a certain mean (without being able, however, to accurately determine said mean). Therefore, the first ground rule for extrapolations is to avoid them whenever possible, for example by selecting suitable benchmark instruments that cover the whole range of required maturities.

Such problems don't exist at the *short end* of the yield curve, i.e. for short maturities of just a few days, since, at least in the major currencies, liquid benchmark quotes are available for short-running o/n- and   t/n-Deposits (overnight and tomorrow-next, with start date today or the next trading day and maturity and redemption one trading day later). Besides, it is certain that the discount factor for a maturity in 0 days is exactly 1, since, of course, no interest accumulates over zero time. If the yield curve is constructed with discount factors, one can therefore always insert a vertex at $T = t$ with $B(t, t) = 1$. With log-linear interpolation for the discount factors this procedure corresponds to a constant extrapolation of the continuously compounded zero rate of the first vertex:

$$
\ln B(t, T) = \ln B(t, t) + \frac{\ln B(t, T_1) - \ln B(t, t)}{T_1 - t}(T - t)
$$

$$
= -\frac{R_1(T_1 - t)}{T_1 - t}(T - t)
$$

$$
= -R_1(T - t) ,
$$

where $\ln B(t, t) = \ln 1 = 0$ was used. R_1 denotes the continuously compounded zero rate at the first vertex T_1 with $B(t, T_1) = \exp(-R_1 (T_1 - t))$.

The extrapolation at the long end is more problematic. Life insurance policies require very long time horizons of up to 70 years to model their interest rate risks and even beyond that. However, no liquidly traded benchmark instruments are available for such long maturities and thus, extrapolations can't be avoided. Since no reliable data are available anyway, very simple extrapolation approaches are commonly used. Very often, the interest rates for all maturities later than the last vertex are assumed to be constant and equal to the interest rate R_n associated with that last vertex (the discount factor cannot be kept constant as this implies forward interest rates equal to zero). An alternative method is to simply add an additional vertex for the maximum required maturity and assume the interest rate R_∞ for such very long periods to be the long-time historical mean or an expert's estimate. An explicit asymptotic behavior towards a limiting value R_∞ can be obtained using an ansatz of the form

$$R_\infty + b \exp(-\gamma (T - t))$$

where the parameters b and γ are determined by requiring that the price $B(t, T_n)$ and its first derivative are hit exactly at the last vertex T_n.

29.3 Parametric Yield Curves

An alternative to the value pairs of grid point and discount factor described above are parametric approaches which describe the yield curve by a mathematical function depending on only a few parameters (usually far fewer in number than the benchmark instruments). The parameters are determined via optimization procedures in such a way that the prices of the benchmark instruments are hit as closely as possible. Advantages of this method are that the yield curves can be modeled so that they meet very strict smoothness requirements (e.g. multiple or infinitely often differentiable), and can be calculated very quickly, with no further data needed except three or four parameters. The optimization procedures are not significantly different from those described in Sect. 29.6. A disadvantage, however, is that not all benchmark prices can be reproduced exactly, as the number of parameters is significantly lower than the number of benchmark prices. The best known of these parametric approaches was developed by Nelson and Siegel and is presented in the following section.

Nelson-Siegel-Procedure

Nelson and Siegel have proposed a parameterization of the yield curve with the interest rate $R(t, T)$ being a function of the maturity T as follows [148]:

$$R(t, T) = \beta_0 + \beta_1 \frac{1 - e^{-\gamma(T-t)}}{\gamma(T - t)} + \beta_2 \left(\frac{1 - e^{-\gamma(T-t)}}{\gamma(T - t)} - e^{-\gamma(T-t)} \right) .$$

The parameters β_0, β_1, β_2 and γ of this model are selected or *optimized* in such a way that the prices of all benchmark instruments are hit as accurately as possible. β_0 is a measure for the interest rate for very long maturities, since the other two terms become arbitrarily small for large $T - t$. The parameter γ can be used to control how quickly the second and third term approach 0. The second term is dominant at the short end of the yield curve and the third term at medium maturities. Svensson proposed an extension of the Nelson-Siegel model [177], in which an additional term with two additional parameters is added (also called Nelson-Siegel-Svensson or NSS model).

$$\begin{aligned} R(t, T) =&\beta_0 + \beta_1 \frac{1 - e^{-\gamma(T-t)}}{\gamma(T - t)} + \beta_2 \left(\frac{1 - e^{-\gamma(T-t)}}{\gamma(T - t)} - e^{-\gamma(T-t)} \right) \\ &+ \beta_3 \left(\frac{1 - e^{-\gamma_1(T-t)}}{\gamma_1(T - t)} - e^{-\gamma_1(T-t)} \right) . \end{aligned}$$

At first sight, these two parameterization approaches look like they are not defined for $T = t$ because of the term $T - t$ in the denominator. However, it can easily be shown that $R(t, t) = \beta_0$ at this boundary, i.e. the function can be continuously continued at this point. Although most typical yield curves, such as rising, falling and those with a small hump or local maximum at the short end (occasionally a local minimum) can be modeled by these parametric forms, it is not possible to hit all liquid benchmark quotes of a typical yield curve. That's partly simply because 15 or 20 available prices of benchmark instruments contain much more information than can be expressed by just four or six parameters. Also, it is not easy to find the optimal set of parameters that best meets the quoted prices of a given set of benchmark instruments [76]. An example for an optimization result is shown in Fig. 29.1. These parametric approaches are therefore rarely used today for modeling discount curves, where an accurate valuation of interest rate derivatives such as swaps is important. However, they are still used today in cases where very little market data is

available and simple interpolation and extrapolation would lead to implausible results. This applies, for example, to zero bond spread curves (depending on the market segment and rating class, very few liquid bond prices might be available for calibration). The same applies to spread curves that cannot be observed directly, such as liquidity spread curves. Bond sector curves are another application. The problem here is that bond prices are not only dependent on the current interest rates, but also on other effects, such as liquidity, default risk, etc., some of which are specific to individual bonds. This is called *specific* interest rate risk. In such cases it is not expedient to try to replicate all prices with a single yield curve. Instead, the yield curve should only contain the *general* interest rate risk, meaning the average interest rate for a class or sector of bond issuers.

29.4 Construction of the Curve Hierarchy

In former times there were only a few yield curves per currency, today one has to deal with a whole zoo of yield curves, which is why one speaks of a *yield curve universe*. Unless otherwise stated, in this book we mean by a yield curve or, more precisely, a discount factor curve, the risk neutral or risk free yield curve, free from all effects of risk factors other than pure interest rate risks. In particular, this means that the curve does not contain any default, liquidity, basis or other spread risk components. This already explains how the risk-free discount factor curve differs from the other curves: it contains none of these other risk components. Since the present value of a future cash flow decreases as the interest rate increases, almost any additional risk factor that reduces the present value of the cash flow can be taken into account by adding a premium (or spread) to the risk-free interest rate.

In order to organize this yield curve universe, a hierarchical structure is suitable in which the risk-free curve represents the base curve and is referred to in the following as $R(t, T)$, for a fixed date t. The effect of an additional risk factor on the present value of future cash flows can then generally be modeled as a spread curve (for example, $S_L(t, T)$ for liquidity risk or $S_D(t, T)$ for default risk) over the risk-free yield curve. A risky yield curve $R_{R+L+D}(t, T)$ with liquidity and default risk components is then obtained simply by adding spreads to the risk-free interest rate:

$$R_{R+L+D}(t, T) = R(t, T) + S_L(t, T) + S_D(t, T) \ .$$

Using continuous compounding for risk-free interest rates and the spreads, the corresponding discount factors have the following simple relationship (where we omitted the time arguments of interest rates and spreads for the sake of clarity)

$$B_{R+L+D}(t, T) = e^{-R_{R+L+D}(T-t)} = e^{-(R+S_L+S_D)(T-t)}$$

$$= B(t, T)e^{-S_L(T-t)}e^{-S_D(T-t)} = B(t, T)B_{S_L}(t, T)B_{S_D}(t, T) .$$

The spreads can thus also be represented as multiplicative factors $B_{S_L}(t, T)$ and $B_{S_D}(t, T)$ which, multiplied by the risk free discount factor, result in the risky discount factor. For other interest rate conventions, the conversion is not as simple. In particular, the spread factors are then no longer independent of the risk-free interest rate. For example, within discrete annual compounding, the risky discount factor becomes

$$B_{R+D}(t, T) = (1 + R + S_D)^{-(T-t)} = (1 + R)^{-(T-t)} \left(\frac{1 + R + S_D}{1 + R} \right)^{-(T-t)}$$

$$= B(t, T) \left(\frac{1 + R + S_D}{1 + R} \right)^{-(T-t)} .$$

This can lead to surprising effects. If, for example, the present value of the default risk of a zero bond is expressed as a premium over the interest rate in discrete annual compounding, this spread will change if the risk-free interest rates changes—even if the survival probability remains the same. The construction of a suitable yield curve hierarchy also facilitates the calculation of risk sensitivities, because only then a sharp separation of effects from the various risk factors such as default risk or liquidity risk is possible.

The hierarchy of yield curves proposed here also implies a sequence that should be observed when generating the yield curves:

1. Generation of the risk-free discount factor curve, which forms the basis for all other yield curves.
2. Generation of the forward curves or base spread curves used to calculate forward interest rates for different maturities; for example, one curve each for forward rates with maturities of one, three, six and twelve months. Care should be taken to start with the curves for which the most liquidly traded quotes are available. In the Euro zone, for example, it makes sense to first generate the 6M forward curve, since swaps with coupon periods of 1Y on

the fixed side against 6M on the variable side are most liquidly traded. To generate the 3M curve one can then use basis swaps with 3M variable against 6M variable. But this requires that the 6M curve is already available. In USD, on the other hand, it makes more sense to start with the 3M curve, since swaps with 6M fixed against 3M variable in USD are traded more frequently than 1Y fixed against 6M variable.

3. Generating cross-currency basis curves for all required foreign currencies to ensure that all products with foreign currency components can also be valuated.

4. Finally, all other required yield curves can be generated.

29.5 Selection of Benchmark Instruments

Two key aspects are of particular importance when selecting appropriate benchmark instruments. On the one hand, the most liquidly traded instruments available should be selected. Illiquid instruments contain little to no information about the interest rates actually traded. Since the yield curves play such a central role in all valuation issues, one cannot afford gross inaccuracies. This is especially true for the risk-free discount factor curve. On the other hand, it is necessary that all benchmark instruments belong to the same risk class. For the generation of the risk-free yield curve, only instruments can be considered which depend solely on risk-free interest rates and no other risk factor, particularly no basis or default risks. To generate a 3M forward curve, only instruments that are dependent on 3M forward rates (e.g. 3M-EURIBOR) and on interest rates that are already known can be considered. If these instruments depend, for example, on 6M forward rates, or on risk-free interest rates for discounting cash flows, then these curves must have been generated beforehand, see Sect. 29.4. If the benchmark instruments differ in their risk profile, then jumps in the yield curve can occur, which should be avoided at all costs. For the generation of a 3M yield curve, for example, 3M deposits, 3M FRAs or 3M against 6M Basis Swaps could be used. For the latter, in addition to the risk-free yield curve, the 6M curve must have been generated already(assuming liquidity of the instruments). However, the simultaneous use of 3M futures quotes is prohibited, since 3M futures contracts are secured by variation margins and therefore belong to a different risk class. It is therefore indispensable to precisely analyze the products behind the benchmark quotes in order to determine exactly all costs and dependencies on risk factors.

A third aspect, also not to be neglected, is the synchronicity of the market prices used. All prices and quotes should be determined within as short a

time period as possible. Due to the hierarchical structure, this applies to all benchmark instruments of all curves of the yield curve universe, as otherwise the spreads between the various yield curves would be incorrect due to interim market movements. This applies in particular to the cross-currency basis spread curves, which is problematic in practice because liquid market quotes are not available at the same time for all currency areas.

29.5.1   Risk Free Interest Rates

One question remained unanswered as yet: Which benchmark instruments should be used for the risk-free yield curve which is the basis for all other curves. First of all, it should be noted that no financial instrument is completely free of all but pure interest rate risks. At best, you can find liquidly traded products that come as close as possible to the ideal. Due to their liquidity, the first and foremost candidates are Overnight Index Swaps (OIS) and Repo Rates. OIS, like ordinary swaps, have a variable side and a side with a fixed coupon. However, the variable side is based on the overnight (o/n) interest rate and will be either compounded over the entire interest period (e.g. EONIA in the euro zone) or it will be averaged over the interest period (e.g. in USD area). The underlying o/n interest rate is unsecured and, like e.g. EURIBOR, is determined on a daily basis by a committee consisting of a group of banks with good credit ratings. Since a default of one of the participating banks within the short period of just one day is extremely unlikely (especially since a participating bank must resign from the committee as soon as its rating deteriorates), these swaps do not carry any significant default risk.

Repo rates are interest rates for securities lending transactions (repurchase agreements). As they are collateralized transactions, repo transactions are considered to be free of default risk. However, the securities, such as equities or bonds, used as underlying instruments for repo transactions are subject to strong price fluctuations, so that phases of underprotection are possible. A further disadvantage is that supply and demand of repo transactions on certain underlyings (e.g. to enter into short positions) can lead to a distortion of prices or repo rates. These arguments suggest that OIS rates are preferable as the best proxy for risk-free interest rates.

In fact, OIS rates have established themselves on the market as a quasi-standard for building risk-free yield curves. Repo curves can then be defined as spread curves over the OIS curve (where these spreads could also be negative). However, the differences between repo curves and OIS curves are usually much

smaller than, for example, the difference between an OIS curve and the 6M forward curve.

The situation is different if the transactions are collateralized with cash (see Sect. 20.2.1). Nowadays, unsecured transactions tend to be the exception. They have disappeared completely from interbank trading. Only in trades between banks and non-banks, states or municipalities are unsecured transactions still common. For a collateralized transaction, the interest rate on the collateral account is decisive. In fact, the collateral account then assumes the role of the risk-free bank account, i.e. the numeraire used in the derivation of the Black-Scholes equation [136, 155]. The whole derivation remains consistent and generalizable if the risk free bank account is replaced by the collateral account. As a consequence, the risk-neutral interest rate hast to be replaced by the interest rate applicable to the collateral account.

29.6 Determination of the Discount Factor Curve

In the following, we will now explicitly generate an OIS yield curve. There are two fundamentally different approaches. The classic *Bootstrapping* and the optimization procedure. Using these two methods, we will generate an OIS curve for the EUR market.

29.6.1 Bootstrapping

The classic bootstrapping begins with the benchmark instrument with the shortest maturity and successively adds another grid point to the yield curve with each additional benchmark instrument. The discount factors or zero rates are calculated in such a way that the prices of the benchmark instruments are precisely met. Benchmark instruments with more than one cash flow use the known portion of the yield curve to discount earlier cash flows. This ensures that all benchmark prices can be replicated exactly with the finished yield curve once the calculation has been completed. This, however, requires that the interpolation method used does not depend on discount factors at later grid points. This condition can be fulfilled for linear interpolation, for example, but not for spline methods, which is why spline interpolations are not consistent with bootstrapping procedures. Likewise, parametric approaches are excluded. In addition, the bootstrapping algorithm has to be adapted if a different instrument type is to be used.

Table 29.1 Benchmark instruments and resulting OIS yield curve in the form of discount factors for Bootstrapping and Nelson-Siegel-Svensson as of 28 November 2013 (the quotes are fictitious and incomplete)

Benchmark	Maturity	Grid Point [Days]	Quote	Boot-strapping	Nelson-Siegel
		0		1.000,000,00	1.000,000,00
Deposit	o/n	1	0.131%	0.999,996,36	0.999,995,95
Deposit	t/n	4	0.134%	0.999,985,19	0.999,983,95
Swap	1W	11	0.146%	0.999,956,81	0.999,956,78
Swap	1M	35	0.130%	0.999,873,26	0.999,871,97
Swap	3M	95	0.118%	0.999,687,01	0.999,707,08
Swap	6M	186	0.102%	0.999,469,80	0.999,539,38
Swap	9M	278	0.095%	0.999,262,67	0.999,400,55
Swap	1Y	369	0.089%	0.999,083,66	0.999,225,72
Swap	18M	551	0.097%	0.998,513,06	0.998,544,21
Swap	2Y	734	0.134%	0.997,272,92	0.997,129,64
Swap	3Y	1100	0.250%	0.992,402,63	0.991,242,46
Swap	5Y	1831	0.701%	0.964,891,99	0.965,522,88
Swap	7Y	2561	1.098%	0.924,215,91	0.924,038,80
Swap	10Y	3658	1.604%	0.846,676,48	0.846,414,05
Swap	15Y	5485	2.103%	0.717,100,43	0.717,476,99
Swap	20Y	7309	2.290%	0.615,928,89	0.613,790,10
Swap	30Y	10961	2.376%	0.472,776,72	0.474,268,82

For the yield curve we choose log-linear interpolated discount factors. As suggested in Sect. 29.2.2, we work around the problem of extrapolation at the short end by inserting the point $T = t$ with $B(t, t) = 1$. The benchmark instruments are listed in Table 29.1. The reference date for the calculation is Thursday, November 28, 2013. The selected benchmark instruments are not a comprehensive list of all available instruments, rather they represent a sufficiently realistic selection to demonstrate the process of yield curve generation. In particular, the "negative hump" with a minimum at one year term is not untypical. Although the Nelson-Siegel-Svensson method used can principally not hit the vertices exactly, the agreement with the bootstrapping result is very good.

The o/n deposit starts at the time t and ends on the following bank working day, in this case Friday, November 29th, one calendar day later. The quoted interest rate is $Q_{o/n} = 0.131\%$ in simple compounding with day count convention actual/360 and refers to a present value of $V_{o/n} = 1$. That is how this quote is to be interpreted. Since the deposit has only one future cash flow, this can be used directly to determine the discount factor. Thus, we get for our

first grid point $T_1 = t + 1$ day:

$$V_{o/n}(t) = 1 = B(t, T_1)\left(1 + Q_{o/n}(T_1 - t)\right)$$

$$\Leftrightarrow \quad B(t, T_1) = \frac{1}{1 + Q_{o/n}(T_1 - t)} = \frac{1}{1 + \frac{Q_{o/n}}{360}} \approx 0{,}99999636 \ .$$

The next benchmark instrument is the tomorrow-next (t/n) deposit, which starts at the time T_1 and ends at the following bank business day T_2. Because of the weekend on November 30 and December 1st that's December 2nd, 2013, three calendar days later. The value of the t/n deposit $V_{t/n}(T_1)$ at time T_1 is 1,[2] analogous to the o/n deposit. Otherwise, the quotation is to be interpreted analogously to the o/n deposit. We can thus establish the following equation:

$$V_{t/n}(t) = B(t, T_1)V_{t/n}(T_1) = B(t, T_1) = B(t, T_1)B(T_1, T_2|t)\left(1 + Q_{t/n}(T_2 - T_1)\right)$$

$$= B(t, T_2)\left(1 + Q_{t/n}(T_2 - T_1)\right)$$

$$\Leftrightarrow \quad B(t, T_2) = \frac{B(t, T_1)}{1 + Q_{t/n}(T_2 - T_1)} = \frac{B(t, T_1)}{1 + \frac{3Q_{o/n}}{360}} \approx 0{,}99998519 \ .$$

Since the t/n deposit starts in the future, we need Gl. 2.7 to calculate the desired zero bond $B(t, T_2)$.

All other benchmark instruments are EONIA swaps and start at the time T_2, which corresponds to the typical two valuta days. Swaps with terms of up to one year have only one coupon period, otherwise the coupons are paid annually both on the variable (EONIA) side and on the fixed side. EONIA swaps with maturities between 12 and 24 months (like the 18-month swap in Table 29.1) have a shortened first period. The coupon C_{EONIA} of the EONIA cash flow from T_i to T_j is calculated using compounded o/n rates of the individual interest rate periods over that time, i.e.:

$$C_{EONIA}(T_i, T_j) = \prod_{m=1}^{M-1}\left(1 + \frac{R_m(b_{m+1} - b_m)}{360}\right) - 1 \ . \tag{29.3}$$

Here, b_m with $m = 1, 2, 3, \ldots, M$, $b_1 = T_i$, and $b_M = T_j$ denote the banking days between T_i and T_j. R_m is the forward o/n rate that starts on the day b_m and ends on the following bank working day b_{m+1}. Since we have defined the EONIA swaps as our benchmark for risk-free interest, we can

---

[2]For simplicity's sake, all present values refer to a Nominal of €1.

express R_m directly by a forward discount factor without having to add an additional base spread. According to Table 2.5, the parenthesis expression in Eq. 29.3 is just equal to the inverse forward discount factor, so that:

$$C_{\text{EONIA}}(T_i, T_j) = \prod_{m=1}^{M-1} B^{-1}(b_m, B_{m+1}|t) - 1 \tag{29.4}$$

$$= B^{-1}(T_i, T_j|t) - 1 . \tag{29.5}$$

The quoted interest rate $Q_{1\text{W}}$, on the other hand, corresponds to the interest rate of the fixed side. Since the swap is usually quoted in such a way that it has zero value, the coupon on the fixed side must be equal to the coupon on the variable side. Therefore, it follows further

$$Q_{1\text{W}}(T_j - T_i) = B^{-1}(T_i, T_j|t) - 1$$

$$\Leftrightarrow \quad B(T_i, T_j|t) = \frac{1}{1 + Q_{1\text{W}}(T_j - T_i)}$$

$$\Leftrightarrow \quad B(t, T_j) = B(t, T_i)\left(\frac{1}{1 + Q_{1\text{W}}(T_j - T_i)}\right) .$$

In the last line, the equation was multiplied by the already known zerobond $B(t, T_i)$ to obtain the sought after zerobond price $B(t, T_j)$. In the same way, the discount factors for all maturities up to and including one year can be calculated.

For maturities of more than one year, several cash flows happen on both sides. Generally, we'll look at an EONIA swap with L cash flows at times T_l with $l = 1, 2, 3, \ldots, L$ and T_0 as the beginning of the first coupon period. With the exception of $B(t, T_L)$, all discount factors $B(t, T_l)$ with $l = 0, 1, 2, \ldots, L - 1$ are already known. The quotation for the L-EONIA swap is denoted as Q_{L}. Since both sides of the swap have the same present value, it follows that

$$\sum_{l=1}^{L} B(t, T_l)Q_{\text{L}}(T_l - T_{l-1}) = \sum_{l=1}^{L} B(t, T_l)C_{\text{EONIA}}(T_{l-1}, T_l)$$

$$= \sum_{l=1}^{L} B(t, T_l)\left(B^{-1}(T_{l-1}, T_l|t) - 1\right)$$

$$= \sum_{l=1}^{L} (B(t, T_{l-1}) - B(t, T_l)) = B(t, T_0) - B(t, T_L) ,$$

where Gl. 29.5 and again Gl. 2.7 were used. If we now put all terms with $B(t, T_L)$ on the same page, we continue with

$$
B(t, T_L)\,(1 + Q_{\mathrm{L}}(T_L - T_{L-1})) = B(t, T_0) + \sum_{l=1}^{L-1} B(t, T_l)\,Q_{\mathrm{L}}(T_l - T_{l-1})
$$

$$
\Leftrightarrow \quad B(t, T_L) = \frac{B(t, T_0) + \sum_{l=1}^{L-1} B(t, T_l)\,Q_{\mathrm{L}}(T_l - T_{l-1})}{1 + Q_{\mathrm{L}}(T_L - T_{L-1})}\ .
$$

$$(29.6)$$

Since all quantities on the right side are already assumed to be known, $B(t, T_L)$ can be calculated.

However, in our derivation we had assumed that all $B(t, T_l)$ with $l = 0, 1, 2, \ldots, L - 1$ were already known. If you look at Table 29.1, however, you will notice that quotes are not available for every year, especially for long-dated swaps. For example, a quote with a term of 6 years is missing. For maturities longer than 10 years there are often only quotes at 5-year intervals. We must therefore calculate the needed intermediate steps at the same time as we calculate the discount factor for the grid point $B(t, T_L)$. For the finished curve to accurately meet the benchmark prices, this must be done in a way that is consistent with the chosen interpolation method. It is therefore generally not possible to commit to an interpolation method only after bootstrapping has been completed, since the interpolation method influences the bootstrapping algorithm. This is now exemplified by an example. We calculate again the vertex $B(t, T_L)$ using a swap with Maturity T_L. To price that swap we need the discount factors $B(t, T_l)$ with $l = 0, 1, 2, \ldots, L$, where the discount factors up to and including $l = L - 2$ are already known. So, we still need $B(t, T_{L-1})$ to calculate $B(t, T_L)$. To determine $B(t, T_{L-1})$ we apply our selected interpolation procedures. With linear interpolation, for instance, we get

$$
B(t, T_{L-1}) = B(t, T_{L-2}) + \frac{B(t, T_L) - B(t, T_{L-2})}{T_L - T_{L-2}}(T_{L-1} - T_{L-2})\ .
$$

This inserted into Eq. 29.6 yields:

$$B(t, T_L) = \frac{B(t, T_0) + \sum_{l=1}^{L-2} B(t, T_l) Q_{\mathrm{L}}(T_l - T_{l-1})}{1 + Q_{\mathrm{L}}(T_L - T_{L-1})}$$

$$+ \frac{\left(B(t, T_{L-2}) + \dfrac{B(t, T_L) - B(t, T_{L-2})}{T_L - T_{L-2}} (T_{L-1} - T_{L-2}) \right) Q_{\mathrm{L}}(T_{L-1} - T_{L-2})}{1 + Q_{\mathrm{L}}(T_L - T_{L-1})}$$

$$= \frac{1}{1 + Q_{\mathrm{L}}(T_L - T_{L-1})} \left(B(t, T_0) + \sum_{l=1}^{L-2} B(t, T_l) Q_{\mathrm{L}}(T_l - T_{l-1}) \right.$$

$$+ B(t, T_{L-2}) Q_L \left(1 - \frac{T_{L-1} - T_{L-2}}{T_L - T_{L-2}} \right) (T_{L-1} - T_{L_2})$$

$$\left. + B(t, T_L) \frac{(T_{L-1} - T_{L-2})^2}{T_L - T_{L-2}} \right) .$$

Solving for $B(t, T_L)$ we get

$$B(t, T_L) = \left(1 - \frac{(T_{L-1} - T_{L-2})^2}{(T_L - T_{L-2})(1 + Q_{\mathrm{L}}(T_L - T_{L-1}))} \right)^{-1}$$

$$\left(B(t, T_0) + \sum_{l=1}^{L-2} B(t, T_l) Q_{\mathrm{L}}(T_l - T_{l-1}) \right.$$

$$\left. + B(t, T_{L-2}) Q_L \left(1 - \frac{T_{L-1} - T_{L-2}}{T_L - T_{L-2}} \right) (T_{L-1} - T_{L_2}) \right) .$$

Now all unknown quantities on the right side are eliminated and $B(t, T_L)$ can be calculated.

Finally, we will look at an example with log-linear interpolation of discount factors. The interpolated discount factor $B(t, T_{L-1})$ is according to Eq. 29.2

$$\ln B(t, T_{L-1}) = \ln B(t, T_{L-2}) + \frac{\ln \left(\frac{B(t, T_L)}{B(t, T_{L-2})} \right)}{T_L - T_{L-2}} (T_{L-1} - T_{L-2})$$

or after exponentiation

$$B(t, T_{L-1}) = B(t, T_{L-2}) \left(\frac{B(t, T_L)}{B(t, T_{L-2})} \right)^{\frac{T_{L-1} - T_{L-2}}{T_L - T_{L-2}}} .$$

With this expression inserted into Eq. 29.6, the equation can no longer be solved for $B(t, T_L)$. Thus, depending on the chosen interpolation method, calculating $B(t, T_L)$ can become quite difficult or even impossible, so that $B(t, T_L)$ can only be determined numerically, for example using a numerical root search. For classic bootstrapping this effort is only worthwhile if a multidimensional optimization of the yield curve (as described in the following Sect. 29.6.2) is ruled out for performance reasons. However, this should usually not be the case for most applications in practice. Figure 29.1 shows that the optimization error for the method presented here is minimal. This does not apply to parametric procedures, though.
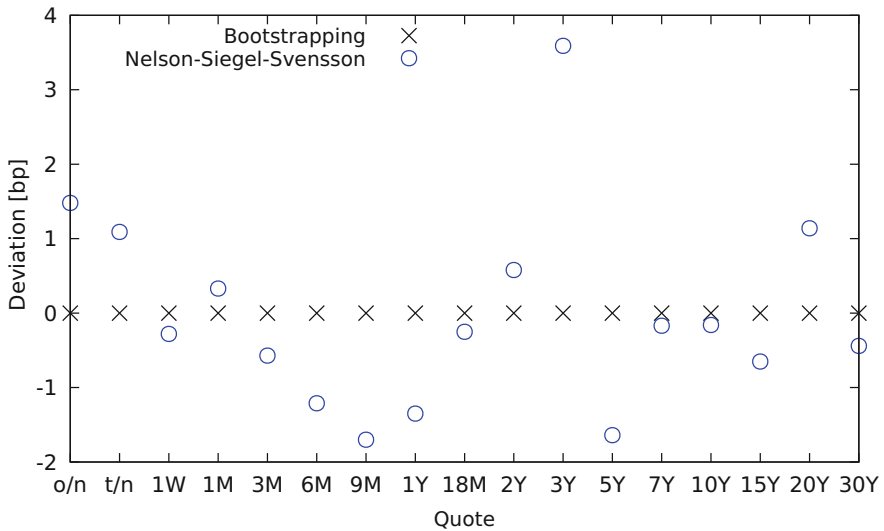
Fig. 29.1 Deviation (in basis points) of the theoretically calculated prices from the actual quoted prices. While the deviation is less than 10^{-10} base points for bootstrapping (i.e. optimization of log-linearly interpolated discount factors at fixed vertices), the Nelson-Siegel-Svensson-procedure deviates by 1 or 2 and in one case even by more than 3 basis points

29.6.2 Optimisation

A more general procedure, which works without the limitations of classical bootstrapping mentioned at the beginning of the previous section, is multidimensional optimization. This additional flexibility is achieved through higher computing costs. For many applications, however, this additional effort is hardly significant. With this optimization procedure, a yield curve is first generated, either as a parametric form or by defining individual grid points, where each benchmark instrument is assigned to exactly one grid point. In the latter case, the interest rates at the individual grid points can be regarded as parameters of the model, and the parametric form is then given by specifying the interpolation and extrapolation. Here it is also possible without any problems to create splines or apply other procedures that depend on all parameters (the interest rates at the grid points) at the same time. Plausible start values are then assigned to the parameters of the model, so that for each future time T a discount factor or a zero rate can be calculated. Finally, a function is defined that calculates the differences between the theoretical benchmark prices calculated using the modeled yield curve and the actual market prices of the benchmark instruments. Those differences are arranged in an array. With the help of a multidimensional optimizer, the parameters are varied until the square sum of all differences in that array becomes minimal. Of course, the benchmark prices will usually not be hit exactly. However, the remaining differences are in general too small to have any practical relevance, since other sources of error (such as synchronicity of the data) have a much greater influence. Also, the accuracy of the input market data is often limited to four figures only.

The procedure therefore comprises the following steps:

1. Parameterization of the yield curve to be generated, either as a parametric form or as grid points with (plausible) initial values.
2. Theoretical calculation of benchmark prices using the modeled yield curve.
3. Calculation of the sum of squares of the differences between theoretical prices and the respective market prices or quotes.
4. If the deviation is less than the allowed tolerance, then the calculation is finished because a suitable set of parameters was found. Otherwise, the calculation continues with the next step.
5. Creation of a new set of parameters.
6. The procedure continues at step 2 above.

The tricky point of this procedure is step 5. The goal is to minimize the number of recalculations of all benchmark instruments, as this step is computationally most expensive. A good optimization process will estimate a new parameter set leading to a significantly smaller square sum of the differences. The process can be visualized as searching for the lowest point or valley (the minimum) on a surface. The height of the surface corresponds to the square sum of the differences, while the model parameters determine the location on this surface. The way in which the optimization procedure determines the new parameter set for the next cycle is the essential difference between different optimization methods.

The *gradient method* calculates (numerically) the multidimensional slope (gradient) of the surface to determine the direction in which one should move through the space spanned by the model parameters in order to move "down" on the surface towards a lower square sum of differences. One problem of the gradient method is its strong dependence on the starting values. If the surface has many small troughs, there is a great danger that you will not be able to find the absolute minimum: The process gets stuck in one of these troughs, because at the bottom of the trough all gradients always point "up", because in every direction you look, the sum of squares is larger than at the bottom of the trough. Another method, called (*Simulated Annealing*), simulates with the help of Monte Carlo methods a slowly cooling gas that settles in the various troughs. This method does not have the problem of finding only one local minimum, but usually requires significantly more calculation steps. The same applies to *genetic* algorithms that simulate "mutations" in the parameters with the hope of breaking out of the local minima described above. Finding the ideal optimization method often borders on art rather than science. In any case, it is advisable to use one of the freely or commercially available optimization routines. For simple problems, the optimizer available in Excel is often sufficient. In Table 29.1, the result of such an optimization is shown for the representation of the yield curve using grid points as well as using the parametric Nelson-Siegel-Svensson method. The Excel-Workbook Bootstrapping.xlsx from the download area [50] shows an exemplary optimization result.

29.7 Forward Curves

As already mentioned several times (e.g. in Sect. 15.6.1), interest rates that are used, for example, as an underlying for the variable side of a swap, are quoted at a premium, the so-called *basis spread*, above the risk-free yield curve. In general, basis spreads are time-dependent, so they are modeled as a basis spread curve.

In addition, the amount of the basis depends on the terms of the forward rates that are to be calculated with it. Consequently, each possible term of forward rates needs their own forward, tenor or basis spread curve, like, e.g. for terms of one, three, six or twelve months.

Generating a forward curve is done analogously to the yield curve bootstrapping or optimization procedures described above (see Sects. 29.6.1 and 29.6.2). The first step is to select appropriate benchmarking instruments. The present values of these benchmark instruments may depend only on the to-be-determined forward curve and other, already known yield curves. One of these other curves is, for example, the risk-free discount factor curve. Further forward curves may also be required if, for example, basis swaps are used as benchmark instruments.

As an example, we consider the creation of EUR forward curves, assuming the EUR discount factor curve is given, i.e. all risk-free discount factors $B(t, T)$ are known or can be calculated. Since swaps with a period of six months on the variable side and one year on the fixed side are more liquid in the euro zone than other maturities on the variable side, we generate the 6M forward curve first. We designate the corresponding discount factors with $B_{6M}(t, T)$, so that the 6M forward rate $F(t, T_i, T_j)$, which starts at time T_i and ends at time $T_j = T_i + 6$ Months reads:

$$
\begin{aligned}
F(t, T_i, T_j) &= \frac{1}{T_j - T_i} \left(B_{6M}^{-1}(T_i, T_j | t) - 1 \right) \\
&= \frac{1}{T_j - T_i} \left(B^{-1}(T_i, T_j | t) - 1 \right) + s(t, T_i, T_j) .
\end{aligned} \tag{29.7}
$$

Where $s(t, T_i, T_j)$ is the basis spread for the forward rate from T_i to T_j as seen at time t. The spread is given here in simple compounding, as this corresponds to the usual quotation. Equating the two right sides yields

$$
\begin{aligned}
\frac{1}{T_j - T_i} \left(B_{6M}^{-1}(T_i, T_j | t) - 1 \right) &= \frac{1}{T_j - T_i} \left(B^{-1}(T_i, T_j | t) - 1 \right) + s(t, T_i, T_j) \\
\Leftrightarrow \quad B_{6M}(T_i, T_j | t) &= \frac{1}{B^{-1}(T_i, T_j | t) + s(t, T_i, T_j)(T_j - T_i)} \\
&= \frac{B(T_i, T_j | t)}{1 + B(T_i, T_j | t) s(t, T_i, T_j)(T_j - T_i)} \\
&= B_s(T_i, T_j | t) B(T_i, T_j | t) ,
\end{aligned}
$$

where we have introduce the "spread discount factor" $B_s(T_i, T_j|t) = 1/(1 + B(T_i, T_j|t)s(t, T_i, T_j)(T_j - T_i))$. With the basis spread $s_s(t, T_i, T_j)$ quoted in continuous instead of simple compounding we have the simple relation

$$B_s(T_i, T_j|t) = e^{-s_s(t, T_i, T_j)(T_j - T_i)}$$

with

$$B_{6M}(T_i, T_j|t) = B_s(T_i, T_j|t)B(T_i, T_j|t) = e^{-s_s(t, T_i, T_j) + R_s(t, T_i, T_j)(T_j - T_i)} \ ,$$

where $R_s(t, T_i, T_j)$ is the continuously compounded risk-free forward rate for 6 months. Thus, in continuous compounding, adding a basis spread to the interest rate is equivalent to multiplying the risk-free discount factor with the "base spread discount factor". For other conventions, this relationship is more complicated. Unfortunately, there are indeed a confusingly large number of conventions in use for quoting spreads—depending on which spread it is, a different one. Whether the basis spread curve is represented as basis spreads $s(T_i, T_j|t)$ in simple compounding, or as $s_s(T_i, T_j|t)$ in continuous compounding, or as pseudo discount factors $B_s(T_i, T_j|t)$, or if one even chooses a completely different convention, is ultimately irrelevant as long as all prices or quotes of the selected benchmark instruments are met. It is also possible to model the forward curve directly in form of the *tenor* discount curve $B_{6M}(t, T)$. But that has the disadvantage that it is no longer immediately clear which contributions to the forward rate come from the risk-free interest rate or the basis spread. If one is also interested in risk sensitivities with regard to the basis spreads, or even wants to model the basis spread as a separate, independent risk factor, the representation by $B_{6M}(t, T)$ is rather unsuitable.

Using Eq. 29.7 the present value for all benchmark instruments can now be specified. For fixed against variable swaps the equation is 15.22. For deposits or FRAs on 6-month periods, the fair interest rate can be directly calculated using the Eq. 29.7. Now the bootstrapping can start step by step, beginning with the short maturities and solving the respective equations for the unknown basis spread. If there are gaps in the benchmark instruments, so that more than one spread is unknown, the missing values for the spreads must be interpolated analogously to the procedure described in Sect. 29.6.1, in consistency with the interpolation method used for the curve. Alternatively, one can start an optimization procedure with a set of initial values and then vary the basis spreads or the parameters of a parametric approach until the differences between the calculated and quoted benchmark prices become minimal.

After the 6M curve has been generated in this way, the next forward curves can be calculated. Now, it is also possible to use quotes for basis swaps, e.g. 3M variable versus 6M variable, since the 6M curve has already been generated and therefore only the 3M curve is missing in the valuation of the underlying swaps. Of course, assuming sufficient liquidity, 3M deposits and 3M FRAs as well as swaps with 3M variable against 6M fixed can be used to generate the 3M curves. In this case, the existence of the 6M curves would not be a prerequisite.

29.8 Cross Currency Curves

For each currency with investment positions in that currency a hierarchy of yield curves as described in the previous sections is required. Each position can then be calculated with the yield curves of its respective currency. It becomes more complicated, however, when a single security involves cash flows in different currencies. As already mentioned in Sect. 16.5, such transactions are priced using modified discount curves which take the cross-currency basis into account. Here, the USD is usually regarded as the reference currency, i.e. the risk-free USD discount curve (e.g. the OIS-USD curve) is used unchanged, whereas all other discount curves are adjusted. To distinguish those adjusted curves from the original discount curves, they are preceded by the designation *CCY*. The modified EUR discount curve is then called the CCY-EUR discount curve, for example.

To calculate a CCY discount curve, we need liquid instruments that depend on the currencies concerned. For the main currencies, those are FX swaps and cross currency interest rate swaps (see Sect. 16.2), which are traded with sufficient liquidity and together cover both the short-term and long-term range. Since the USD discount curve is the only discount curve that is not adjusted, it is assumed to be known and used as the base curve. If one or both sides of the selected benchmark cross-currency swaps has coupons based on variable interest rates, the corresponding forward yield curves must already be known.

As a first step and completely analogous to Sect. 29.7, the CCY discount curves are determined via optimization methods so that all the selected benchmark instruments can be valuated. In a further Step, additional CCY discount curves, for which no forward FX transactions against USD exist, can be determined, provided that for one of the two involved currencies a CCY discount curve has already been calculated.

As we've seen in Sect. 29.5.1, for transactions with cash collateral the interest rate applicable to that collateral should be used as the risk-free interest rate. As a rule, this is the OIS rate in the collateral currency. Generalizing this to foreign currency transactions means that the collateral currency is also the base currency for the determination of the CCY curves. In fact, it may also happen that a trade in currency A is collateralized with a cash collateral in currency B. A consistent valuation would then require that the B discount curve be used as the base discount curve and all cash flows of the instrument are discounted using the CCY-B discount curve. Thus, cash collateral in foreign currency implicitly brings CCY basis risk into play. A single cross currency swap that exchanges variable cash flows in currency A against variable cash flows in currency B and is collateralized in currency C, would then have to be calculated using five yield curves!

It is not uncommon that one or both counterparties have the right to choose the currency in which cash collateral is to be provided. Depending on the situation, it might be more advantageous to use one or the other currency for collateral purposes. This right represents an option which certainly can have significant value. However, the valuation of such collateral currency options can be quite difficult [155].

29.9 Survival Probabilities

The probability that a counterparty or issuer will still be able to meet its payment obligation over a specified period of time is referred to as the survival probability.[3] As already shown in Sect. 20.1, survival probabilities are formally similar to discount factors. This can be exploited by expressing the time dependence of survival probabilities as a curve formally resembling a discount curve. This survival probability curve can then be generated from liquid instruments with an explicit default risk in the same way as a discount factor curve. For this, we will use the notation introduced in Sect. 20.3 for survival probabilities $Q(t, T)$, recovery rates R, risky zero bonds or discount factors $\bar{B}(t, T)$ and the abbreviation $\tilde{B}(t, T) = B(t, T) Q(t, T)$.

---

[3]There are also other events used to define "survival", such as filing for bankruptcy, restructuring or similar. For the determination of the present value, however, it is only important whether the contractually agreed upon cash flows actually happen or whether they are completely or partially lost.

29.9.1   CDS Spread Curves

The present value of a credit default swap (CDS) depends directly on the survival probability of the reference address, which can be seen directly from Eq. 20.12 or even better from the approximation Eq. 20.13. This is why liquid CDS quotes are ideally suited for building a curve of survival probabilities. The risk-free yield curve for calculating the discount factors must already be known. Then the recovery rate R remains as the only unknown quantity, which is however usually also quoted and thus given exogenously. Since changing the quotation to up-front payments during the "Big Bang" in the CDS market, a complete CDS quote contains the following information:

- Reference address/obligation,
- Duration,
- Up-Front-Payment (equal to the cash value of the CDS),
- Premium rate (from a list of few fixed values),
- Recovery rate.

The generation of the survival probability curve is then carried out in the same way as the generation of an interest rate curve, in which the values of an initial curve are varied until the theoretical present values of the CDS benchmarks match the up-front payments of the corresponding CDS quotes. As with the yield curves, a grid point is created for each benchmark quote at the maturity of the CDS.

In contrast to the interest rate market, in the CDS market a uniform standard has emerged for the internal presentation of survival probability curves. This standard is based on piecewise constant hazard rates, analogous to the representation of discount factor curves by means of piecewise constant forward rates. The hazard rates defined by Gl. 20.3 are completely analogues to instantaneous forward rates. The method of piecewise constant interpolation of instantaneous forward rates (or in this case hazard rates) was already described above in the section on constant interpolation.

29.9.2   Synthetic CDS Spreads

Unfortunately, CDS spreads are not available in sufficient liquidity for all reference addresses. This is due to the increasing regulation of the markets and the decline in trading in credit and credit portfolio derivatives such as *Credit Debt Obligations* (*CDO*s).

If the CDS market for certain reference addresses can no longer be regarded as active and liquid, the question arises how the CDS spread information can be generated "synthetically" by other means. An obvious approach is to use real-world default probabilities (PDs), such as those provided by rating agencies as long-term averages of actual defaults per rating class in a particular country, currency and industrial sector. A comparison of real PDs with those implied from CDS quotes shows, however, that even for liquid data, those implicit PDs are systematically higher than real PDs. In the following, we present a procedure how real PDs can nevertheless be used to determine synthetic CDS spreads.

For this approach we disassemble the "fair" or synthetic CDS spread s_{synth} into three components

$$s_{\text{synth}} = s_{\text{EL}} + s_{\text{UL}} + s_{\text{L}} \, ,$$

where s_{EL} denotes the portion of the CDS spread resulting from the *expected* loss due to default risk (expected loss spread), s_{UL} denotes the portion due to the *unexpected* loss (Unexpected Loss Spread) and s_{L} denotes an additional liquidity premium. So the assumption is that, in addition to the expected loss due to default risk expressed by s_{EL}, further surcharges, taking into account unexpected default risks and liquidity, are required to explain the traded CDS spreads. This structure will now be made more plausible.

CDS Spread Component Due to Expected Losses

The default probabilities required for the calculation of s_{EL} are calculated on the basis of historically observed actual defaults. Typically, rating agencies such as Moody's, S&P or Fitch determine and provide such default probabilities based on different criteria such as industry sector, country, currency and *rating class* (or *rating* for short). A rating class is a rough classification of default risk, where "AAA" is the best rating with the lowest default probability and "CCC" means that default is imminent (see Table 29.2).

Table 29.3 shows examples of default probabilities for maturities of up to 6 years, as provided by rating agencies. The specified PDs already include *Migration Risk*. Migration risk is the risk that the rating of a company improves or deteriorates. The table shows, for example, that a company with a BB rating has a default probability of 4.58% over a term of 3 years, or to put it another way: Of 10,000 companies with a BB rating, on average 458 will default within three years. To calculate this default probability, all companies

Table 29.2 Comparison of the rating scales of Moody's, S&P and Fitch. As each rating agency uses its own methodology, the mapping presented here is by no means accurate. A company with a Moody's rating of Aa2 therefore does not necessarily have an AA rating from S&P and Fitch

Moody's	S&P	Fitch
Aaa	AAA	AAA
Aa1	AA+	AA+
Aa2	AA	AA
Aa3	AA−	AA−
A1	A+	A+
A2	A	A
A3	A−	A−
Baa1	BBB+	BBB+
Baa2	BBB	BBB
Baa3	BBB−	BBB−
Ba1	BB+	BB+
Ba2	BB	BB
Ba3	BB−	BB−
B1	B+	B+
B2	B	B
B3	B−	B−
Caa-C	CCC+	CCC+

Table 29.3 Example of expected average default probabilities in % for different rating classes. The probability indicates the percentage of companies in a respective rating class, that are in default after one, two, etc. years

	1Y	2Y	3Y	4Y	5Y	6Y
AAA	0	0.015	0.051,667	0.098,333	0.150,333	0.307,5
AA+	0	0.02	0.02	0.068,667	0.107	0.198,5
AA	0.006,667	0.018,667	0.114,667	0.284,667	0.412	0.555
AA−	0.046	0.106,333	0.166	0.232,333	0.285,667	0.541,5
A+	0.046,667	0.158,333	0.337,333	0.500,333	0.676	0.936,5
A	0.068,667	0.230,667	0.406	0.623,333	0.802,667	0.990,5
A−	0.110,667	0.261,667	0.450,667	0.595,667	0.816	1.13
BBB+	0.147	0.369,333	0.645	0.939,667	1.277	1.63
BBB	0.172,667	0.597,333	1.069,333	1.671,667	2.185,333	2.406,5
BBB−	0.377,333	1.065	1.876,333	2.722,333	3.617,333	4.346
BB+	0.771,333	2.129	3.639	5.174,333	6.237,333	7.399,5
BB	0.823,667	2.606	4.587,333	6.440,333	8.091	9.448
BB−	1.482	3.889,333	6.671	9.046,667	11.004,67	16.278,5
B+	2.033,333	5.890,333	9.687,333	12.899	15.066,67	21.458
B	4.050,333	9.258	13.812,33	17.045,67	19.637	26.967,5
B−	5.964,333	11.857,67	16.975,67	20.702,33	23.66	35.082,5
CCC+	22.772	32.105	38.503	41.605,67	44.595,67	52.134,5

are counted, which had a BB rating to start with and then were defaulted after three years, regardless of which rating they had in the meantime (e.g., after a year or two). Typically, most defaulting companies will see one or more deteriorations of their rating in the years and months prior to the default. The curve for survival probabilities $Q(t)$ can now be calculated from the grid points $Q(1Y) = 1 - PD(1Y)$, $Q(2Y) = 1 - PD(2Y)$ etc., using the log-linear interpolation method. Thus, the logarithms of the survival probabilities are interpolated linearly (analogous to the discount factors in 29.2). The expected loss spread s_{EL} can then be calculated as a Par CDS spread according to Eq. 20.14, or using Approximation 20.15.

CDS Spread Component Due to Unexpected Losses

The calculation of the unexpected loss spread is based on the idea that *economic capital* (or *EC* for short) must be assigned to each asset to cover the risks associated with that asset. The amount EC of additional capital required corresponds to the amount of the unexpected loss. The provision of this EC induces costs (cost of capital or COC), which are generally higher than the risk-free interest rate. The assumption is now, that these additional costs are passed on to the protection buyer, which increases the credit spread. This is then the unexpected loss component s_{UL} of the synthetic spread s_{synth}:

$$s_{UL} = EC \cdot COC .$$

EC is expressed as a percentage of the nominal insured by the CDS. The regulatory capital RE can be used as the lower limit for the required capital. This limit is estimated as

$$RE = RWA \cdot T1Q$$

Here, T1Q denotes the Tier 1 quota and RWA the (*Risk Weighted Assets*) according to the regulatory formula from the Basel II regulations [8]. According to this framework, RWA is a function of the expected loss, the maturity of the asset and the counterparty class to which the reference address belongs. This formula can also be used in our context. The cost of capital COC should be the long-term average return on equity of companies belonging to the same region and industrial sector as the reference address. The capital is thus assumed to cost as much as it could generate return if used otherwise. In reality however, the cost of capital varies considerably from company to company, even within

the same industry sector and the same region. For the fair value analysis carried out here, it nevertheless appears appropriate to use an average COC to arrive at something like a "generally accepted" cost of capital.

Liquidity Premium on Top of the CDS Spread

The third component of the synthetic CDS spread is the liquidity spread s_{L}. This spread can be interpreted as an additional fee that counterparties charge each other, even in liquid markets, as compensation for the comparatively lower liquidity of derivatives markets. The liquidity spread is usually rather small (about 0–30 basis points), depending on the complexity of the derivatives. The assumption of a flat, constant spread is often sufficient here.

Calibrating the Parameters

A way to determine the COC parameter is based on macroeconomic considerations, such as the average expected return on capital of companies in a sector. Another option would be to calibrate to a basket of liquid CDS quotes and extrapolate to other reference addresses. The liquidity spread can either also be calibrated or simply be fixed. As a rule, the effect of the COC dominates the liquidity spread by far.

   The procedure presented here certainly has clear weaknesses. The uncertainty in determining COC is not negligible, as the result of the COC calibration often depends very much on the market situation and the selection of the benchmark portfolio. An even greater inaccuracy lies in the use of rating data. Ratings change much less frequently than credit spreads and only represent a comparatively rough grid. This is reflected, for example, in the fact that quite often companies with identical ratings and in the same industry sector and currency region are traded at very different par CDS spreads. However, if liquidly traded CDS spreads are not available, the approach shown here provides a practicable method for the determination of synthetic CDS spreads.

29.10   The Old Yield Curve World

Prior to 2007, it was common practice to use exactly one discount curve per currency. Basis and default risks have often only been considered for such products whose explicit purpose was the trading of such risks; like for example

basis swaps (Sect. 17.1) or credit default swaps (Sect. 20.6). That was possible, since basis and credit spreads were very small (at least in trading between banks) and often could be neglected compared to other risks. Since about 2007 however, this has not been the case anymore and it doesn't seem likely that the spreads will ever go back to the low levels of before.

The methods for generating a yield curve today and before 2007 are fundamentally not very different. The essential difference is the selection of the benchmark instruments used to construct the yield curve. For example, prior to 2007 it was common to use deposits of different maturities up to one year at the short end of the yield curve—occasionally supplemented by money market futures. For maturities longer than two years, plain vanilla par swap rates were used. Therefore, the curve was often been named *swap curve*. For some markets, swap rates are quoted with different period lengths for the coupons on the fixed and variable side. In that case, the more liquid quotes were chosen for the curve construction. For the euro area for example, these are swap rates with one year on the fixed side and six months on the variable side.

However, this approach is problematic if the base spreads are significant compared to the interest rates. Between short-running o/n- and t/n-deposits (overnight and tomorrow-next, with start time today or the next trading day and maturity and repayment one trading day later) and the 6M forward rates (e.g. 6M-EURIBOR) underlying the swap rates, the basis spread cannot be neglected. In the middle segment of the curve, deposits with maturities between one week and one year exist. Money market futures usually refer to a three-month rate (e.g. 3M-EURIBOR). But the reference rate, e.g. the 3M-EURIBOR, refers to an unsecured transaction, while the futures themselves are always collateralized. These multiple changes in the interest period of the forward rates lead to jumps in the yield curve due to the different basis spreads, which are unacceptable for many applications. In addition, even the valuation of simple products such as swaps based on a swap curve, which is being used both for the discounting of the cash flows as well as for the calculation of the forward rates, is no longer in line with the market practice. Therefore, OIS discounting should be used as much as possible.

29.11   The Future of Yield Curves

The currently still predominant term rates used as underlying for the floating swap's leg were originally based on rates like EURIBOR, USD LIBOR or other "*IBOR" rates. Here, IBOR stands for interbank offer rate, the rate at which

one bank could borrow money from another bank. However, these rate indices are not based on real transaction. Instead, they are calculated as the average of estimates submitted by a pool of leading banks in the respective currency market.

As an estimate, these rate indices can not be easily verified by an independent party. In the aftermath of the financial crises of 2008, traded and quoted rates significantly diverged. As later turned out, this was, at least partially, due to manipulation of the rate estimates. Small changes of the quoted IBOR rates could have significant effects on some bank's swap portfolio, because the swap pay-offs depend directly on this fixed quotes of IBOR rates. This has been abused by some of the involved traders to manipulate the IBOR rates.

The method for the calculation of IBOR rates have since than been modified in order to prevent these manipulations. Still, regulators seek to replace all IBOR rates by transaction and OIS based rates. OIS or overnight index swap rates are those rates, at which banks lend each other money overnight. Though these overnight rates are based on uncollateralized transactions, the credit risk is very small because of the short lending period. In addition, they are based on averaged real transactions rather than estimates.

Switching from IBOR rates to OIS based term rates (called ATR or alternative term rates) will have a significant impact on how term rates have to be calculated and fixed. Various methods of calculating term rates based on overnight rates are under discussion, involving compounding, average, and different fixings methods. Most of these methods will require the application of convexity adjustments (see Sect. 14.5), and the calculation of ATRs will be more complex. On the other hand, if all term rates would be based on overnight rates, the world would become simpler with respect the fact that we would be back in a world with only one interest rate curve per currency.

30

Volatility

30.1 Implied Volatilities

30.1.1 Smiles and Volatility Indices

Risk neutral or *implied* volatilities are required for the valuation of derivatives with optionalities, while real world volatilities are used for market risk calculations. A volatility is called "implied" if it is calculated implicitly from quoted prices of liquidly traded options. The reader may wonder, why implied and real world volatilities differ at all, since a major result of the considerations in Chap. 7 was that in the Black Scholes PDE (Eq. 7.8) the real drift has been replaced by the risk free interest rate, but not so the volatility, which remains unchanged. Therefore, the difference between real world and implied volatility is no consequence of the theory. Nevertheless, this difference is observed in practice. One reason for the difference is the fact that real world volatilities are determined in hindsight based on the historic time series of observed actual prices, while implied volatilities are *forward looking*, since they are implied from options with expiry dates in the future. Another possible explanation is that actually traded option prices include add ons or discounts to incorporate other risks like liquidity or cluster risks (caused by one-sided demand) that are difficult to measure. With all other parameters fixed, these add ons and discounts will impact the calculated implied volatility. Whatever the reasons for the difference are, for risk neutral pricing implicit volatilities have to be used, since liquidly traded plain vanilla options are used to hedge Vega risk (i.e. the risk of changing volatility) caused by structured financial products.

Implied volatilities are extracted from liquidly traded plain vanilla option prices which, because of their liquidity, have usually small bid-ask spreads which improves accuracy of the calculated volatility. Given these prices, it is possible to calculate the implied volatilities, if all other parameters like interest rates, underlying price, etc., which enter the standard market option price function, are well known. The price function depends on the market but is often of Black-Scholes type or some of its variants.[1] The procedure is to vary the volatility, keeping all other parameters constant, untile the price function yields the same price as the traded market price. This is the *implied volatility*. If all market makers in a certain market segment agree on what price function to use and if all other parameters besides the volatility are easily observable, we have a 1:1 relationship between option price and implied volatility: If either one entity is known, the other can easily be calculated by means of the market price function. Therefore, in some markets, it is common to quote the implied volatilities instead of the option prices. For example, this is the case for swaptions, caps/floors or FX options and also for options on stocks or stock indices. For these liquidly traded, simple options, the Black-Scholes or Black-76 formula should be considered as a mere quotation vehicle that allows to transform implied volatilities into prices and vice versa rather than a pricing model, since the traded prices are just the result of offer and demand anyway.

Often, the assumptions on which the pricing models are based (for example, the assumption of continuous trading) are satisfied reasonably well for options *at* the money, but not for options which are far in or out of the money. This has the effect that the pricing model yields different implied volatilities for options that are not at the money. As a consequence, if the implied volatilities are plotted as a function of the strike price, the curve obtained is not the graph of a constant function as assumed by the Black-Scholes model. For most options we observe the effect that the implied volatilities increase with increasing distance from the at-the-money strike (in other words, when the option is increasingly in or out of the money); the resulting volatility curve tends to look like a *smile*. This expression has come to be the accepted term for describing the implied volatility as a function of the *strike* and is independent of the form of the curve. In general, the strike dependence is not symmetric around at the money, which is expressed by the term *skew*.

---

[1]For interest rates, the Bachelier model is nowadays also used, since it assumes a normally distributed underlying and therefore allows for negative interest rates.

Options with different *expiries* also show different implied volatilities even if the strike is the same. In analogy to interest rates, the dependency of the volatilities on the time to option expiry is known as the *volatility term structure*.

The variables expiry date and strike price are *absolute* variables and as such are not very suitable for the comparison of the implied volatilities of different options on the same underlying; the variables *time to* expiry and *relative* strike price are preferable. The relative strike is the strike expressed as a percentage of the current underlying price. A call with a strike of 84 euros on an underlying with a spot price of 80 euros is 5% out of the money and therefore has a relative strike of −5%. A corresponding put would have a relative strike of +5%. This relative strike is called *moneyness*. The moneyness is thus defined as the relative distance between the strike K and the at-the-money point

$$\text{moneyness} = \frac{K - S(t)}{S(t)} \ ,$$

where $S(t)$ is the current spot price[2] of the underlying. The implied volatility of options on the same underlying (for example, the DAX) as a function of time to expiry and relative strike is called the *volatility surface.* The volatility surfaces obtained from liquid option prices can be used to price illiquid options or OTC contracts on the same underlying.

For options on interest rates or interest rate instruments such as a cap or an option on Bund futures, the implied volatility depends on a third parameter. This is the time to expiry of the underlying. For such options, a complete volatility surface is determined by the volatility as a function of the options' expiries and the maturities of the underlying as well as the options' relative strikes.[3] A typical example are two caps on distinct variable interest rate indices, but otherwise identical parameters. One cap my refer to a quarterly rate, e.g. 3M-EURIBOR, the other on a semi-annual rate like 6M-EURIBOR. These two caps have to be considered as different products and will have indeed different prices resp. implied volatilities.

Building such structure requires many liquid options. Therefore, these structures usually contain a large number of gaps. The gaps on the volatility structure of an underlying, i.e., the implied volatilities for the parameter

---

[2]Sometimes the at-the-money-point is defined to be at the current *forward* price $S(t, T)$ of the underlying, where T is the expiry of the option (*at-the-money forward*). Correspondingly, the moneyness is then defined as the relative distance between the strike and this forward price.

[3]It is also possible to consider underlyings with different expiries at distinct *underlyings*. Then, there is per underlying one volatility structure depending on the option parameters strike and option expiry only.

combination for which no liquid options are available, are usually closed with clever (or courageous!) interpolation schemes.

An example of an implied volatility published by an exchange, a *volatility index,* is the *VDAX* published since December 5, 1994 by the DTB (later renamed Eurex). This index refers to one single point of the volatility structure of the DAX, namely the volatility of the DAX for a (fictitious) at the money option with a time to expiry of 45 days. Not the current DAX value is considered at the money but rather the forward price of the DAX in 45 days. This forward price is available since the prices of DAX futures contracts are known (or can be determined from the prices of DAX options using the put-call parity Eq. 6.8). The volatility of the fictitious ATM option is then determined as follows:

- For each expiry date, only those option series are considered whose strikes are at most 100 points away from the ATM point. For a series of options with strikes increasing in 25 point increments, this allows a maximum of 16 options (eight calls and eight puts) per expiry date.
- The theoretical option price is calculated using the Black-Scholes Eq. 8.7 with the same volatility for all options with the same expiry date. This volatility is then varied as an input parameter until the sum of the squared difference between the theoretical option prices per expiry date attain a minimum value (*least squares fit*). In this way, a fitted implied volatility for each expiry date is obtained.
- The actual VDAX at time t is then obtained by interpolating the implied volatilities for the two surrounding expiries T' and T'' of the expiry given by $T = t + 45$ as follows

$$\text{VDAX}(t) = \sqrt{\frac{T'' - T}{T' - T}\sigma^2(T') + \frac{T - T'}{T'' - T'}\sigma^2(T'')}$$

with $T' \leq T < T''$ and $T = t + 45$ days.

30.2 Local Volatility Surfaces

The volatility dependency on the strike is not consistent with the Black-Scholes model, since the underlying process should not depend on the strikes at which options on this underlying are traded. One approach to deal with this inconsistency is to make volatilities local, i.e. depending on the underlying

price as well as on the time to expiry. Then, the underlying would follow a log-normal process only approximately. This idea was first published by Dupire [58]. Therefore, local implied volatilities are sometimes called Dupire-Volatilities.

For general underlyings modeled as Ito processes of the form 2.19 in Sect. 2.4, the transition probabilities and the *Dupire volatilities*, also referred to as *local volatilities*, can be derived by differentiating the option price. The reader is advised to become familiar with Sects. 2.4 and 7.2 as the material in these sections are essential for a sound comprehension of what follows.

30.2.1 Implicit Transition Probabilities

The central relation holding for all Ito processes of the form 2.19 between the transition probability p defined in Eq. 2.33a and the price V of a financial instrument (an option, for example) has already been indicated in Eq. 7.16 of Sect. 7.2:

$$V(S, t, T) = B(t, T) \int_{-\infty}^{\infty} f(S', T) p(S', T \,|\, S, t) dS' \,, \qquad (30.1)$$

where the function $f(S', T)$ corresponds to the payoff of the financial instrument at expiry T. This equation can be inverted to determine the transition probabilities p from the prices of liquid options observed in the market; in doing so we determine the *implied transition probabilities*. We assume here that the market prices of a series of liquid plain vanilla options (both calls and puts) with different strikes K and expiries T are available.

For a call with a payoff profile $f(S', t' = T) = \max(S' - K, 0)$, Eq. 30.1 becomes

$$V_{\text{call}} = B(t, T) \int_{K}^{\infty} (S' - K) p(S', T \,|\, t, S) dS' \,.$$

The price of the call is now differentiated with respect to the limit of integration K for the purpose of extracting p from the integral. This yields the integrand evaluated at the (lower) limit of integration $S' = K$ and, since K itself appears in the integrand, a second term arises, involving the integral of the derivative of the integrand with respect to the strike:

$$\frac{\partial V_{\text{call}}}{\partial K} = -B(t, T) \underbrace{(K - K)}_{0} p(K, T \,|\, S, t) - B(t, T) \int_{K}^{\infty} p(S', T \,|\, t, S) dS' \,.$$

Thus

$$\frac{\partial V_{\text{call}}}{\partial K} = -B(t, T) \int_K^\infty p(S', T \,|\, t, S) dS' . \tag{30.2}$$

For a plain vanilla put with a payoff profile $f(S') = \max(K - S', 0)$ we proceed analogously to obtain

$$\frac{\partial V_{\text{put}}}{\partial K} = +B(t, T) \int_{-\infty}^K p(S', T \,|\, t, S) dS' .$$

A further derivative with respect to K yields the same result for both puts and calls:

$$\frac{\partial^2 V}{\partial K^2} = B(t, T) p(K, T \,|\, t, S) . \tag{30.3}$$

This can now be easily solved for the transition probability p[4]

$$p(K, T \,|\, t, S) = B^{-1}(t, T) \frac{\partial^2 V}{\partial K^2} . \tag{30.4}$$

In this way, the implicit transition probabilities can (at least theoretically) be obtained from the observed option prices (with fixed t and S, the strike and maturity date being variable). Through Eq. 30.4 we can determine (at time t and known S) the transition probability p associated with the points (K, T) for which call or put prices exist and thus (after an appropriate interpolation between these points) an *implied transition probability surface.* Noted that not only the price but also the *second derivative* of the price function with respect to the strike must be known. Since the numerical determination of derivatives on the basis of measured data is quite prone to error, this method can only be implemented in practice if an analytic formula for the observed option prices as a function of the strike price is available; the derivative with respect to the strike can then be performed analytically. For example, the derivative with respect to the strike price can be calculated from the Black-Scholes equation

---

[4]Observe that in the derivation of this equation we assumed certain (plain vanilla) payoff profiles for the options but we did not assume any specific underlying process. Thus, everything is valid for plain vanilla calls and puts on arbitrary underlyings of the form 2.19.

for the price of a call[5]

$$\frac{\partial V_{\text{call}}}{\partial K} = -B(t, T)N(x - \sigma\sqrt{T - t})$$

with x as in Eq. 8.5. The second derivative is then

$$\frac{\partial^2 V}{\partial K^2} = \frac{B(t, T)}{K\,\sigma\sqrt{T - t}}N'(x - \sigma\sqrt{T - t})\,.$$

Because of the put-call parity, the second derivative of the put price can immediately be calculated by taking the derivative with respect to K on both sides of the equation $\frac{\partial V_{\text{put}}}{\partial K} = \frac{\partial V_{\text{call}}}{\partial K} + B(t, T)$ this shows again that the second derivative with respect to K of the put and call are identical. Hence, for the Black-Scholes case, the implicit transition probabilities are given by

$$p(K, T \,|t, S) = \frac{\exp\left\{-\frac{1}{2}\left(x - \sigma\sqrt{T - t}\right)^2\right\}}{K\sqrt{2\pi\sigma^2(T - t)}}\,. \tag{30.5}$$

The observed option prices no longer appear explicitly in this expression but the volatilities do. These are the implied Black-Scholes volatilities which yield the observed option price when substituted into the Black Scholes pricing formula. The implied Black-Scholes volatilities are nothing other than the observed option prices quoted in another form; the Black-Scholes formula is the mapping between the two ways of quoting (on the one hand the option price, on the other the implied volatility). In the Excel workbook IMPLIED-VOLANDTRANSPROB.xlsm from the download section [50], the computation of the implied Black-Scholes volatilities from quoted ODAX options (from February 2000) are explicitly performed by means of the *Newton-Raphson method* along with the subsequent calculation of the transition probabilities.

In this way, we have constructed the implicit transition probabilities enabling, theoretically, the determination of the price of other (OTC) options via Eq. 30.1. However, we need a range of S' (the first argument of p in Eq. 30.1) from $-\infty$ to ∞. The first argument of p corresponds to the strike, when determining p via Eq. 30.4. Therefore we need traded options for strikes

[5]Similar derivatives have already been calculated in Sect. 12.5.2.

ranging from $-\infty$ to ∞ and thus cannot proceed without extrapolation arguments.

Observe that Eq. 30.5 agrees exactly with the transition probability 2.38 for the process 2.24 given earlier. Thus, everything is consistent. This can be seen by making the obvious associations $t' = T$ and $S' = K$ in 2.38 and substituting the risk-neutral drift $\widetilde{\mu}$ (see Eq. 9.25 in connection with Eq. 2.26):

$$\widetilde{\mu} = \frac{1}{T-t} \ln \left(\frac{B_q(t,T)}{B(t,T)} \right) . \tag{30.6}$$

30.2.2 Implicit Local Volatility Surfaces

The *local volatilities* can be determined in a similar manner as illustrated above for finding the implicit transition probabilities from observed market prices of derivatives. These are to be distinguished from the Black-Scholes volatilities (which are nothing other than a way of quoting option prices where the mapping from one quotation to the other is defined by the Black-Scholes formula). The method introduced here goes back to a procedure developed by Dupire. *Local volatilities* are therefore also referred to as *Dupire volatilities.*

We again assume that market prices $V(S, t, K, T)$ of a series of plain vanilla options (calls and puts) on an underlying with price $S(t)$ are known at time t for various strikes K and expiries $t' = T$. Starting from the central, general equation 30.1, we now take the derivative with respect to T and then use the forward equation 2.34 to substitute for $\partial p / \partial T$. For example, for a call with a payoff profile given by $f(S', T) = \max(S' - K, 0)$:

$$\frac{\partial V_{\text{call}}}{\partial T} = \frac{\partial B(t,T)}{\partial T} \frac{V_{\text{call}}}{B(t,T)} + B(t,T) \int_K^\infty (S' - K) \frac{\partial p(S', T \,|\, S, t)}{\partial T} dS'$$

$$= \frac{\partial \ln B(t,T)}{\partial T} V_{\text{call}}$$

$$+ B(t,T) \int_K^\infty (S' - K) \left[\frac{1}{2} \frac{\partial^2}{\partial S'^2} \left[b^2(S', T) p \right] - \frac{\partial}{\partial S'} \left[a(S', T) p \right] \right] dS'$$

$$= \frac{\partial \ln B(t,T)}{\partial T} V_{\text{call}} + B(t,T)(\frac{1}{2} I_1 - I_2) .$$

The derivative with respect to the variable of integration in the integrand suggests to solve the integrals I_1 and I_2 by integrating by parts:

$$I_1 \equiv \int_K^\infty (S'-K) \frac{\partial^2 (b^2 p)}{\partial S'^2} dS'$$

$$= (S'-K) \frac{\partial (b^2 p)}{\partial S'} \Big|_K^\infty - \int_K^\infty \frac{\partial (S'-K)}{\partial S'} \frac{\partial (b^2 p)}{\partial S'} dS' \ .$$

A reasonable assumption would be that, starting from S, the probability p (and its derivatives) for reaching S' in a finite time $(T-t)$ should converge towards zero very fast for $S' \to \infty$. The upper limit $S' = \infty$ of the first term on the right-hand side then makes no contribution to the evaluation of the integral. The expression evaluated at the lower limit $S' = K$ is obviously equal to zero as well. Since $\partial (S'-K)/\partial S' = 1$, the above expression for I_1 reduces to

$$I_1 = - \int_K^\infty \frac{\partial (b^2 p)}{\partial S'} dS' = - b^2 p \Big|_K^\infty = b^2(K,T) p(K,T \,|\, S,t) \ ,$$

where we have again made use of the fact that $p(S',T \,|\, S,t)$ converges towards zero very fast as $S' \to \infty$ and that therefore the upper limit contributes nothing to the integral. Proceeding analogously with the second integral I_2 yields after integration by parts

$$I_2 \equiv \int_K^\infty (S'-K) \frac{\partial (ap)}{\partial S'} dS' = \underbrace{(S'-K) ap \Big|_K^\infty}_{0} - \int_K^\infty \underbrace{\frac{\partial (S'-K)}{\partial S'}}_{1} a\, p\, dS'$$

$$= - \int_K^\infty a(S',T) p(K,T \,|\, S,t) dS' \ .$$

In total, $\partial V / \partial T$ becomes

$$\frac{\partial V_{\text{call}}}{\partial T} = \frac{\partial \ln B(t,T)}{\partial T} V_{\text{call}}$$

$$+ B(t,T) \left[\frac{1}{2} b^2(K,T) p(K,T \,|\, S,t) + \int_K^\infty a(S',T) p(K,T \,|\, S,t) dS' \right] \ .$$

Analogously for puts

$$\frac{\partial V_{\text{put}}}{\partial T} = \frac{\partial \ln B(t, T)}{\partial T} V_{\text{put}}$$
$$+ B(t, T) \left[\frac{1}{2} b^2(K, T) p(K, T \mid S, t) - \int_{-\infty}^{K} a(S', T) p(K, T \mid S, t) dS' \right],$$

where an analogous assumption has been made that $p(S', T \mid S, t)$ for $S' \to -\infty$ converges towards zero fast enough and that, consequently, the lower limit makes no contribution to the integral.

As yet, our deliberations hold for every general Ito process of the form 2.19. From now on, however, an explicit stochastic process must be assumed in order to continue with the computation. For the special case of a *risk-neutral* random walk, i.e., if the parameters are chosen as in Eq. 7.1 with a volatility $\sigma = \sigma(S, t)$ and a risk-neutral drift, Eq. 30.6, we obtain for the call

$$\frac{\partial V_{\text{call}}}{\partial T} = \frac{\partial \ln B(t, T)}{\partial T} V_{\text{call}} + B(t, T) \frac{1}{2} \sigma^2(K, T) K^2 p(K, T \mid S, t)$$
$$+ \frac{B(t, T)}{T - t} \ln \left(\frac{B_q(t, T)}{B(t, T)} \right) \int_{K}^{\infty} p(S', T \mid S, t) S' dS'$$

and for the put

$$\frac{\partial V_{\text{put}}}{\partial T} = \frac{\partial \ln B(t, T)}{\partial T} V_{\text{put}} + B(t, T) \frac{1}{2} \sigma^2(K, T) K^2 p(K, T \mid S, t)$$
$$- \frac{B(t, T)}{T - t} \ln \left(\frac{B_q(t, T)}{B(t, T)} \right) \int_{-\infty}^{K} p(S', T \mid S, t) S' dS' .$$

Using Eqs. 30.1 and 30.2, the remaining integral for the call can be written as

$$\int_{K}^{\infty} p(S', t' \mid S, t) S' dS' = \int_{K}^{\infty} p(S', t' \mid S, t)(S' - K) dS'$$
$$+ K \int_{K}^{\infty} p(S', t' \mid S, t) dS'$$
$$= B^{-1}(t, T) V_{\text{call}} - K B^{-1}(t, T) \frac{\partial V_{\text{call}}}{\partial K}$$

and analogously for the corresponding integral for the put

$$\int_{-\infty}^{K} p(S', t' \,|\, S, t) S' dS' = -\int_{-\infty}^{K} p(S', t' \,|\, S, t)(K - S')dS'$$

$$+ K \int_{-\infty}^{K} p(S', t' \,|\, S, t)dS'$$

$$= -B^{-1}(t, T)V_{\text{put}} + K B^{-1}(t, T)\frac{\partial V_{\text{put}}}{\partial K} .$$

The remaining p is replaced by the expression in Eq. 30.4, which holds for both puts and calls, to obtain the same differential equation for call and put prices with respect to the variables K and T

$$\frac{\partial V}{\partial T} = \frac{\partial \ln B(t, T)}{\partial T} V + \frac{1}{2}\sigma^2(K, T) K^2 \frac{\partial^2 V}{\partial K^2}$$

$$+ \frac{1}{T - t} \ln\left(\frac{B_q(t, T)}{B(t, T)}\right)\left(V - K\frac{\partial V}{\partial K}\right) . \tag{30.7}$$

For continuous compounding this simplifies further to

$$\frac{\partial V}{\partial T} = \frac{1}{2}\sigma^2(K, T) K^2 \frac{\partial^2 V}{\partial K^2} - (r - q)K\frac{\partial V}{\partial K} - qV .$$

This equation can be easily solved for $\sigma^2(K, T)$ thereby allowing implicit local volatilities to be calculated from the option prices observed in the market. As for the implicit transition probabilities described above, Eq. 30.7 can be used (at time t for known values of S) to determine the value of $\sigma(K, T)$ for all points (K, T) for which the price of a call or put is observable in the market; an appropriate interpolation yields a *local volatility surface*. Of course, we need more than just the option price to complete this construction; the derivatives with respect to the strike price and the expiry must be known as well. Here we *always* must rely on the *numerical* determination of these derivatives. This is because σ (as opposed to p) itself appears in any analytical pricing formula (as for example in the Black-Scholes formula) and substitution of such price

formulas into the above equation thus ultimately just yields the trivial[6] identity "$\sigma = \sigma$". As mentioned before, such numerical computations of (second!) derivatives are prone to error.

The volatilities obtained through numerical differentiation of the option prices appearing in Eq. 30.7 are particularly unstable for options that are either well in or well out of the money since the (especially error prone!) second derivative $\partial^2 V / \partial K^2$ is quite close to zero for these values of K. Since isolating $\sigma^2(K, T)$ involves dividing by this very small (and error prone) second derivative, extremely large errors can be expected for the resulting local volatility. We can avoid at least this problem (i.e., the problem of dividing by a very small variable) by expressing the option prices in terms of their implied Black-Scholes volatilities $\widehat{\sigma}$ (which are just a different quotation of option prices). We then obtain an expression for the local volatility with a denominator of the form $1 + \varepsilon$ with $\varepsilon > 0$. For example, for continuous compounding and with x as given in Eq. 8.5

$$\sigma^2(K, T) = \frac{\widehat{\sigma}^2 + 2(T - t)\widehat{\sigma}\frac{\partial\widehat{\sigma}}{\partial T} + 2rK(T - t)\widehat{\sigma}\frac{\partial\widehat{\sigma}}{\partial K}}{\left[1 + Kx\sqrt{T - t}\frac{\partial\widehat{\sigma}}{\partial K}\right]^2 + K^2(T - t)\widehat{\sigma}\left[\frac{\partial^2\widehat{\sigma}}{\partial K^2} - x(\frac{\partial\widehat{\sigma}}{\partial K})^2\sqrt{T - t}\right]} .$$

But even in this form the local volatility remains very instable as long as the derivatives are approximated numerically from market data.[7] Practitioners usually attempt to smooth the Black-Scholes volatility surface by using one of several approaches. One example is to fit an analytic function (for example, a polynomial) $u(K, T)$ to the observed Black-Scholes volatilities. The derivatives of this function can then be calculated analytically when constructing the *Dupire local volatility surface.*

Having constructed the local volatility surface, the valuation of further options (OTCs) can be accomplished with a finite difference scheme, for example, as described in Chap. 10. The Excel workbook

---

[6]This can be shown quite easily for continuous compounding with the help of Eqs. 8.6, 30.2 and 30.3. The one additional differentiation needed is

$$\frac{\partial V_{\text{call}}}{\partial T} = -qe^{-q(T-t)}SN(x) + re^{-r(T-t)}KN(x - \sigma\sqrt{T - t})$$

$$+ e^{-r(T-t)}\frac{\sigma(K, T)}{2\sqrt{T - t}}K N'(x - \sigma\sqrt{T - t}) .$$

[7]Such a numerical approximation is done by approximating the differential quotient with a difference quotient, i.e. by $\frac{\partial V}{\partial K} \approx \frac{V_1(K_1) - V_2(K_2)}{K_1 - K_2}$, for instance, where V_1 and V_2 are two observed market prices, etc. An analogous procedure is performed for the Black-Scholes volatilities.

FINITEDIFFERENCEMETHOD.XLSM from the download section [50] is structured such that it is immediately obvious at what stage a volatility surface enters into the computation.

As an alternative to the Dupire approach, Andersen and Brotherton-Radcliffe have published a method which calibrates the local volatility surface directly on the grid [3].

30.3 Volatility Transformations

30.3.1 Transformation Between Relative and Absolute Volatility

The volatility of a risk factor with a price $S(t)$ at time t is usually measured as the standard deviation (per year) of the *logarithmic* price changes $\delta \ln S(t)$ (e. g., implicitly extracted from the option prices as discussed in Sects. 30.1 and 30.2 or else through the analysis of historical time series as in Chap. 31) or obtained from data providers. For the *absolute* risk, in other words, for the actual monetary amount of a potential loss, the change of the price itself is needed and not its logarithm. Furthermore, the functional relationships (the sensitivities, for instance) between the price of a financial instrument and the risk factor S are usually given as functions of S and not in terms of $\ln S$. Therefore, there are many reasons for transforming the volatility of logarithmic changes into one of absolute changes $\delta S(t)$. In linear approximation, this is quite simply accomplished through[8]

$$\delta \ln S(t) \equiv \ln S(t + \delta t) - \ln S(t) = \ln \left(\frac{S(t + \delta t)}{S(t)} \right)$$

$$= \ln \left(\frac{S(t) + \overbrace{[S(t + \delta t) - S(t)]}^{\delta S(t)}}{S(t)} \right)$$

$$= \ln \left(1 + \frac{\delta S(t)}{S(t)} \right) = \frac{\delta S(t)}{S(t)} - \frac{1}{2} \left(\frac{\delta S(t)}{S(t)} \right)^2 \pm \cdots$$

$$\approx \frac{\delta S(t)}{S(t)}$$

---

[8]The following expansion of the logarithm function is used: $\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \pm \cdots$

and hence,

$$\delta S(t) \approx S(t)\, \delta \ln S(t) \; . \tag{30.8}$$

Thus, in linear approximation, *relative* (i.e., percentage) price changes are equal to *logarithmic* price changes

$$\frac{S(t+\delta t) - S(t)}{S(t)} \equiv \frac{\delta S(t)}{S(t)} \approx \delta \ln S(t) \tag{30.9}$$

where the approximation becomes exact for $\delta t \to 0$. This, incidentally, is exactly the point which provides an intuitive picture of the logarithm: The logarithmic change over a time span is nothing other than the corresponding relative change if this time span is infinitesimally short (see also Sect. 2.3.2).

According to Eq. 30.8, the absolute price change follows from the logarithmic change multiplied by the current market price $S(t)$, called the *current level*. This is obviously consistent with the change in a risk factor over a finite time interval as given by Ito's lemma (see Eqs. 2.28 and 2.29) as can be seen by Taylor-expanding the exponential function appearing there for small δt. We have seen this approximation several times already in this book, see for example Eq. 21.17.

Note that $S(t)$ is known at time t and as such is not a random variable. On the right-hand side of Eq. 30.8, only $\delta \ln(S(t))$ is a random variable. Therefore, the variance of the absolute price change is simply

$$\text{var}\,[\delta S(t)] \approx \text{var}\,\Big[\underbrace{S(t)}_{\text{known}}\ \underbrace{\delta \ln S(t)}_{\text{random variable}} \Big] \tag{30.10}$$

$$= S(t)^2\, \text{var}\,[\delta \ln S(t)]$$

$$= S(t)^2 \sigma_i^2(t)\delta t \; ,$$

where in the last step the variance of the logarithmic price changes over a time span δt in accordance with the definition in Eq. 2.15 was expressed in terms of the volatility σ (in %) over this time span.

30.3.2   Summation of Volatilities

A sum (a portfolio) of *multiple* risk factors S_i, containing N_i of each of the risk factors S_i has a value given by

$$V(t) \equiv \sum_{i=1} N_i S_i(t)$$

The volatility σ_V of the value V of this portfolio is defined exactly as was done for individual risk factors as the standard deviation[9] of the *logarithmic* value changes of the portfolio. Equation 30.8 can be applied to approximate the logarithmic changes of V as well as of the individual risk factors, yielding

$$V(t)\, \delta \ln(V(t)) \approx \delta V(t) = \delta \left(\sum_{i=1}^n N_i S_i(t) \right) = \sum_{i=1}^n N_i \delta S_i(t)$$

$$\approx \sum_{i=1}^n N_i S_i(t)\, \delta \ln(S_i(t)) \,,$$

and hence

$$\delta \ln(V(t)) \approx \sum_{i=1}^n \underbrace{\left(\frac{N_i S_i(t)}{V(t)} \right)}_{\text{weight}} \underbrace{\delta \ln(S_i(t))}_{\text{random variable}} . \qquad (30.11)$$

The logarithmic change of the sum of risk factors is thus the weighted sum over the logarithmic changes of each of the individual risk factors. The weight of each individual risk factor is equal to the contribution (in %) of the current value of the pertinent risk factor to the total current value of the portfolio. It is thus a *present value weighting*, similar to that appearing in the *duration* in Eq. 5.8.

Each of the current prices $S_i(t)$ are known at time t and as such are not random variables. The logarithmic changes of the individual risk factors, on the other hand, are random variables. From Eq. A.17, we know that the

---

[9]More precisely: the standard deviation per square root of the time over which the change occurs.

variance of a sum of random variables is equal to the sum of the *covariances* of these random variables:

$$\text{var}\left[\delta \ln (V)\right] \approx \frac{1}{V^2} \text{var} \left[\sum_{i=1}^{n} N_i \, S_i \, \delta \ln (S_i) \right] \tag{30.12}$$

$$= \frac{1}{V^2} \sum_{i,j=1}^{n} N_i \, N_j \, S_i \, S_j \, \text{cov}\left[\delta \ln (S_i) , \delta \ln \left(S_j\right)\right] \ .$$

We now express the variances of the logarithmic price changes over a time span of length δt again as in Eq. 2.15 in terms of the corresponding volatilities and use Eq. A.14 for the covariance:

$$\text{var}\left[\delta \ln (V)\right] = \sigma_V^2 \delta t, \ \ \text{var}\left[\delta \ln (S_i)\right] = \sigma_i^2 \delta t$$

$$\text{cov}\left[\delta \ln (S_i) , \delta \ln \left(S_j\right)\right] = \rho_{i,j} \, \sigma_i \sqrt{\delta t} \, \sigma_j \sqrt{\delta t} \ .$$

Thus, we finally obtain the desired relation between the volatility σ_V of the sum of risk factors and the volatilities σ_i of the individual risk factors:

$$\sigma_V^2 \approx \sum_{i,j=1}^{n} \underbrace{\left(\frac{N_i \, S_i(t)}{V(t)} \right)}_{\text{weight } i} \sigma_i \, \rho_{i,j} \, \sigma_j \underbrace{\left(\frac{N_j \, S_j(t)}{V(t)} \right)}_{\text{weight } j} . \tag{30.13}$$

30.3.3 Transformation Between Yield and Price Volatility

As was discussed in detail in Sect. 22.2, the same holds (in linear approximation) for a portfolio with sensitivities $\Delta_i = \partial V / \partial S_i$ with respect to the risk factor S_i with the obvious association $\Delta_i = N_i$. This substitution leads from Eq. 30.13 immediately to Eq. 22.14.

Specifically, a change in the price V of a financial instrument with a sensitivity Δ_i with respect to a *single* risk factor S_i is in linear approximation given by

$$\delta V(t) \approx \Delta_i(t) \, \delta S_i(t) \approx \Delta_i(t) \, S_i(t) \, \delta \ln (S_i(t)) \ . \tag{30.14}$$

The first approximation is the delta-normal approximation for the price of the financial instrument as a function of the risk factor. The second approximation

is a result of the linear approximation 30.8 for the logarithm of the risk factor change. The linear approximation of the logarithmic change in V is likewise equal to the relative change $\delta V / V$. Dividing Eq. 30.14 by V yields the relationship between the *logarithmic* change of V and that of S:

$$\delta \ln (V(t)) \approx \frac{\delta V(t)}{V(t)} \approx \frac{\Delta_i(t)}{V(t)} S_i(t) \, \delta \ln (S_i(t)) \ . \tag{30.15}$$

At time t, all of the variables appearing in the equation are known except for the changes $\delta \ln(V)$ and $\delta \ln(S)$. The variance of this equation is thus

$$\text{var} [\delta \ln (V(t))] \approx \frac{\Delta(t)^2}{V(t)^2} S_i(t)^2 \text{var} [\delta \ln (S_i(t))] \ .$$

Expressing the variances through the volatilities as in Eq. 2.15, we obtain the relationship between the volatility σ_V of the price of a financial instrument induced by the volatility σ_i of a risk factor and the volatility of the risk factor itself:

$$\sigma_V(t) = \left| \frac{\Delta_i(t)}{V(t)} S_i(t) \right| \sigma_i(t) \ . \tag{30.16}$$

The absolute value sign indicates explicitly that the volatility, being the square root of the variance, is always positive.

The correlation of *another* risk factor S_j with this financial instrument (also dependent on S_i) is the same as its correlation with the risk factor S_i itself since

$$\sigma_V \, \sigma_j \, \rho_{V\,j} \, \delta t$$

$$= \text{cov} \left[\delta \ln (V), \ \delta \ln (S_j) \right] = \text{cov} \left[\frac{\Delta_i}{V} S_i \, \delta \ln (S_i), \ \delta \ln (S_j) \right]$$

$$= \frac{\Delta_i}{V} S_i \, \text{cov} \left[\delta \ln (S_i), \ \delta \ln (S_j) \right] = \underbrace{\frac{\Delta_i}{V} S_i \, \sigma_i}_{\sigma_V} \, \sigma_j \, \rho_{ij} \, \delta t$$

$$= \sigma_V \, \sigma_j \, \rho_{ij} \, \delta t \ ,$$

and hence

$$\rho_{Vj} = \rho_{ij} \ . \tag{30.17}$$

The transformation between the yield and price volatility is a special case of Eq. 30.16. Both the volatilities σ_R for interest rates $R(t, T)$, called *yield volatilities*, as well as volatilities σ_B for the prices (*price volatilities*) of the corresponding zero bonds $B_R(t, T)$ for the various terms $T - t$ are available in the market. The transformation between these two volatilities is accomplished with Eq. 30.16, where the relative sensitivity Δ / V for bonds is, as is well known, the (negative) *modified duration*, Eq. 5.10. The transformation for zero bonds specifically is thus

$$\sigma_B(t, T) = D_{\mathrm{mod}} R(t, T)\, \sigma(t, T) \qquad (30.18)$$

$$= \begin{cases} \dfrac{R(t, T)(T - t)}{1 + R(t, T)} \sigma_R(t, T) & \text{discrete compounding} \\[2ex] \dfrac{R(t, T)(T - t)}{1 + R(t, T)(T - t)} \sigma_R(t, T) & \text{simple compounding} , \end{cases}$$

where Table 15.1 in Sect. 15.1 provides the explicit form for the two most frequently used compounding methods.

30.3.4 Currency Transformation of Volatilities and Correlations

The General Transformations for All Risk Factors

Transformation of the Volatility Many risk factors are prices referring to a currency (and just one), even if quoted in %. E.g., a German government bond quoted at, say, 98% has a value of 98% of its face value in EUR. Interest rates refer to interest payments in a specific currency, too. The volatility of a quoted price does not have a currency itself, but still depends heavily on the currency in which the price or risk factor is quoted. The same holds for correlations, with the difference that a correlation refers to two risk factors and therefore two currencies may be involved. In the following, we use the notion *reference currency* to denote the currency used in which the present values, prices and risk factors shall be calculated. If the reference currency differs from the currency of the risk factor (*foreign currency*), then the risk factor, expressed in terms of the reference currency, is also exposed to variations of the FX rate between the reference and foreign currency. This impacts volatilities and correlations related to this risk factor as well. The risk factor's volatility differs if the risk factor is quoted in foreign currency rather than in in reference currency. The

reference currency could be, e.g., equal to the *accounting currency* (sometimes also denoted as *base currency* or *domestic currency*) or to the currency of the cash flows of a specific derivative, even if the derivative's underlying is quoted in another currency. Please note, that for quotation of FX rates a different notation is used, see Sect. 30.3.4 for details.

It will be shown that the volatility transformation to another currency is similar (but not identical!) to the estimation of the volatility of portfolio with two risk factors. Here, the first risk factor is the price in foreign currency and the second risk factor is the exchange rate (*FX rate*) of the foreign currency in terms of the reference currency. First, we analyse the volatility transformation for lognormally distributed risk factors. Afterwards, we consider the transformation of correlations.

Let S be the price of a risk factor in foreign currency other than the reference currency and D the exchange rate with respect to the reference currency (i.e., D units of the reference currency are exchanged for one unit of the foreign currency). The value of the risk factor expressed in terms of the reference currency is simply the product DS of the exchange rate and S. We now calculate the variance of the logarithmic change of this value in the reference currency. We make use of the logarithm's property to turn products into sums. Then, the variance of the resulting sum of random variables can be expressed as the sum of the covariances as in Eq. A.17:

$$\text{Var}\left[\delta \ln\left(D(t)S(t)\right)\right]$$

$$= \text{Var}\left[\underbrace{\delta \ln(D(t))}_{\text{random variable } X_1} + \underbrace{\delta \ln\left(S(t)\right)}_{\text{random variable } X_2}\right] = \sum_{a,b=1}^{2} \text{cov}\left[X_a, X_b\right]$$

$$= \text{Var}\left[\delta \ln\left(D(t)\right)\right] + \text{var}\left[\delta \ln\left(S(t)\right)\right] + 2\text{cov}\left[\delta \ln\left(D(t)\right), \delta \ln\left(S(t)\right)\right] \ .$$

With these variances, it follows immediately from Eq. 2.15 that the volatility of the product DS is given by

$$\sigma_{DS} = \sqrt{\sigma_D^2 + \sigma_S^2 + 2\rho_{S,D}\,\sigma_D\,\sigma_S} \ . \tag{30.19}$$

This is the volatility transformation: on the left-hand side, we have the desired volatility in the reference currency, on the right, the volatility with respect to the foreign currency, the volatility of the exchange rate and the correlation between them.

While Eq. 30.13 gives the volatility of a *sum* of risk factors, Eq. 30.19 is the volatility of the *product* of risk factors where the exchange rate D and the price S each represent a risk factor.

Although the results are quite similar, there are two fundamental differences:

- In contrast to Eq. 30.13, no approximation of the logarithm was made in the derivation of Eq. 30.19. It is therefore an exact equality.
- In Eq. 30.19, only the volatilities and correlations appear and not the current levels of the risk factors D and S.

Let us consider the following example: a portfolio composed of a British bond with a face value of N =GBP 100, 000 whose price and volatility is quoted in British Pounds (GBP). The risk of the portfolio *expressed in GBP* is simply the product of the given daily volatility σ = 0.026% and the current level 99.704% of the face value N (multiplied by the square root of the liquidation period and the percentile for the confidence level c):

$$\begin{aligned}
\text{VaR}(c, t, T) &= Q_{1-c}\sqrt{T - t}\ S\,N\,\sigma_S \\
&= Q_{1-c}\sqrt{T - t} \times 97,704\% \times N \times 0,00026 \\
&= Q_{1-c}\sqrt{T - t} \times 25,38\ \text{GBP} .
\end{aligned}$$

The risk is now to be converted into USD. The transformation in Eq. 30.19 intuitively means that a further risk factor must be taken into consideration, the USD/GBP exchange rate D. This value and its daily volatility are: $D =$ 1.469 USD/GBP and $\sigma_D =$ 0.404%. The correlation between the two risk factors is $\rho_{S,D} = -0.11574$. The value at risk in USD (with respect to a confidence level of c and a liquidation period $T - t$) is essentially the product of the current portfolio value in USD and the portfolio volatility (with respect to USD). The portfolio's value is just the product (and not the sum; herein lies the reason for the difference between this and Eq. 30.13) of the two risk factors S and D. The portfolio volatility with respect to USD is given by the transformation in Eq. 30.19. Thus, the value at risk in USD becomes

$$\begin{aligned}
\text{VaR}(c, t, T) &= Q_{1-c}\sqrt{T - t}D(t)S(t)N\sigma_{SD} \\
&= Q_{1-c}\sqrt{T - t}D(t)S(t)N\sqrt{\sigma_D^2 + \sigma_S^2 + 2\rho_{S,D}\,\sigma_D\,\sigma_S}
\end{aligned}$$

$$= Q_{1-c}\sqrt{T-t} \times 97,704, \text{GBP} \times 1.469\frac{\text{USD}}{\text{GBP}}$$

$$\times \sqrt{0,00404^2 + 0,00026^2 - 2 \times 0,11574 \times 0,00404 \times 0,00026}$$

$$= Q_{1-c}\sqrt{T-t} \times 577,39\,\text{USD}\,.$$

This is very different from the (wrong) result obtained had we (incorrectly) converted the GBP-based value at risk into the value at risk in USD by simply multiplying it by the current GBP/USD exchange rate. This would have resulted in a completely incorrect VaR of "25.38 GBP ×1.469 USD/GBP = 37.28 USD", which differs significantly from the correct result. Thus, the risk of the position mainly stems from the fact that it is a position in a foreign currency (as seen from USD). The specific price risk of the bond with respect to its home currency only plays a minor role.

Transformation of the Correlation Having presented the transformation of the volatility in detail, we now consider the correlations. The transformation for the correlations are somewhat more complex since two risk factors are involved, these being the two factors to which the correlation refers. We must differentiate between two cases: a risk factor is given in the reference currency and the other in the foreign currency, or both risk factors are quoted in (different) foreign currencies.

Case 1 The risk factor with price S_1 is given in the foreign currency, the second risk factor S_2 in the reference currency. The foreign currency has the exchange rate D with respect to the reference currency (D units of the reference currency are exchanged for one unit in the foreign currency). The price of the first risk factor in the reference currency is therefore equal to DS_1. We now consider the covariance of the logarithmic value changes of these prices with those of the second risk factor. Exploiting the linearity of the covariance (Eq. A.13), it is easy to see that

$$\text{cov}\,[\delta \ln(D\,S_1),\; \delta \ln(S_2)]$$
$$= \text{cov}\,[\delta \ln(D) + \delta \ln(S_1),\; \delta \ln(S_2)]$$
$$= \text{cov}\,[\delta \ln(D),\; \delta \ln(S_2)] + \text{cov}\,[\delta \ln(S_1),\; \delta \ln(S_2)]\,.$$

With Eq. A.14, the covariances can be expressed in terms of the volatilities and the correlations:

$$\rho_{DS_1,\,S_2}\sigma_{DS_1}\,\sigma_{S_2} = \rho_{D,\,S_2}\,\sigma_D\,\sigma_{S_2} + \rho_{S_1,\,S_2}\,\sigma_{S_1}\,\sigma_{S_2}\,.$$

Applying Eq. 30.19 for the volatility of the product DS_1 finally yields the desired transformation

$$\rho_{DS_1, S_2} = \frac{\rho_{D, S_2} \sigma_D + \rho_{S_1, S_2} \sigma_{S_1}}{\sqrt{\sigma_D^2 + \sigma_{S_1}^2 + 2\rho_{D, S_1} \sigma_D \sigma_{S_1}}} . \tag{30.20}$$

The correlation transformation of a price S given in the foreign currency and the exchange rate D itself is a special case of Eq. 30.20 with $S_1 := S$ and S_2 set equal to D:

$$\rho_{DS, D} = \frac{\sigma_D + \rho_{D,S} \sigma_S}{\sqrt{\sigma_D^2 + \sigma_S^2 + 2\rho_{D, S} \sigma_D \sigma_S}} = \frac{\sigma_D + \rho_{D,S} \sigma_S}{\sigma_{DS}} , \tag{30.21}$$

where Eq. 30.19 was used in the last step.

Case 2 The risk factor with a price S_1 is quoted in a foreign currency, the second risk factor S_2 is likewise given in a foreign currency. These two foreign currencies may be different. The foreign currency associated with S_1 has exchange rate D_1 with respect to the reference currency, the second foreign currency has exchange rate D_2. Then the price of the first risk factor in the reference currency equals $D_1 S_1$. Correspondingly the price of the second risk factor is $D_2 S_2$. As before, we consider the covariance of the logarithmic value changes of the prices

$$\begin{aligned}
&\text{cov}\left[\delta \ln (D_1 S_1), \delta \ln (D_2 S_2)\right] \\
&= \text{cov}\left[\delta \ln(D_1) + \delta \ln(S_1), \delta \ln(D_2) + \delta \ln(S_2)\right] \\
&= \text{cov}\left[\delta \ln(D_1), \delta \ln(D_2)\right] + \text{cov}[\delta \ln(S_1), +\delta \ln(S_2)] \\
&\quad + \text{cov}\left[\delta \ln(D_1), \delta \ln(S_2)\right] + \text{cov}[\delta \ln(S_1), \delta \ln(D_2)] .
\end{aligned}$$

Applying Eq. A.14, the covariances again can be expressed in terms of the volatilities and the correlations:

$$\begin{aligned}
\rho_{D_1 S_1, D_2 S_2} \sigma_{D_1 S_1} \sigma_{D_2 S_2} &= \rho_{D_1, D_2} \sigma_{D_1} \sigma_{D_2} + \rho_{D_1, S_2} \sigma_{D_1} \sigma_{S_2} \\
&\quad + \rho_{D_2, S_1} \sigma_{D_2} \sigma_{S_1} + \rho_{S_1, S_2} \sigma_{S_1} \sigma_{S_2} .
\end{aligned}$$

Using Eq. 30.19 for the volatilities of the products $D_1 S_1$ and $D_2 S_2$ finally yields the desired transformation

$$\rho_{D_1 S_1, D_2 S_2} \tag{30.22}$$

$$= \frac{\rho_{D_1, D_2}\sigma_{D_1}\sigma_{D_2} + \rho_{D_1, S_2}\sigma_{D_1}\sigma_{S_2} + \rho_{D_2, S_1}\sigma_{D_2}\sigma_{S_1} + \rho_{S_1, S_2}\sigma_{S_1}\sigma_{S_2}}{\sqrt{\sigma_{D_1}^2 + \sigma_{S_1}^2 + 2\rho_{D_1, S_1}\sigma_{D_1}\sigma_{S_1}} \sqrt{\sigma_{D_2}^2 + \sigma_{S_2}^2 + 2\rho_{D_2, S_2}\sigma_{D_2}\sigma_{S_2}}} \, .$$

This is the correlation with respect to the reference currency between both of the risk factors S_1 and S_2 quoted in their respective (different) currencies.

We are frequently confronted with the special case of this equation in which the two foreign currencies are identical, implying $D_1 = D_2 = D$. In this case the transformation reduces to

$$\rho_{DS_1, DS_2} = \frac{\sigma_D^2 + \left(\rho_{D, S_2}\sigma_{S_2} + \rho_{D, S_1}\sigma_{S_1}\right)\sigma_D + \rho_{S_1, S_2}\sigma_{S_1}\sigma_{S_2}}{\sqrt{\sigma_D^2 + \sigma_{S_1}^2 + 2\rho_{D, S_1}\sigma_D\sigma_{S_1}} \sqrt{\sigma_D^2 + \sigma_{S_2}^2 + 2\rho_{D, S_2}\sigma_D\sigma_{S_2}}} \, . \tag{30.23}$$

Transformations for the Exchange Rates and Cross Rates

The transformations described above are frequently used in practice. Currencies are quoted most often against USD, though some are also quoted against EUR or other currencies. For a given pair of currencies, e.g. EUR/USD, the first currency (i.e. EUR) is called *base currency*, while the second currency is called *quote currency* (i.e. USD). The quoted price is given as the price of one unit of base currency in terms of the quote currency. If the quote currency differs from the reference currency, the FX quotes need to be converted to a convention where the base currency equals the foreign currency and the quote currency equals the reference currency, before they can be used for valuation or risk management. Here, *cross rates* come into play. For example, for an USD-based investor, the exchange rate from EUR into Yen (i.e. EUR/JPY), is a cross rate. Cross rates are obtained by dividing the desired rate (in this case USD/JPY) by the exchange rate of the new base currency (for example USD/EUR). To illustrate this concept further, consider the following examples of exchange rates with respect to the USD

$$D_i(t) = \frac{USD}{EUR}, \quad D_j(t) = \frac{USD}{GBP}$$

$$D_n(t) = \frac{USD}{JPY}, \quad D_m(t) = \frac{USD}{CHF}$$

From these, cross rates for arbitrary combinations can be established as follows:

$$D_{ji}(t) = \frac{D_j(t)}{D_i(t)} = \frac{EUR}{GBP}$$

$$D_{ni}(t) = \frac{D_n(t)}{D_i(t)} = \frac{EUR}{JPY}$$

$$D_{mn}(t) = \frac{D_j(t)}{D_n(t)} = \frac{JPY}{CHF} \ , \quad \text{etc.}$$

Thus, the subtlety in transforming volatilities and correlations of cross rates lies in the fact that, as opposed to the previous section, not the products but the *quotients* of risk factors (exchange rates) must be considered. The quotient of two values a and b can, of course, be seen as the product of a and $1/b$. The arguments of the previous section can then be applied. Consider the volatilities and correlations of an "inverse risk factor" (exchange rate) $1/D$:

$$\text{var}\left[\delta\ln\left(\frac{1}{D(t)}\right)\right] = \text{var}\left[-\delta\ln(D(t))\right] = (-1)^2\text{var}\left[\delta\ln(D(t))\right]$$

$$\text{cov}\left[\delta\ln\left(\frac{1}{D(t)}\right), \delta\ln(S(t))\right] = \text{cov}\left[-\delta\ln(D(t)), \delta\ln(S(t))\right]$$

$$= -\text{cov}\left[\delta\ln(D(t)), \delta\ln(S(t))\right] \ .$$

This implies that $1/D$ has the same volatility as D, and that the correlations of $1/D$ with an arbitrary risk factor S is just the negative correlation of D with S (the correlation between two inverse prices is then the same as the correlation between the prices themselves):

$$\sigma_{\frac{1}{D}} = \sigma_D \ , \quad \rho_{\frac{1}{D},S} = -\rho_{D,S} \ , \quad \rho_{\frac{1}{D},\frac{1}{S}} = \rho_{D,S} \ . \quad (30.24)$$

Now, the transformations of the previous section can directly be applied. Let D_b denote the quoted exchange rate of the investor's reference currency quoted in terms of some other quote currency. For example, if the data are quoted with respect to USD and the investor's reference currency is EUR then $D_b = \text{USD/EUR}$. Let D_j be the quoted exchange rate of another currency. The corresponding volatilities and correlation quoted by the data provider are

denoted by σ_b, σ_j and $\rho_{b,j}$. The volatility of D_j with respect to the investors reference currency is the volatility σ_{jb} of the cross rate

$$D_{jb} := D_j/D_b \ . \tag{30.25}$$

When applying the equations of the previous sections we generally have to associate D_j with a risk factor S quoted in the foreign currency and $1/D_b$ with the exchange rate D from that foreign currency to our domestic or home currency.

- From Eq. 30.19 (with the substitution $S = D_j$ and $D = 1/D_b$), the volatility of a cross rate is:

$$\sigma_{jb} = \sqrt{\sigma_j^2 + \sigma_b^2 - 2\rho_{j,b}\,\sigma_j\,\sigma_b} \tag{30.26}$$

- From Eq. 30.22 (with the substitution $S_1 = D_j$, $D_1 = 1/D_i$, $S_2 = D_n$ and $D_2 = 1/D_m$), the correlation between two cross rates $D_{ji} = D_j/D_i$ and $D_{nm} = D_n/D_m$ is:

$$\rho_{ji,\,nm} = \frac{\rho_{i,m}\sigma_i\,\sigma_m - \rho_{n,i}\sigma_n\,\sigma_i - \rho_{j,m}\sigma_j\,\sigma_m + \rho_{j,\,n}\sigma_j\,\sigma_n}{\sqrt{\sigma_j^2 + \sigma_i^2 - 2\,\rho_{j,i}\,\sigma_j\,\sigma_i}\,\sqrt{\sigma_n^2 + \sigma_m^2 - 2\,\rho_{n,\,m}\,\sigma_n\,\sigma_m}} \tag{30.27}$$

$$= \frac{\rho_{i,m}\sigma_i\,\sigma_m - \rho_{n,i}\sigma_n\,\sigma_i - \rho_{j,m}\sigma_j\,\sigma_m + \rho_{j,\,n}\sigma_i\,\sigma_m}{\sigma_{ji}\,\sigma_{nm}} \ ,$$

where in the last step Eq. 30.26 was used. The correlation of two cross rates with respect to the same quote currency can now be obtained by setting $m = i = b$:

$$\rho_{jb,\,nb} = \frac{\sigma_b^2 - (\rho_{n,b}\sigma_n + \rho_{j,b}\sigma_j)\,\sigma_b + \rho_{j,\,n}\sigma_j\,\sigma_n}{\sqrt{\sigma_j^2 + \sigma_b^2 - 2\,\rho_{j,b}\,\sigma_j\,\sigma_b}\,\sqrt{\sigma_n^2 + \sigma_b^2 - 2\,\rho_{n,\,b}\,\sigma_n\,\sigma_b}} \tag{30.28}$$

$$= \frac{\sigma_b^2 - \left(\rho_{n,b}\sigma_n + \rho_{j,b}\sigma_j\right)\,\sigma_b + \rho_{j,\,n}\sigma_j\,\sigma_n}{\sigma_{jb}\,\sigma_{nb}} \ .$$

- From Eq. 30.20 (with the substitution $D = 1/D_b$, $S_1 = D_j$ and $S_2 = S$), the correlation of a cross rate D_{jb} with a risk factor S still quoted in its original foreign currency is:

$$\rho_{jb,S} = \frac{\rho_{j,S}\,\sigma_j - \rho_{b,S}\,\sigma_b}{\sqrt{\sigma_b^2 + \sigma_j^2 - 2\,\rho_{b,S}\,\sigma_b\,\sigma_S}} = \frac{\rho_{j,S}\,\sigma_j - \rho_{b,S}\,\sigma_b}{\sigma_{jb}} \,. \qquad (30.29)$$

We emphasize that in this equation the risk factor S is still quoted in its original foreign currency.

- From Eq. 30.19 (with the substitution $D = D_{jb}$), the volatility of a *transformed* risk factor S (i.e., of a risk factor quoted in the investor's home currency by applying the corresponding (cross) exchange rate D_{jb}) is

$$\sigma_{D_{jb}S} = \sqrt{\sigma_{jb}^2 + \sigma_S^2 + 2\rho_{jb,S}\,\sigma_{jb}\,\sigma_S} \qquad (30.30)$$

$$= \sqrt{\sigma_{jb}^2 + \sigma_S^2 + 2\left(\sigma_j\rho_{j,S}\,\sigma_S - \sigma_b\rho_{b,S}\sigma_S\right)}$$

with σ_{jb} from Eq. 30.26 and $\rho_{jb,S}$ from Eq. 30.29.

- The correlation of a risk factor S quoted in the investor's reference currency (by applying the corresponding (cross) exchange rate D_{jb}) with a cross rate $D_{nb} = D_n/D_b$ results from Eq. 30.20 with the substitutions $D = D_{jb}$, $S_1 = S$ and $S_2 = D_{nb}$:

$$\rho_{D_{jb}S,\,D_{nb}} = \frac{\rho_{jb,nb}\,\sigma_{jb} + \rho_{nb,S}\,\sigma_S}{\sqrt{\sigma_{jb}^2 + \sigma_S^2 + 2\,\rho_{jb,S}\,\sigma_{jb}\,\sigma_S}} \qquad (30.31)$$

with $\rho_{jb,nb}$ as in Eq. 30.28, $\rho_{jb,S}$ as in Eq. 30.29 and σ_{jb} as in Eq. 30.26. This can of course also be derived from first principles:

$$\text{cov}\left[\delta\ln\left(D_{jb}S\right),\,\delta\ln(D_{nb})\right] = \text{cov}\left[\delta\ln\left(D_{jb}\right) + \delta\ln(S),\,delta\ln(D_{nb})\right]$$

$$= \text{cov}\left[\delta\ln\left(D_{jb}\right),\,\delta\ln(D_{nb})\right]$$

$$+ \text{cov}\left[\delta\ln(S),\,\delta\ln(D_{nb})\right]$$

$$\rho_{D_{jb}S,\,D_{nb}}\sigma_{D_{jb}S}\,\sigma_{nb} = \rho_{jb,nb}\sigma_{jb}\sigma_{nb} + \rho_{nb,S}\,\sigma_S\sigma_{nb} \,.$$

A factor σ_{nb} cancels in the last equation and for $\sigma_{D_{jb}S}$ we use Eq. 30.30 to arrive at Eq. 30.31. In the special case that both cross rates are the same, $D_{nb} = D_{jb}$, we have:

$$\rho_{D_{jb}S, D_{jb}} = \frac{\sigma_{jb} + \rho_{jb, S}\, \sigma_S}{\sqrt{\sigma_{jb}^2 + \sigma_S^2 + 2\, \rho_{jb, S}\, \sigma_{jb}\, \sigma_S}} \; . \tag{30.32}$$

- The correlation of a risk factor S_1 (quoted in the investor's home currency by applying the corresponding (cross) exchange rate D_{jb}) with another risk factor S_2 (quoted in the investor's home currency by applying the corresponding (cross) exchange rate D_{ib}) results from Eq. 30.22 with the substitutions $D_1 = D_{jb}$ and $D_2 = D_{ib}$:

$$\rho_{D_{jb}S_1, D_{ib}S_2} = \frac{\rho_{jb, ib}\sigma_{jb}\, \sigma_{ib} + \rho_{jb, S_2}\sigma_{jb}\, \sigma_{S_2} + \rho_{ib, S_1}\sigma_{ib}\, \sigma_{S_1} + \rho_{S_1, S_2}\sigma_{S_1}\, \sigma_{S_2}}{\sqrt{\sigma_{jb}^2 + \sigma_{S_1}^2 + 2\, \rho_{jb, S_1}\, \sigma_{jb}\, \sigma_{S_1}}\, \sqrt{\sigma_{ib}^2 + \sigma_{S_2}^2 + 2\, \rho_{ib, S_2}\, \sigma_{ib}\, \sigma_{S_2}}} \tag{30.33}$$

with $\rho_{jb, ib}$ as in Eq. 30.28, the correlations between cross rates and risk factors (like for instance ρ_{jb, S_2}, ρ_{ib, S_1}, etc.) as in Eq. 30.29 and the cross rate volatilities σ_{jb} and σ_{ib} as in Eq. 30.26. The correlation ρ_{S_1, S_2} and volatilities σ_{S_1} and σ_{S_2} are the risk factor data quoted in their respective *original* currencies, as quoted externally. This can of course also be derived from first principles:

$$\begin{aligned}
\text{cov}\left[\delta \ln\left(D_{jb}S_1\right), \; \delta \ln\left(D_{ib}S_2\right)\right] &= \text{cov}\left[\delta \ln\left(D_{jb}\right) + \delta \ln\left(S_1\right), \right.\\
&\qquad \left. \delta \ln\left(D_{ib}\right) + \delta \ln\left(S_2\right)\right] \\
&= \text{cov}\left[\delta \ln\left(D_{jb}\right), \delta \ln\left(D_{ib}\right)\right] \\
&\quad + \text{cov}\left[\delta \ln\left(D_{ib}\right), \delta \ln\left(S_1\right)\right] \\
&\quad + \text{cov}\left[\delta \ln\left(D_{jb}\right), \delta \ln\left(S_2\right)\right] \\
&\quad + \text{cov}\left[\delta \ln\left(S_1\right), \delta \ln\left(S_2\right)\right] \\
\rho_{D_{jb}S_1, D_{ib}S_2}\sigma_{D_{jb}S_1}\, \sigma_{D_{ib}S_2} &= \rho_{jb, ib}\sigma_{jb}\sigma_{ib} + \rho_{ib, S_1}\sigma_{ib}\sigma_{S_1} \\
&\quad + \rho_{jb, S_2}\sigma_{jb}\sigma_{S_2} \; .
\end{aligned}$$

Dividing by $\sigma_{D_{jb}S_1}$ and $\sigma_{D_{ib}S_2}$ and using Eq. 30.30 for those two volatilities we arrive at Eq. 30.33. In the special case that both cross rates are the same, $D_{ib} = D_{jb}$, we have:

$$\rho_{D_{jb}S_1, D_{jb}S_2} = \frac{\sigma_{jb}^2 + \left(\rho_{jb,\,S_2}\sigma_{S_2} + \rho_{jb,\,S_1}\sigma_{S_1}\right)\sigma_{jb} + \rho_{S_1,\,S_2}\sigma_{S_1}\,\sigma_{S_2}}{\sqrt{\sigma_{jb}^2 + \sigma_{S_1}^2 + 2\,\rho_{jb,\,S_1}\,\sigma_{jb}\,\sigma_{S_1}}\,\sqrt{\sigma_{jb}^2 + \sigma_{S_2}^2 + 2\,\rho_{jb,\,S_2}\,\sigma_{ib}\,\sigma_{S_2}}} \,.$$

$$(30.34)$$

The Excel workbook VALUEATRISK.XLSX from the download section [50], demonstrates the application of the transformations introduced above explicitly. Table 30.1 shows the original quotes. The transformations of these data into EUR-based data proceeds in four steps:

1. The volatilities and correlations of the cross rates with each other with respect to EUR are determined in accordance with Eqs. 30.26 and 30.28.
2. Now the correlations $\rho_{jb,\,S}$ between these cross rates and the risk factors are determined in an intermediate step using Eq. 30.29. Observe that here the risk factors are not yet converted by any exchange rate but are still quoted in there original currencies, i.e., the Japanese 5-year zero bond is quoted in JPY, the Japanese 7-year zero bond is quoted in JPY and the British 6-month zero bond is quoted in GBP.
3. Only after these two steps can we apply Eqs. 30.31 and 30.32 to determine the correlations between the risk factors (now quoted in the reference currency) and the cross rates (also quoted in the reference currency).
4. Finally the EUR-based volatilities of the *transformed* risk factors (i.e., of the risk factors quoted in the investor's home currency) are calculated via Eq. 30.30 and the correlations of the transformed risk factors with one another are determined via Eq. 30.33 (or Eq. 30.34 for the correlation between the two Japanese bonds).

Table 30.1 The original price volatilities and correlations. The data are for relative daily changes

	USD/EUR	USD/JPY	USD/GBP	JPY.Z05	JPY.Z07	GBP.R180
USD/EUR	**0.680%**	0.49105	0.76692	0.13130	0.03701	−0.01863
USD/JPY	0.49105	**0.731%**	0.51365	−0.02040	−0.02092	−0.07130
USD/GBP	0.76692	0.51365	**0.404%**	0.12458	0.10580	−0.11574
JPY.Z05	0.13130	−0.02040	0.12458	**0.093%**	0.89410	0.17171
JPY.Z07	0.03701	−0.02092	0.10580	0.89410	**0.177%**	0.00987
GBP.R180	−0.01863	−0.07130	−0.11574	0.17171	0.00987	**0.026%**

The bold values are in the main-diagonal of the matrix and are thus volatilities (quoted as percentages) while off-diagonal elements are correlations (quoted as numbers)

Table 30.2 The price volatilities and correlations with respect to EUR

	EUR/JPY	EUR/GBP	JPY.Z05	JPY.Z07	GBP.R180
EUR/JPY	**0.713%**	0.49406	0.99141	0.96976	0.49221
EUR/GBP	0.49406	**0.452%**	0.48798	0.49552	0.99834
JPY.Z05	0.99141	0.48798	**0.705%**	0.99006	0.48739
JPY.Z07	0.96976	0.49552	0.99006	**0.725%**	0.49386
GBP.R180	0.49221	0.99834	0.48739	0.49386	**0.450%**

The bold values are in the main-diagonal of the matrix and are thus volatilities (quoted as percentages) while off-diagonal elements are correlations (quoted as numbers)

Table 30.3 The price volatilities and correlations with respect to USD

	USD/EUR	USD/JPY	USD/GBP	JPY.Z05	JPY.Z07	GBP.R180
USD/EUR	**0.680%**	0.49105	0.76692	0.50501	0.48829	0.76987
USD/JPY	0.49105	**0.731%**	0.51365	0.99192	0.97158	0.51183
USD/GBP	0.76692	0.51365	**0.404%**	0.52664	0.52665	0.99794
JPY.Z05	0.50501	0.99192	0.52664	**0.735%**	0.99058	0.52632
JPY.Z07	0.48829	0.97158	0.52665	0.99058	**0.748%**	0.52515
GBP.R180	0.76987	0.51183	0.99794	0.52632	0.52515	**0.402%**

The bold values are in the main-diagonal of the matrix and are thus volatilities (quoted as percentages) while off-diagonal elements are correlations (quoted as numbers)

Observe that for Steps 3 and 4 we need the EUR-based correlations and volatilities of the cross rates from Step 1 and the correlations between the risk factors (still in their original currencies) and the cross rates obtained in Step 2. The results of these transformations are presented in Table 30.2 and were already used in Table 24.3 of Chap. 24 for calculating the value at risk.

As expected, from the perspective of an EURO-based investor there is a very high correlation between the Japanese bonds and the EUR/JPY exchange rate. In fact, the price movements (in EURO) of these bonds are almost exclusively determined by the exchange rate. This is in stark contrast to the *original* correlations of those bonds with the exchange rates in Table 30.1 which are very low. But the currency transformations correctly capture the fact, that those bond prices, if seen from outside Japan, are very much dependent on the JPY exchange rate.

The original data refer to *different* currencies: the exchange rates all refer to the USD, the Japanese bonds to JPY and the British bond to GBP. Thus, even if the reference currency was USD, the volatilities and correlations of the Japanese and British bonds must be converted. This is accomplished using Eq. 30.19 for the volatilities, Eqs. 30.20 (with S_2 being an exchange rate) and 30.21 for the correlations between the bonds and the exchange rates and Eqs. 30.22 and Eq. 30.23 for correlations of the bonds with one another. The corresponding results are presented in Table 30.3.

31

Check for
updates

Market Parameter from Historical Time Series

Having shown in the previous sections how statistical parameters such as the volatility can be obtained *implicitly* from the prices of derivatives traded in the market (if they are not quoted directly anyway), we now proceed with presenting in the following section, how such statistical figures, one of which is the volatility, could be determined by analyzing the historical time series. These volatilities are called *real world* or *historic* volatilities in contrast to *implicit* or *risk-neutral* volatilities.

Depending on the usage, either risk-neutral or real-world volatilities are applied. The risk-neutral or arbitrage-free valuation of derivatives requires risk-neutral volatilities (as one might have guessed), since otherwise it would be impossible to match the price quotes of (simple) options, which are used for dynamic hedging that is essential for the risk-neutral valuation approach. On the other hand, the purpose of risk management is to calculate the risk of a potential real-world loss especially of those risks, which are not explicitly (statically) hedged. Here, real-world volatilities are needed.

Time series analysis is a broad topic in the field of *statistics* whose application here will be limited to those areas which serve the purposes of this book. A much more general and wide-reaching presentation can be found in [86], for example.

31.1 Historical Averages as Estimates for Expected Values

Based on a *time series* of some risk factor (e.g. a stock price), we start with calculating for every price change the logarithm of the ratio of new and old price, which equals in linear approximation the *relative* price change, see Eq. 30.9. The statistical measures we estimate are then related to the logarithm of risk factor, e.g. the stock price. This is always handy, if we consider lognormally distributed risk factors. For stock prices, this is the case, at least in first order approximation. Interest rates used to be considered as lognormally distributed, too, as long as the interest rates does not become to low or even negative. For normally distributed risk factors (e.g. interest rates in domains with low or negative interest rates), the *absolute* difference of price for risk factor changes would be used. The statistical methods described below work in both cases, and for a historical as well as for a simulated time series.

From the expectation and the variance of this variable the *historical volatility* σ and the *historical mean return* μ can be determined by recording the relative changes over a period of length δt along a historical (or simulated) path:

$$\mu = \frac{1}{\delta t} \, \mathrm{E}\,[X] \;, \quad \sigma^2 = \frac{1}{\delta t} \, \mathrm{Var}\,[X]$$

$$\text{with} \;\; X = \ln\left(\frac{S(t+\delta t)}{S(t)}\right) \approx \frac{S(t+\delta t) - S(t)}{S(t)} \;. \tag{31.1}$$

The *historical correlation* between two price processes is likewise determined from the relative changes. With Y denoting the change of the second price (analogous to X as defined above), the correlation, as defined in Eq. A.14, is

$$\rho = \frac{\mathrm{cov}\,[XY]}{\sqrt{\mathrm{Var}\,[X]\,\mathrm{Var}\,[Y]}} \;.$$

These equations make it apparent that a procedure is required enabling the determination of the expectation and variance from time series data. It is well known from statistics that from n measurements X_i, which are realizations of a random variable X, the *mean* $\langle X \rangle$ (and more generally the mean $\langle f(X) \rangle$ of a function of X) can be computed in the following way

$$\langle X \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i \;, \quad \langle f(X) \rangle = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \;.$$

The *law of large numbers* states that these means approximate the desired *expectations* and variances as follows[1]:

$$\langle f(X) \rangle \xrightarrow{n \to \infty} E[f(X)]$$

$$\langle f(X)^2 \rangle - \langle f(X) \rangle^2 \xrightarrow{n \to \infty} \frac{n-1}{n} \operatorname{var}[f(X)] . \tag{31.2}$$

This result may seem trivial or the difference between the actual parameters (the right-hand side) and the measured approximation (left-hand side) may be unclear to the reader. This difference, however, is fundamental: the theoretical value (the right-hand side) can never be precisely known; it can only be *estimated* more or less exactly through the computation of means of measured data. Such "means of measured data" are called *estimators* in statistics to distinguish them from the true values. Examples of estimators are the expectations on the left-hand sides in Eq. 31.2. More estimators will be introduces in the following sections.

The expectation and variance of X are needed for the description of the risk factor X. For this, the means of the realizations of the random variables X and X^2 are needed. The determination of the *error* made in making these estimates requires X^4 as well (see Sect. 31.2). Therefore, for any time series analysis the computation of the following measures are especially helpful:

$$\langle X \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i , \quad \langle X^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i^2 , \quad \langle X^4 \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i^4 . \tag{31.3}$$

These measures are called the 1., 2. and 4. *moment* of the distribution. In general, the n-th moment m_k of a distribution is defined as

$$m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k .$$

If multiple correlated prices are involved, the following means for each pair of prices are also necessary:

$$\langle XY \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i , \quad \langle (XY)^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i^2 Y_i^2 . \tag{31.4}$$

---

[1]The factor $(n-1)/n$ of the variance is necessary if the estimator for the variance is to be *unbiased*. See any introductory statistics textbook for more on this subject.

These are the *composite moments* $m_{1,1}$ and $m_{2,2}$ of the time series X and Y. From these values, historical estimates for the mean return, volatility and correlation can be obtained:

$$\mu = \frac{1}{\delta t}\, \mathrm{E}\,[X] \approx \frac{1}{\delta t}\,\langle X \rangle \tag{31.5}$$

$$\sigma = \frac{1}{\sqrt{\delta t}}\sqrt{\mathrm{var}\,[X]} \approx \frac{1}{\sqrt{\delta t}}\sqrt{\frac{n}{n-1}}\sqrt{\langle X^2 \rangle - \langle X \rangle^2}$$

$$\rho = \frac{\mathrm{cov}\,[XY]}{\sqrt{\mathrm{var}\,[X]}\,\sqrt{\mathrm{var}\,[Y]}} \approx \frac{\langle XY \rangle - \langle X \rangle\,\langle Y \rangle}{\sqrt{\langle X^2 \rangle - \langle X \rangle^2}\,\sqrt{\langle Y^2 \rangle - \langle Y \rangle^2}}\,.$$

31.2 Error Estimates

Estimating the error in the measured values is essential for evaluating their meaningfulness. It is therefore insufficient to provide a value as the result of making observations since the actual theoretical value is not obtained (and will never be obtained) from measured data. It is more appropriate to find an interval on the basis of observed data within which the theoretical value lies.

A claim such as "the historical price volatility is 20% per year" is not very meaningful if nothing is said about the error associated with such a claim, for example 0.5% or 50%. A valid statement on the other hand might be "the historical price volatility is 20% ± 3% per year". This means that the actual volatility lies with high probability somewhere between 17% and 23% per year. The safer you need to be, i.e. the higher the probability should be that the value lies indeed in this interval, the greater the interval will be, if the range of parameter values is not restricted for other reasons (e.g., in case of a dice the result of a valid throw is with 100% probability between 1 and 6).

In this section, several simple methods for determining the statistical error are introduced and its calculation will be demonstrated explicitly for the volatility and correlation. We begin by assuming that data in a time series (referred to as *observed values* or *measurements*) are pair wise independent and in consequence uncorrelated. Finally, we will briefly show how to test whether measurements are independent and how to proceed in the case that they are not, i.e., how to account for autocorrelations. Naturally, the subject is quite technical. Error estimation is theoretically quite simple but lengthy. Nonetheless, anyone who wishes to conduct a serious analysis of historical or simulated data should understand and apply this material.

There are two different types of error. The first is the *statistical error*, which is a consequence of the fact that only a finite number of measurements are taken. The second is the *systematic error*. These are errors arising from a fundamental mistake (for example, a programming error in a Monte Carlo simulation). The failure to decrease with an increasing number of measurements is characteristic of a systematic error (as opposed to the statistical error). Only the statistical error will be dealt with in this section.

31.2.1 Uncorrelated Measurements

The determination of expectations is accomplished through calculating the mean of the observed values as in Eq. 31.5. For a sufficiently large number n of measurements

$$E[X] \approx \langle X \rangle , \quad \text{var}[X] \approx \frac{n}{n-1} \left(\langle X^2 \rangle - \langle X \rangle^2 \right) \tag{31.6}$$

holds. The central question is: what is the (statistical) error involved in estimating these parameters as above? The n observations are realizations of the random variables X_i, $i = 1, \ldots n$. The mean is the weighted sum over these random variables X_i and as such is again a random variable. The statistical error will be defined as the *standard deviation* of this new random variable, or equivalently, the error is the square root of the variance of the mean. A fundamental result from the field of statistics is

$$\langle X \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \Rightarrow \tag{31.7}$$

$$\text{var}[\langle X \rangle] = \frac{1}{n} \text{var}[X] \quad \text{if } E[X_i X_j] = 0 \quad \text{for } i \neq j .$$

The variance of the *mean* of uncorrelated, identically distributed random variables is equal to the variance of the random variable itself divided by the number of observations. This result combined with the approximation in Eq. 31.6 for the variance yields the statistical error (denoted below by the symbol δ) defined as the standard deviation of the mean

$$\delta \langle X \rangle \equiv \sqrt{\text{var}[\langle X \rangle]} = \sqrt{\frac{1}{n} \text{var}[X]} \approx \sqrt{\frac{\langle X^2 \rangle - \langle X \rangle^2}{n-1}} . \tag{31.8}$$

This holds in general for the mean of a function of the random variable X:

$$\delta \langle f(X) \rangle = \sqrt{\frac{1}{n} \operatorname{var}[f(X)]} \approx \sqrt{\frac{\langle f(X)^2 \rangle - \langle f(X) \rangle^2}{n-1}} . \tag{31.9}$$

The statistical error in the estimation of the mean return as defined in Eq. 31.5 is thus[2]

$$\delta \mu \approx \frac{1}{\delta t \sqrt{n-1}} \sqrt{\langle X^2 \rangle - \langle X \rangle^2} . \tag{31.10}$$

The mean of a function is to be distinguished from the function of the mean if the function is not linear: $\langle f(X) \rangle \neq f(\langle X \rangle)$. Thus, the *error of a function* of an uncertain value z (in the case under discussion, $z = \langle X \rangle$) cannot be computed directly in general. A Taylor series expansion of the function can provide assistance in such cases. The *propagation of error* can be obtained from this Taylor series:

$$f = f(z) \text{ mit } z = z \pm \delta z \implies \delta f = \left| \delta z \frac{\partial f}{\partial z} \right| + \frac{1}{2} \left| (\delta z)^2 \frac{\partial^2 f}{\partial z^2} \right| + \cdots$$

The vertical lines in the above equation indicate the absolute value. This can be generalized for functions of multiple variables with associated errors. Such expressions quickly become very lengthy. Restricting our consideration to the first (linear) terms, we obtain what is known as the *quadratic error propagation.*

$$f = f(z_1, z_2, \ldots, z_k) \text{ with } z_i = z_i \pm \delta z_i$$

$$\implies \quad \delta f \approx \left| \sum_{i=1}^{k} \delta z_i \frac{\partial f}{\partial z_i} \right| = \sqrt{\sum_{i=1}^{k} \left(\delta z_i \frac{\partial f}{\partial z_i} \right)^2} . \tag{31.11}$$

---

[2]Here, the δ in δt denotes the length of a time interval between two data points in the time series and *not* the "error in t".

This law will be required in order to determine the error involved in measuring the variance since

$$\text{var}[X] \approx \frac{n}{n-1} \left(\langle X^2 \rangle - \langle X \rangle^2 \right) = f(z_1, z_2) \quad \text{with}$$

$$z_1 = \langle X \rangle, \quad z_2 = \langle X^2 \rangle, \quad f(z_1, z_2) = \frac{n}{n-1} \left(z_2 - z_1^2 \right)$$

$$\Rightarrow \quad \frac{\partial f}{\partial z_1} = -\frac{2n}{n-1} z_1, \quad \frac{\partial f}{\partial z_2} = \frac{n}{n-1}.$$

The error of f, calculated with quadratic error propagation, is then

$$\delta f \approx \sqrt{ \left(\delta z_1 \frac{\partial f}{\partial z_1} \right)^2 + \left(\delta z_2 \frac{\partial f}{\partial z_2} \right)^2 }$$

$$= \sqrt{ \left(\delta z_1 \frac{2n}{n-1} z_1 \right)^2 + \left(\delta z_2 \frac{n}{n-1} \right)^2 }$$

$$= \frac{n}{n-1} \sqrt{ 4 z_1^2 (\delta z_1)^2 + (\delta z_2)^2 }.$$

Since z_1 and z_2 are means of X and $f(X) = X^2$ respectively, their errors are respectively,

$$z_1 = \langle X \rangle \quad \Rightarrow \quad \delta z_1 = \delta \langle X \rangle \approx \frac{1}{\sqrt{n-1}} \sqrt{ \langle X^2 \rangle - \langle X \rangle^2 }$$

$$z_2 = \langle X^2 \rangle \quad \Rightarrow \quad \delta z_2 = \delta \langle X^2 \rangle \approx \frac{1}{\sqrt{n-1}} \sqrt{ \langle X^4 \rangle - \langle X^2 \rangle^2 }.$$

Substituting these into the expression for δf yields

$$\delta f \approx \frac{1}{\sqrt{n-1}} \frac{n}{n-1} \sqrt{ 4 \langle X \rangle^2 \left(\langle X^2 \rangle - \langle X \rangle^2 \right) + \left(\langle X^4 \rangle - \langle X^2 \rangle^2 \right) }$$

$$= \frac{1}{\sqrt{n-1}} \frac{n}{n-1} \sqrt{ \langle X^4 \rangle - \langle X^2 \rangle^2 + 4 \langle X^2 \rangle \langle X \rangle^2 - 4 \langle X \rangle^4 }.$$

This is the statistical error made in measuring the variance var[X]. In order to determine this error when analyzing historical time series, the means of X, X^2 and X^4 must be measured.

Likewise, the error of the volatility σ can be determined through the following deliberations

$$\sigma\,[X] \equiv \frac{1}{\sqrt{\delta t}}\sqrt{\mathrm{var}\,[X]} \approx \frac{1}{\sqrt{\delta t}}\sqrt{\frac{n}{n-1}}\sqrt{\langle X^2\rangle - \langle X\rangle^2} = \frac{1}{\sqrt{\delta t}}g(z_1, z_2)$$

with $z_1 = \langle X\rangle$, $\quad z_2 = \langle X^2\rangle$, $\quad g(z_1, z_2) = \sqrt{\frac{n}{n-1}}\sqrt{z_2 - z_1^2}$ $\Rightarrow$

$$\frac{\partial g}{\partial z_1} = -\sqrt{\frac{n}{n-1}}\frac{z_1}{\sqrt{z_2 - z_1^2}}, \qquad \frac{\partial g}{\partial z_2} = \frac{1}{2}\sqrt{\frac{n}{n-1}}\frac{1}{\sqrt{z_2 - z_1^2}}.$$

An analogous calculation as above yields

$$\delta\sigma \approx \frac{1}{\sqrt{\delta t}}\frac{1}{\sqrt{n-1}}\sqrt{\frac{n}{n-1}}\frac{\sqrt{\langle X^4\rangle - \langle X^2\rangle^2 + 4\langle X^2\rangle\langle X\rangle^2 - 4\langle X\rangle^4}}{2\sqrt{\langle X^2\rangle - \langle X\rangle^2}}$$

$$(31.12)$$

for the error in the measured volatility. The expression for the error of the correlation between two prices X and Y is even longer as it is represented by a function of *five* means:

$$\rho[X, Y] \equiv \frac{\mathrm{cov}[XY]}{\sqrt{\mathrm{var}[X]\mathrm{var}[Y]}} \approx \frac{\langle XY\rangle - \langle x\rangle\langle Y\rangle}{\sqrt{\langle X^2\rangle - \langle X\rangle^2}\sqrt{\langle Y^2\rangle - \langle Y\rangle^2}} = \rho(z_1, z_2, z_3, z_4, z_5)$$

where $z_1 = \langle X\rangle$, $\quad z_2 = \langle Y\rangle$, $\quad z_3 = \langle X^2\rangle$, $\qquad z_4 = \langle Y^2\rangle$, $\quad z_5 = \langle XY\rangle$

$$\rho = \frac{z_5 - z_1 z_2}{\sqrt{z_3 - z_1^2}\sqrt{z_4 - z_2^2}}$$

The derivatives of the correlation with respect to the z_i are

$$\frac{\partial\rho}{\partial z_1} = -\rho\left(z_1 + \frac{z_2}{z_5 - z_1 z_2}\right)$$

$$\frac{\partial\rho}{\partial z_2} = -\rho\left(z_2 + \frac{z_1}{z_5 - z_1 z_2}\right)$$

$$\frac{\partial \rho}{\partial z_3} = \frac{\partial \rho}{\partial z_4} = \frac{\rho}{2}$$

$$\frac{\partial \rho}{\partial z_5} = \frac{\rho}{z_5 - z_1 z_2}$$

The errors of the z_i are as in Eq. 31.9

$$\delta z_1^2 = \frac{\langle X^2 \rangle - \langle X \rangle^2}{n-1}, \quad \delta z_2^2 = \frac{\langle Y^2 \rangle - \langle Y \rangle^2}{n-1}$$

$$\delta z_3^2 = \frac{\langle X^4 \rangle - \langle X^2 \rangle^2}{n-1}, \quad \delta z_4^2 = \frac{\langle Y^4 \rangle - \langle Y^2 \rangle^2}{n-1}$$

$$\delta z_5^2 = \frac{\langle X^2 Y^2 \rangle - \langle XY \rangle^2}{n-1} .$$

All these results inserted into Eq. 31.11 yields the statistical error of the correlation

$$\partial \rho \approx \sqrt{\sum_{i=1}^{5} \left(\delta z_i \frac{\partial \rho}{\partial z_i} \right)^2} .$$

Table 31.1 illustrates the application of Eq. 31.5 for the measurement of the mean return and the volatility from a data set with $n = 250$ observations and the estimation of their statistical errors in accordance with Eqs. 31.10 and 31.12. In addition to the relative price changes X, the second and fourth powers are measured. Using these means, the above equations are used to compute the mean return and the volatility as well as the errors involved in estimating these two parameters.

The data set was generated by a simulated random walk with a yield of 6.00% and a volatility of 20.00%. The measured values could thus be compared with the "true values" (a luxury naturally not at our disposal when using historical data). The true values lie within the error of the measurement.

The error is naturally large since the number of measurements taken is so small. As the number of measurements gets bigger, we see that the error decreases as the inverse of square root of the number of measurements; a tenfold decrease in the statistical error can thus only be accomplished if a sample size 100 times as large is placed at our disposal.

Table 31.1 Measuring the yield, the volatility and their errors from a (simulated) data series of 250 "measurements". The parameters used to simulate the data set are shown for comparison

	Yield per δt		Vol per δt	
Simulated	6.00%		20.00%	
Measured	5.99%		19.70%	
Error	1.25%		1.00%	

x	$x2$		$x4$		
		Averages			
0.05986719	0.04223756		0.00502627		
		Data			n
−0.070,8944,77	0.005,026,027		2.526,09	$\times 10^{-05}$	1
−0.014,768,650	0.000,218,113		4.757,33	$\times 10^{-08}$	2
0.011,417,976	0.000,130,37		1.699,64	$\times 10^{-08}$	3
−0.066,321,993	0.004,398,607		1.934,77	$\times 10^{-05}$	4
−0.113,237,822	0.012,822,804		0.000,164,424		5
0.089,718,194	0.008,049,354		6.479,21	$\times 10^{-05}$	6
0.200,262,728	0.040,105,16		0.001,608,424		7
0.329,164,502	0.108,349,270		0.011,739,564		8
−0.003,557,309	1.265,44	$\times 10^{-05}$	1.601,35	$\times 10^{-10}$	9
−0.066,032,319	0.004,360,267		1.901,19	$\times 10^{-05}$	10
0.266,072,855	0.070,794,764		0.005,011,899		11
−0.173,980,700	0.030,269,284		0.000,916,23		12
0.193,003,141	3.725,02	$\times 10^{-02}$	0.001387578		13
−0.044,716,037	0.001,999,524		3.998,1	$\times 10^{-06}$	14
−0.038,558,758	0.001,486,778		2.2105,1	$\times 10^{-06}$	15
−0.033,664,767	0.001,133,317		1.284,41	$\times 10^{-06}$	16
…	…		…		…

It is essential to be aware that errors are also only statistical quantities. Therefore we can *not* be *sure* that the interval defined by the error actually contains the true value of the estimated parameter; there is only a certain *probability* that this is the case. If we have reason to believe that the measured estimator, such as the mean return, is normally distributed, then there is approximately 68% probability that the true parameter will lie within the error interval, since the error is defined as *one* standard deviation. Thus, the probability for the true value to lie outside the range obtained from the statistical error is approximately 32% in this case. If this uncertainty is too large, we could of course define the statistical error to correspond to two or three standard deviations or more. We only need to multiply the error in the derivations above by the desired multiplicative factor. For example, if we define the error as two standard deviations and the measured estimator is normally distributed, then the probability that the true value will lie within the new error interval is 95.4%.

As is clear from the above, we need the probability distribution of the measured estimator if we want to assign confidence levels to error intervals. It can by no means be taken for granted that the measured estimator has the same distribution as the random variables in the time series. If, for instance, the random variables X in the time series are *uniformly* distributed on a finite interval $[a, b]$ then the estimator for μ as defined in Eq. 31.5 is (for large n) approximately *normally* distributed (because of the *central limit theorem*, see Sect. A.4.3). If, however, the random variables X in the time series are *normally* distributed, then the estimator for μ is also *normally* distributed. However, the estimator for the *variance* is in this case a χ^2-distributed variable. This can be seen as follows: according to Eqs. 31.6 and 31.3, the estimator for the variance is

$$\mathrm{var}\,[X] \approx \frac{n}{n-1}\left(\left\langle X^2\right\rangle - \langle X\rangle^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \langle X\rangle^2\right)\ .$$

The X_i are all normally distributed and the mean $\langle X\rangle$ is also normally distributed in this case. Therefore the estimator for the variance is a sum of *squared* normally distributed random variables and as such χ^2-distributed (see Sect. A.4.6).

31.2.2   Error of Autocorrelated Measurements

The methods described above for the determination of the statistical error hold only for *uncorrelated* measurements. That means, it has been tacitly assumed that the n measurements in a time series come from *independent* observations. However, independence is an assumption which often cannot be made, in particular in time series analysis (for example, in the case of moving average methods, see Sect. 33.4). Daily changes in a moving 30-day price average, for example (the mean of prices observed over the previous 30 days is computed) will remain small from one day to the next since in each daily adjustment, only the oldest price is replaced by the most recently observed value, the other 29 prices in the average remain the same. The measurement of such a variable is highly correlated with the measurement made on the previous day. In such a situation, error considerations must be modified significantly.

Autocorrelation and Autocovariance

A correlation of one and the same variable *with itself* is called *autocorrelation*.[3] Just as *correlation* measures the (linear) dependency between two *different* random processes, *auto*correlation measures the (linear) dependency between a process has *on itself*. The autocorrelation is defined by

$$\varrho(t, h) = \frac{\mathrm{cov}(X_{t+h}, X_t)}{\mathrm{cov}(X_t, X_t)} = \frac{\mathrm{cov}(X_{t+h}, X_t)}{\mathrm{var}(X_t)} \ , \tag{31.13}$$

where the definition of the *autocovariance* is completely analogous to the definition of the covariance between two different random variables (see Eq. A.10)

$$\mathrm{cov}(X_{t+h}, X_t) = \mathrm{E}\left[(X_{t+h} - \mathrm{E}\left[X_{t+h}\right])(X_t - \mathrm{E}\left[X_t\right])\right] \tag{31.14}$$

$$= \mathrm{E}\left[X_{t+h} X_t\right] - \mathrm{E}\left[X_{t+h}\right] E\left[X_t\right] \ .$$

We arrive at the final equality by merely taking the product in the previous expression and using the linearity of the expectation. The resulting equation corresponds to A.11. Definition 31.14 immediately establishes the relation between the autocovariance and the variance

$$\mathrm{cov}(X_t, X_t) = \mathrm{E}[(X_t - \mathrm{E}\left[X_t\right])(X_t - \mathrm{E}\left[X_t\right])] = \mathrm{var}(X_t) \ .$$

which has already been used in establishing the last equality in Eq. 31.13.

The autocorrelation in Eq. 31.13 also corresponds to the ordinary correlation known from statistics. If the time series is stationary[4] (in particular, if the variance of the process remains constant), $\mathrm{var}(X_t) = \mathrm{var}(X_{t+h})$ holds and ρ can be written in a form corresponding to Eq. A.14:

$$\varrho(h) = \frac{\mathrm{cov}(X_{t+h}, X_t)}{\sqrt{\mathrm{var}(X_{t+h})}\sqrt{\mathrm{var}(X_t)}} \ .$$

---

[3]Autocorrelations do not appear merely in certain measurement methods but are in general inherent to non-Markov processes, i.e. for processes whose current value is influenced by past values. See Sect. 32.1 for more on this topic. It should be noted that correlation measures only linear dependencies, though.

[4]This means intuitively that the parameters describing the time series are time independent, see Chap. 32.

This is just the definition of the correlation of the random variable X_{t+h} with the random variable X_t, irrespective of whether these random variables belong to the same time series.

Autocorrelation Time and Error Estimates

With the autocorrelations above we can estimate an autocorrelation *time* which specifies the number of time steps needed between two measurements in order to guarantee that the two measurements are (at least approximately) uncorrelated. The *autocorrelation time τ* is defined through the autocorrelation Eq. 31.13 in the following manner

$$\tau(t) \equiv \frac{1}{2} \sum_{h=-\infty}^{\infty} \varrho(t, h) = \frac{1}{2} \sum_{h=-\infty}^{\infty} \frac{\text{cov}(X_{t+h}, X_t)}{\text{cov}(X_t, X_t)} \; . \tag{31.15}$$

Usually, only *stationary* time series are investigated (see Chap. 32). Then ϱ is only dependent on the time difference h between the observations and not on the time point t. The autocorrelation time is then likewise independent of t, i.e., a constant. For *un*correlated observations we have $\varrho(t, h) = \delta_{h,0}$ (where $\delta_{h,0}$ again denotes the Kronecker delta) and therefore simply $\tau = 1/2$.

If the number n of observations is much larger than the autocorrelation time τ it can be shown that

$$\delta \langle X \rangle \approx \sqrt{\frac{2\,\tau}{n} \, \text{var}\,[X]} \approx \sqrt{\frac{2\,\tau}{n-1} \left[\langle X^2 \rangle - \langle X \rangle^2 \right]} \qquad \text{for} \quad n >> \tau \tag{31.16}$$

holds for the mean error of X. This reduces to Eq. 31.8 if the measurements are uncorrelated since then we have $\tau = 1/2$.

The autocorrelation time (and thus the autocorrelations) must be measured if the error in Eq. 31.16 is to be determined. An estimator for the *autocovariance* is given by

$$\text{cov}(X_{t+h}, X_t) \approx \frac{1}{n - |h|} \sum_{\substack{i,j \\ i-j=|h|}} X_i X_j - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2 \; . \tag{31.17}$$

For $h = 0$ this estimator is similar to Eq. 31.6 but the factor $n/(n - 1)$ is not reproduced (for large n however, this factor is very close to 1):

$$\text{var}\,[X_t] = \text{cov}(X_t, X_t) \approx \frac{1}{n} \sum_{\substack{i,j \\ i=j}} X_i X_j - \left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)^2 = \left\langle X_t^2 \right\rangle - \langle X_t \rangle^2 \ .$$

As can be seen from Eq. 31.15, the autocorrelations between $h = -\infty$ and $h = -\infty$ must in theory be measured for the calculation of the autocorrelation *time*. In practice, a suitable *cutoff* $(n - \tilde{n})$ can be chosen to limit the sum in Eq. 31.15 to a finite one, neglecting the autocorrelations that are so small that they contribute almost nothing to the sum. Substituting the estimator for the autocovariance, Eq. 31.17, then yields the estimator for the autocorrelation time.

$$\tau \approx \frac{1}{2} \sum_{h=-(n-\tilde{n})}^{(n-\tilde{n})} \frac{\text{cov}(X_{t+h}, X_t)}{\text{cov}(X_t, X_t)} \quad \text{with} \ \ \tau \ << \tilde{n} << n$$

$$\approx \frac{1}{2} \frac{1}{\left\langle X^2 \right\rangle - \langle X \rangle^2} \sum_{h=\tilde{n}-n}^{n-\tilde{n}} \left[\frac{1}{n - |h|} \sum_{\substack{i,j \\ i-j=|h|}} X_i X_j - \langle X \rangle^2 \right] \ .$$

All this substituted into Eq. 31.16 finally provides a possibility of estimating the error of a mean taking autocorrelations into account:

$$\delta\,\langle X \rangle \approx \sqrt{\frac{1}{n - 1} \sum_{h=\tilde{n}-n}^{n-\tilde{n}} \left[\frac{1}{n - |h|} \sum_{\substack{i,j \\ i-j=|h|}} X_i X_j - \langle X \rangle^2 \right]} \ .$$

There are thus two possibilities of taking the autocorrelations into consideration:

- We wait for at least as long as the autocorrelation time before taking a second measurement; the measurements would then be uncorrelated.
- We do *not* wait for the autocorrelation time to pass (e.g., because it is to short, for instance) and use the above expression for the error of the correlated observations.

31.3 Return and Covariance Estimates

We will now present some commonly used estimators for returns and covariances. The importance of return estimates is unquestionable for any investment decision. Just as important are covariance estimates for quantifying the risk, as shown for instance in Sect. 21.5. In addition, estimates for other quantities like volatilities, correlations and Betas can be derived from the covariance estimates. For all estimates we will use historical risk factor market prices S_i, $i = 0, 1, \ldots K$ at times

$$t_n = t_0 + n \, dt \ \ \text{with } n = 0, \ldots, N \ .$$

The estimates will be done at time

$$t = t_N > t_0$$

and the *window* used for the estimates ranges from t_0 until t_N. Regular sizes of such time windows range from ca. 30 day to ca. 2 years. We will denote by δt the time span over which the estimations will be made. This time span is also called *investment horizon*, *holding period* or *liquidation period*. The time span between to adjacent data in the time series will be denoted by dt. We will present a situation often occurring in practice, namely that the holding period does *not* have the same length as dt. To still have a clear presentation of the issues we will, however, assume that the holding period is a multiple of dt:

$$\delta t = m \, dt \ .$$

Using time series of *daily* settlement prices, for instance, one can calculate estimators for holding periods of m days (e.g., weekly or monthly estimators).

31.3.1 Return Estimates

The risk factor *returns* over the holding period δt, i.e., the logarithmic price changes will be denoted by r_i.

$$r_i(t) = \frac{1}{\delta t} \ln \left[\frac{S_i(t + \delta t)}{S_i(t)} \right] \quad \text{for } i = 0, \ldots, K \ . \tag{31.18}$$

The historical prices $S_i(t_n)$ in a time series with $N + 1$ prices at times t_n with

$$t_n - t_{n-1} = dt \text{ for all } n = 1, \ldots, N$$

can be translated into *historical returns* for all past holding periods (all with length δt):

$$r_i(t_{n-m}) = \frac{1}{\delta t} \ln \left[\frac{S_i(t_n)}{S_i(t_{n-m})} \right] \text{ for } i = 0, \ldots, K \text{ and } n = m, \ldots, N \, .$$

$$(31.19)$$

The return at the *current* time $t = t_N$ over the *next* holding period δt (which still lies in the future) is only known for the risk free investment (this is the risk free rate). For all risky investments, however, the future prices $S_i(t_N + \delta t)$ are not yet known. Thus, the returns cannot be calculated but can only be estimated. We will use $\mu_i(t_N)$ to denote the estimator for $r_i(t_N)$.

The Moving Average (MA) Estimator

The common *moving average* estimator is simply the (equally weighted) mean of all historical returns over past time spans of length δt within the time window used for the estimation.[5]

$$\mu_i(t_N) = \frac{1}{N - m + 1} \sum_{n=m}^{N} r_i(t_{n-m})$$

$$= \frac{1}{(N - m + 1)\,\delta t} \sum_{n=m}^{N} \ln \left[\frac{S_i(t_n)}{S_i(t_{n-m})} \right] \, .$$

The logarithm appearing above can be written as a sum over historical dt-returns since everything except the first and the last term cancels in the

---

[5]Of course, this estimator is strongly autocorrelated since one time step later $N - m$ out of the $N - m + 1$ values in the sum are still the same.

following sum:

$$\sum_{k=1}^{m} \ln \left[\frac{S_i(t_{n-k+1})}{S_i(t_{n-k})} \right] = \ln[S_i(t_n)] - \ln[S_i(t_{n-1})] +$$

$$\ln[S_i(t_{n-1})] - \ln[S_i(t_{n-2})] +$$

$$\dots +$$

$$\ln[S_i(t_{n-m+1})] - \ln[S_i(t_{n-m})]$$

$$= \ln[S_i(t_n)] - \ln[S_i(t_{n-m})] = \ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})} \right].$$

Inserting this into the expression for $\mu_i(t_N)$ and making the index substitution $x = n - k$ yields:

$$\mu_i(t_N) = \frac{1}{(N-m+1)\,\delta t} \sum_{n=m}^{N} \sum_{k=1}^{m} \ln \left[\frac{S_i(t_{n-k+1})}{S_i(t_{n-k})} \right]$$

$$= \frac{1}{(N-m+1)\,\delta t} \sum_{k=1}^{m} \sum_{x=m-k}^{N-k} \ln \left[\frac{S_i(t_{x+1})}{S_i(t_x)} \right]. \qquad (31.20)$$

Now we have expressed $\mu_i(t_N)$ as a sum over dt-returns (in contrast to δt-returns). Again, almost everything cancels in the sum over x

$$\sum_{x=m-k}^{N-k} \ln \left[\frac{S_i(t_{x+1})}{S_i(t_x)} \right] = \ln \left[\frac{S_i(t_{N-k+1})}{S_i(t_{m-k})} \right],$$

leaving us with

$$\mu_i(t_N) = \frac{1}{(N-m+1)\,\delta t} \sum_{k=1}^{m} \ln \left[\frac{S_i(t_{N-k+1})}{S_i(t_{m-k})} \right]. \qquad (31.21)$$

The remaining sum contains only m terms now, instead of N. This clearly shows the fundamental problem with return estimates: only historical prices in the earliest and latest time period δt contribute. All prices in between are simply not used! This becomes especially severe for $dt = \delta t$, i.e., for $m = 1$. In this case only the very first and the very last price of the time window

contribute:

$$\mu_i \left(t_N\right) = \frac{1}{N \, \delta t} \ln \left[\frac{S_i(t_N)}{S_i(t_0)} \right] .$$

Two Alternatives for the Moving Average

One way to circumvent this problem is to use historical returns over time spans with different lengths $(N - n) \, dt$, all ending today. Based on Definition 31.18, we can construct an estimator using these historical returns in the following way:

$$\mu_i \left(t_N\right) = \frac{1}{N} \sum_{n=0}^{N-1} \underbrace{\frac{1}{(N - n) \, dt} \ln \left[\frac{S_i(t_N)}{S_i(t_n)} \right]}_{\text{return from } t_n \text{ to } t_N}$$

$$= \frac{1}{N \, dt} \sum_{n=0}^{N-1} \frac{1}{N - n} \ln \left[\frac{S_i(t_N)}{S_i(t_n)} \right] . \tag{31.22}$$

Observe that in this estimator we have dt (and not δt) appearing in the denominator. All historical prices (even the ones way back in the past) enter with their influence still relevant today, namely with the return over the corresponding time span ending today. The historical returns over long time periods enter with the same weight as returns of shorter time periods. Thus, old prices are effectively under-weighted since the corresponding returns, although belonging to long time spans, have no more influence then returns over shorter time spans (resulting from more recent prices).

If older prices are to be as important as more recent ones, we can give the historical returns weights proportional to the length of the time spans they belong to: the longer the time span, the more weight the corresponding returns gets. An estimator with this feature is

$$\mu_i \left(t_N\right) = \underbrace{\frac{1}{\sum_{k=0}^{N-1} (N - k)}}_{\text{numeraire}} \sum_{n=0}^{N-1} \underbrace{(N - n)}_{\text{time weight}} \cdot \underbrace{\frac{\ln \left[S_i(t_N)/S_i(t_n) \right]}{(N - n) \, dt}}_{\text{yield from } t_n \text{ to } t_N}$$

$$= \frac{2}{N(N + 1) dt} \sum_{n=0}^{N-1} \ln \left[\frac{S_i(t_N)}{S_i(t_n)} \right] , \tag{31.23}$$

where we have used $\sum_{k=0}^{N-1} (N-k) = N(N+1)/2$, which can be shown easily.[6]

The Exponentially Weighted Moving Average (EWMA)

A very well known method is the *exponentially weighted moving average* estimator, or *EWMA* for short. In this estimator, the historical data are weighted less and less the further in the past they lie. This is accomplished by a damping factor λ in the following way:

$$\mu_i\,(t_N) = \frac{1}{M} \sum_{n=m}^{N} \lambda^{N-n} r_i\,(t_{n-m}) \text{ with } 0 < \lambda \leq 1 \text{ and } M := \sum_{k=m}^{N} \lambda^{N-k} \ .$$

(31.24)

With Eq. 31.19 for the historical returns this can be written as

$$\mu_i\,(t_N) = \frac{1}{M\,\delta t} \sum_{n=m}^{N} \lambda^{N-n} \ln\left[\frac{S_i\,(t_n)}{S_i\,(t_{n-m})}\right]$$

(31.25)

In contrast to the simple moving average, Eq. 31.21, the price logarithms of adjacent time periods dt do not cancels each other since they are differently weighted. Thus all prices influence the estimator.

By adjusting the parameter λ, the EWMA estimator can be made similar to the estimator in Eq. 31.22 as well as to the estimator in Eq. 31.23. For $\lambda = 1$ the EWMA estimator equals the simple moving average in Eq. 31.21.

---

[6]On one hand we get with the index transformation $i := N - k$

$$\sum_{k=0}^{N-1} (N-k) = \sum_{i=N}^{i=1} i = \sum_{i=1}^{N} i; \ .$$

On the other hand we have

$$\sum_{k=0}^{N-1} (N-k) = N^2 - \sum_{k=0}^{N-1} k = N^2 - \sum_{i=1}^{N}(i-1) = N^2 + N - \sum_{i=1}^{N} i \ .$$

Equating both results yields

$$\sum_{i=1}^{N} i = N^2 + N - \sum_{i=1}^{N} i \implies \sum_{i=1}^{N} i = N(N+1)/2 \ .$$

All the above return estimates are demonstrated in the Excel-Workbook RETURNESTIMATES.XLS. Although neither the root mean square errors nor the Correlations with the ex post realized returns point out any clear favorite, one can see from the graphical presentation of the return time series that Eq. 31.23 looks like the best compromise between quite stable historical evolvement and still fast reaction to market movements; even when large time windows are used for the estimation.

31.3.2 Covariance Estimates

The entries $\delta \Sigma_{ij}$ in the covariance matrix, Eq. 21.22, i.e., the *risk factor* covariances can be determined via Eq. 21.28 using the *return* covariances

$$\delta \Sigma_{ij} \equiv \mathrm{cov}\left[\delta \ln S_i, \delta \ln S_j\right] = \delta t^2 \, \mathrm{cov}\left[r_i, r_j\right]$$

Similarly to the moving average estimators of the mean returns, the moving average estimators for the covariances of the returns over the holding period with length $\delta t = m dt$ at time t_N are

$$\mathrm{cov}\left[r_i, r_j\right](t_N) = \frac{1}{N-m} \sum_{n=m}^{n=N} \left[r_i(t_{n-m}) - \mu_i(t_N)\right]\left[r_j(t_{n-m}) - \mu_j(t_N)\right] ,$$

with r as in Eq. 31.19 and μ as in Eq. 31.21. Explicitly:

$$\mathrm{cov}\left[r_i, r_j\right](t_N) \tag{31.26}$$

$$= \frac{1}{N-m} \sum_{n=m}^{n=N} \left(\frac{1}{\delta t} \ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] - \mu_i(t_N)\right)\left(\frac{1}{\delta t} \ln\left[\frac{S_j(t_n)}{S_j(t_{n-m})}\right] - \mu_j(t_N)\right) .$$

The EWMA estimator is analogously

$$\mathrm{cov}\left[r_i, r_j\right](t_N) \tag{31.27}$$

$$= \frac{1}{M-1} \sum_{n=m}^{n=N} \lambda^{N-n} \left[\frac{1}{\delta t} \ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] - \mu_i(t_N)\right]\left[\frac{1}{\delta t} \ln\left[\frac{S_j(t_n)}{S_j(t_{n-m})}\right] - \mu_j(t_N)\right]$$

with μ from Eq. 31.25 and M from Eq. 31.24.

From these covariance estimates the *volatility* can be determined according to Eq. 21.30

$$\sigma_i \left(t_N \right) = \sqrt{\delta t \, \text{cov} \left[r_i, r_i \right] \left(t_N \right)} \qquad \text{for} \quad i = 0, \ldots, K \; . \qquad (31.28)$$

With the moving average estimate, Eq. 31.26, this yields the simply moving average volatility estimate. With Eq. 31.27, this yields the EWMA volatility estimate. The difference between these estimates and the ones presented in Sect. 33.4 is that in Sect. 33.4 we assume $\mu_i \equiv 0$.

The *correlations* resulting from the above covariances are according to their definition (see for example Eq. 31.5)

$$\rho_{i,j} \left(t_N \right) = \frac{\text{cov} \left[r_i, r_j \right] \left(t_N \right)}{\sqrt{\text{cov} \left[r_i, r_i \right] \left(t_N \right)} \sqrt{\text{cov} \left[r_j, r_j \right] \left(t_N \right)}} \quad \text{for} \; i, j = 0, \ldots, K$$

$$(31.29)$$

Again, this yields the moving average or the EWMA estimate depending on which estimate is used for the covariance.

In the *Capital Asset Pricing Model* (*CAPM*) there is a ratio *Beta* , which relates the evolvement of risk factors $S_i(t)$ to the evolvement of a benchmark $S_0(t)$. As a rule, this benchmark represents a whole market and is usually an index. Beta is calculated from a regression of the risk factors time series with the benchmark time series, see Sect. 28.1. Estimates for Beta directly follow from the covariance estimates: the Beta of the i-th risk factor at time t_N is

$$\beta_i \left(t_N \right) = \frac{\text{cov} \left[r_i, r_0 \right] \left(t_N \right)}{\text{cov} \left[r_0, r_0 \right] \left(t_N \right)} = \rho_{i,0} \left(t_N \right) \frac{\sigma_i \left(t_N \right)}{\sigma_0 \left(t_N \right)} \quad \text{for} \; i = 1, \ldots, K \; .$$

$$(31.30)$$

Again, this yields the moving average or the EWMA estimate for β, depending on which estimate is used for the covariance.

32

Check for updates

Time Series Modeling

Time series analysis aims to develop a *model*, which describes the time series in all its measurable features. This goes far beyond than merely determining statistical parameters from observed time series data (such as the variance, correlation, etc.) as described in Chap. 31. Estimators such as those appearing in Eq. 31.5 are examples of how *parameters* can be estimated which are subsequently used to *model* the stochastic process governing the time series (for example, a random walk with drift μ and volatility σ). To develop a model that is capable of simulation a time series with similar features is the principle goal of time series analysis. The object is thus to interpret a series of observed data points $\{X_t\}$, for example a historical price or volatility evolution (in this way acquiring a fundamental understanding of the process) and to *model* the processes underlying the observed historical evolution. In this sense, the historical sequence of data points is interpreted as just one *realization* of the time series process. The parameters of the process are then estimated from the available data and can subsequently be used in making *forecasts*, for example.

As much structure as possible should be extracted from a given data sequence and then transferred to the model. Let $\{\widehat{X}_t\}$ be the time series generated by the model process (called the *estimated* time series). The difference between this and the actually observed data points $\{X_t\}$ are called *residues* $\{X_t - \widehat{X}_t\}$. These should consist of only "noise", i.e., they should be unpredictable random numbers.

In order to be able to fit a time series model, the "raw data", i.e., the sequence of historical data points, must sometimes undergo a *pre-treatment*. In this procedure, *trends* and *seasonal components* are first eliminated and a change may

be made to the scale of the data, so that the resulting sequence is a *stationary time series.*[1] A stationary time series is characterized by the *time* invariance of its expectation, variance and covariance. In particular, the expectation and variance are constant. Without loss of generality, the expectation can be assumed to be zero since it can be eliminated during the pre-treatment through a *centering of the time series.* This is accomplished by subtracting the mean $\overline{X} = \frac{1}{T} \sum_{t=1}^{T} X_t$ from every data point in the time series $\{X_t\}$.

As just discussed, the *stationarity* of a time series implies $E[X_t] = E[X] \; \forall t$ and the autocovariance Eq. 31.14 becomes

$$\text{cov}(X_{t+h}, X_t) = E[X_{t+h} \, X_t] - E[X]E[X] = E[X_{t+h} \, X_t] \, . \qquad (32.1)$$

The final equality in the above equation holds if the time series has been centered in the pre-treatment. We will always assume this to be the case. Furthermore, the autocovariance and autocorrelation (just as the variance) are independent of t if the time series is stationary, and therefore depend only on the *time lag h*. We frequently write

$$\gamma(h) := \text{cov}(X_{t+h}, X_t)$$

Likewise, if the time series is stationary we have $\varrho(t, h) = \varrho(h)$ in the autocorrelation Eq. 31.13. The following useful symmetry relations can be derived directly from the stationarity of the time series (this can be shown by substituting t with $t' = t - h$):

$$\gamma(-h) = \gamma(h) \quad , \qquad \varrho(-h) = \varrho(h) \, . \qquad (32.2)$$

From definition 31.14, we can immediately obtain an estimate[2] for the autocorrelation and the autocovariance of a stationary data sequence

$$\widehat{\gamma}(h) = \widehat{\text{cov}}(X_{t+h}, X_t) = \frac{1}{T} \sum_{t=1}^{T-h} (X_{t+h} - \langle X \rangle)(X_t - \langle X \rangle) \quad , \qquad \widehat{\varrho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}$$

$$(32.3)$$

---

[1]More precisely, we are dealing with a *weakly stationary* time series in what follows.

[2]In the following material, we will distinguish the estimator of a parameter from the parameter itself with a "hat" notation.
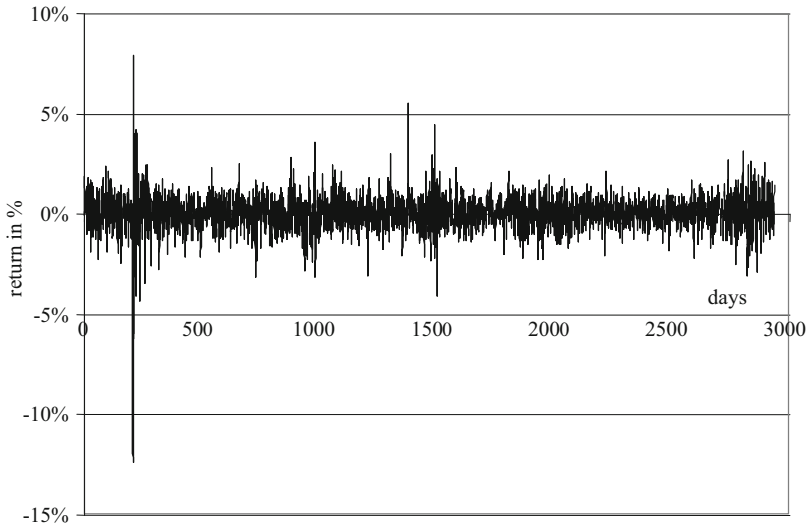
Fig. 32.1 Daily returns of the FTSE index as an example of a stationary data series. The crash in October 1987 is clearly visible

for $h \in \mathbb{N}_0$. The autocovariances (and autocorrelations) are usually computed for at most $h \leq 40$. Note that h has to be substantially smaller than T in all cases; the estimation is otherwise too inexact.[3]

Of course, we can fit *different* time series models to a stationary time series (after having undergone a pre-treatment if necessary) and then compare their goodness of fit and forecasting performance. Thus the following three general steps must be taken when modeling a given sequence of data points

1. Pre-treatment of the data sequence to generate a stationary series (elimination of trend and seasonal components, transformation of scale, etc.).
2. Estimation and/or fitting of the time series model and its parameters.
3. Evaluation of the goodness of fit and forecasting performance on the basis of which a decision is made as to whether the tested model should be accepted or a new model selected (step 2).

Figure 32.1 shows the daily relative change (returns) of the FTSE Index taken from the daily data from Jan-01-1987 through Apr-01-1998 (2,935 days).

---

[3]The fact that only T appears in the denominator in Eq. 32.3 instead of $T - h$, as one might expect, guarantees that the estimator for the covariance matrix $[\hat{\gamma}(i - j)]_{i,j=1}^{T}$ is automatically positive definite.

This sequence of data points is defined as

$$X_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \; , \tag{32.4}$$

where $\{Y_t\}$ represents the original data sequence of FTSE values. The data set $\{X_t\}$ consists of 2,934 values. According to Eq. 30.9, the relative changes in Eq. 32.4 are approximately equal to the difference of the logarithms if the daily changes are sufficiently small:

$$X_t \approx \ln(Y_t) - \ln(Y_{t-1}) \; . \tag{32.5}$$

This is the first difference of the logarithm of the original sequence of FTSE index values. The above example represents a typical pre-treatment procedure performed on the data. Instead of the original data $\{Y_t\}$, which is by no means stationary (drift$\neq 0$ and variance increase with time as $\sim \sigma t$), we generate a stationary data sequence as in Eq. 32.5 through standard transformations in time series analysis. Specifically in our case, what is known as *Box-Cox scaling* (taking the logarithm of the original data) was performed and subsequently the first differences were calculated for the purpose of trend elimination. Stationary time series data like these are then used in the further analysis, in particular, when fitting a model to the data.

The above example should provide sufficient motivation for the pre-treatment of a time series. The interested reader is referred to Chap. 35 for further discussion of pre-treating time series data to generate stationary time series. We will assume from now on that the given time series have already been pre-treated, i.e., potential trends and seasonal components have already been eliminated and scaling transformations have already been performed appropriately, so that the resulting data sequences are stationary. Such a stationary time series is given by a sequence of random variables $\{X_t\}$, $t \in \mathbb{N}$.

32.1 Stationary Time Series and Autoregressive Models

This chapter introduces a basic approach in time series analysis employing a specific time series model, called autoregressive model. We then continue by extending the results to the case of a time-dependent variance (GARCH model) which finds application in modeling *volatility clustering* in financial

time series. This technique is widely used in modeling the *time evolution* of volatilities.

Rather than working under the idealized assumption of time-continuous processes, the processes modeled in this chapter are truly discrete in time. The discussion is geared to the needs of the user. We will forgo mathematical rigor and in most cases the proofs of results will not be given. Not taking these "shortcuts" would increase the expanse of this chapter considerably. However, the attempt will be made to provide thorough reasoning for all results presented.

A process for modeling a time series of stock prices, for example, has already been encountered in this text: the random walk. An important property of the random walk is the *Markov property*. Recall that the Markov property states that the next step in a random walk depends solely on its current value, but not on the values taken on at any previous times. If such a Markov process is unsatisfactory for modeling the properties of the time series under consideration, an obvious generalization would be to allow for the influence of past values of the process. Processes whose current values can be affected by values attained in the past are called *autoregressive.* In order to characterize these processes, we must first distinguish between the *unconditional* and *conditional* variance denoted by $\text{var}[X_t]$ and $\text{var}[X_t|X_{t-1}, \ldots, X_1]$, respectively. The *un*conditional variance is the variance we are familiar with from previous chapters, whereas the conditional variance is the variance of X_t *under the condition* that $X_{t-1}, \ldots, X_1$ have occurred. Analogously, we must differentiate between the *conditional* and *unconditional* expectation denoted by $\text{E}[X_t]$ and $\text{E}[X_t|X_{t-1}, \ldots, X_1]$, respectively, where the last is the expectation of X_t *under the condition* that $X_{t-1}, \ldots, X_1$ have occurred. There is no difference between the two when the process under consideration is independent of its history.

32.1.1 AR(p) Processes

Having made these preparatory remarks and definitions, we now want to consider processes whose current values are influenced by one or more of their predecessors. If, for example, the effect of the p previous values of a time series on the current value is *linear*, the process is referred to as an *autoregressive process of order p,* and denoted by AR(p). The general autoregressive process of pth order makes use of p process values in the past to generate a representation of

today's value, or explicitly

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots \phi_p X_{t-p} + \varepsilon_t \tag{32.6}$$

$$= \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma^2) .$$

The changes ε_t here are independent of all previous time series values X_s, $s < t$, and thus represent an injection of truly new information into the process[4]. In particular, this means that $\mathrm{cov}[X_i, \varepsilon_j]$ is always zero. The conditional variance and conditional expectation of the process are

$$\mathrm{E}[X_t | X_{t-1}, \ldots, X_1] = \sum_{i=1}^{p} \phi_i X_{t-i} \tag{32.7}$$

$$\mathrm{var}[X_t | X_{t-1}, \ldots, X_1] = \mathrm{var}[\varepsilon_t] = \sigma^2 .$$

It can be shown that stationarity is guaranteed if the zeros z_k of the *characteristic polynomial*[5]

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \tag{32.8}$$

lie outside of the closed unit disk, i.e., when the norm $|z_k|$ is larger than 1 for all zeros z_k. In particular, if the process is stationary then the *un*conditional expectation and variance have the following properties: $\mathrm{E}[X_t] = \mathrm{E}[X_{t-i}]$ and $\mathrm{var}[X_t] = \mathrm{var}[X_{t-i}]$. Exploiting this, we can easily calculate explicit expressions for the *un*conditional expectation and variance. The unconditional expectation $\mathrm{E}[X_t]$ is

$$\mathrm{E}[X_t] = \mathrm{E}[\sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t] = \sum_{i=1}^{p} \phi_i \underbrace{\mathrm{E}[X_{t-i}]}_{\mathrm{E}[X_t]} + \underbrace{\mathrm{E}[\varepsilon_t]}_{0} = \mathrm{E}[X_t] \sum_{i=1}^{p} \phi_i .$$

In the first step we have simply used definition 32.6 for X_t. The second step is merely the linearity of the expectation operator. In the third step we have finally used the decisive properties of the process, namely stationarity of the

---

[4]The notation ε_t will always indicate independent, identically $\mathrm{N}(0, \sigma^2)$-distributed random variables. Another common definition is $\varepsilon_t \sim \mathrm{W}(0, \sigma^2)$, where W stands for *white noise*. This is a somewhat more general statement and is used in reference to random variables which are not normally distributed as well.
[5]This polynomial plays a central role in the theory of time series.

expectation and randomness of the residues. The result is therefore

$$E[X_t]\left(1 - \sum_{i=1}^{p}\phi_i\right) = 0$$

This implies that the unconditional expectation must be zero since stationarity guarantees that the sum of the ϕ_i is *not* equal to one.[6]

The unconditional variance can be computed using similar arguments

$$\text{var}[X_t] = \text{var}[\sum_{i=1}^{p}\phi_i X_{t-i} + \varepsilon_t]$$

$$= \sum_{i,j=1}^{p}\phi_i\phi_j\text{cov}[X_{t-i}, X_{t-j}] + \sum_{i=1}^{p}\phi_i\text{cov}[X_{t-i}, \varepsilon_t] + \text{var}[\varepsilon_t]$$

$$= \sum_{i,j=1}^{p}\phi_i\phi_j\text{cov}[X_{t-i}, X_{t-j}] + 0 + \sigma^2$$

$$= \text{var}[X_t]\sum_{i,j=1}^{p}\phi_i\phi_j\varrho(i - j) + \sigma^2 ,$$

where we used Eq. A.17 and—in the last step—definition 31.13 for stationary processes. Solving for $\text{var}[X_t]$ yields immediately

$$\text{var}[X_t] = \frac{\sigma^2}{1 - \sum_{i,j=1}^{p}\phi_i\phi_j\varrho(i - j)} . \tag{32.9}$$

An expression for the autocorrelation function ϱ of the process can be obtained by multiplying both sides of Eq. 32.6 by X_{t-h} and taking the expectation. Here stationarity is used in form of Eqs. 32.2 and 32.1:

$$\varrho(h) = \varrho(-h) = \frac{\text{cov}(X_{t-h}, X_t)}{\text{cov}(X_t, X_t)} = \frac{E(X_{t-h}, X_t)}{E(X_t^2)}$$

$$= \frac{1}{E(X_t^2)}E(X_{t-h}, \sum_{i=1}^{p}\phi_i X_{t-i} + \varepsilon_t)$$

---

[6]This can be shown using the characteristic polynomial.

$$= \frac{1}{\mathrm{E}(X_t^2)} \sum_{i=1}^{p} \phi_i \mathrm{E}(X_{t-h}, X_{t-i}) + \frac{1}{\mathrm{E}(X_t^2)} \underbrace{\mathrm{E}(X_{t-h}, \varepsilon_t)}_{0}$$

$$= \sum_{i=1}^{p} \phi_i \frac{\mathrm{E}(X_{t-h+i}, X_t)}{\mathrm{E}(X_t^2)} = \sum_{i=1}^{p} \phi_i \frac{\mathrm{E}(X_{t-(h-i)}, X_t)}{\mathrm{E}(X_t^2)} \ ,$$

and thus

$$\varrho(h) = \sum_{i=1}^{p} \phi_i \varrho(h - i) \ . \tag{32.10}$$

These are the *Yule-Walker equations* for the autocorrelations ϱ. The autocorrelations can thus be computed recursively by setting the initial condition $\varrho(0) = 1$. Consider the following example of an AR(2) process:

$$\varrho(1) = \phi_1 \varrho(1 - 1) + \phi_2 \varrho(1 - 2) = \phi_1 1 + \phi_2 \varrho(1) \Rightarrow \ \varrho(1) = \frac{\phi_1}{1 - \phi_2}$$

$$\varrho(2) = \phi_1 \varrho(1) + \phi_2 \varrho(0) = \frac{\phi_1^2}{1 - \phi_2} + \phi_2 \ , \text{ and so on.}$$

Here, the symmetry indicated in Eq. 32.2 was used together with Eq. 32.10. Substituting these autocorrelations into Eq. 32.9 finally yields the unconditional variance of an AR(2) process:

$$\mathrm{var}[X_t] = \frac{\sigma^2}{1 - \phi_1^2 \varrho(1 - 1) - \phi_1 \phi_2 \varrho(1 - 2) - \phi_2 \phi_1 \varrho(2 - 1) - \phi_2^2 \varrho(2 - 2)}$$

$$= \frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1 \phi_2 \varrho(1)} = \frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1^2 \phi_2/(1 - \phi_2)}$$

In practice, however, the autocorrelations should be computed from the original data series itself with the aid of Eq. 32.3, instead of from the coefficients ϕ_i, $i = 1, 2, \ldots, p$ which themselves are only estimated values.

The Autoregressive Process of First Order

We now consider the most simple case, namely $p = 1$. Explicitly, the *autoregressive process of first order AR(1)* is defined as

$$X_t = \phi X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) . \tag{32.11}$$

The stationarity condition for this process implies that $|\phi| < 1$ since Eq. 32.8 states simply that

$$1 - \phi z = 0 \text{ for some } z \text{ where } |z| > 1 .$$

The conditional variance and conditional expectation of the process are

$$E[X_t | X_{t-1}, \ldots, X_1] = \phi X_{t-1}$$
$$\text{var}[X_t | X_{t-1}, \ldots, X_1] = \text{var}[\varepsilon_t] = \sigma^2 .$$

The *un*conditional expectation is equal to zero as was shown above to hold for general AR(p) processes. The *un*conditional variance can be calculated as

$$\text{var}[X_t] = \text{var}[\phi X_{t-1} + \varepsilon_t]$$
$$= \phi^2 \text{var}[X_{t-1}] + \phi \text{cov}[X_{t-1}, \varepsilon_t] + \text{var}[\varepsilon_t]$$
$$= \phi^2 \text{var}[X_t] + 0 + \sigma^2 \implies$$
$$\text{var}[X_t] = \frac{\sigma^2}{1 - \phi^2} . . \tag{32.12}$$

Recursively constructing future values via Eq. 32.11 starting from X_t yields

$$X_{t+h} = \phi X_{t+h-1} + \varepsilon_{t+h}$$
$$= \phi^2 X_{t+h-2} + \phi \varepsilon_{t+h-1} + \varepsilon_{t+h}$$
$$\cdots$$
$$= \phi^h X_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}$$

The autocovariance of the AR(1) thus becomes explicitly

$$\text{cov}(X_{t+h}, X_t) = \text{cov}(\phi^h X_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}, X_t)$$

$$= \phi^h \text{cov}(X_t, X_t) + \sum_{i=0}^{h-1} \phi^i \underbrace{\text{cov}(\varepsilon_{t+h-i}, X_t)}_{0}$$

$$= \phi^h \text{var}[X_t]$$

$$= \phi^h \frac{\sigma^2}{1 - \phi^2} \ .$$

The auto*correlation* is therefore simply ϕ^h, and as such is an exponentially decreasing function of h. The same result can of course be obtained from the Yule-Walker equations

$$\varrho(h) = \sum_{i=1}^{1} \phi_i \varrho(h-i) = \phi \varrho(h-1) = \phi^2 \varrho(h-2) = \cdots = \phi^h \underbrace{\varrho(h-h)}_{1} \ .$$

It is worthwhile to consider a *random walk* from this point of view. A (one-dimensional) random walk is by definition constructed by adding an independent, identically distributed random variable (*iid*, for short) with variance σ^2 to the last value of attained in the walk. The random walk can thus be written as

$$X_t = X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{N}(0, \sigma^2) \ .$$

It follows from this definition that the conditional variance of the random walk is σ^2 and the expectation equals zero. The random walk corresponds to an AR(1) process with $\phi = 1$. This contradicts the stationarity criterion $|\phi| < 1$! The random walk is therefore a *non*-stationary AR(1) process. The non-stationarity can be seen explicitly by considering the *un*conditional variance:

$$\text{var}[X_t] = \text{var}[X_{t-1} + \varepsilon_t]$$

$$= \text{var}[X_{t-1}] + \text{cov}[X_{t-1}, \varepsilon_t] + \text{var}[\varepsilon_t]$$

$$= \text{var}[X_{t-1}] + 0 + \sigma^2 \ .$$

Thus, for all $\sigma \neq 0$ we have $\mathrm{var}[X_{t-1}] \neq \mathrm{var}[X_t]$, i.e., the process *cannot* be stationary. Therefore we cannot obtain a closed from expression similar to Eq. 32.12 for the unconditional variance (this can also be seen from the fact that if $\phi = 1$, Eq. 32.12 would imply a division by zero). The unconditional variance can, however, be determined recursively

$$\mathrm{var}[X_t] = k\sigma^2 + \mathrm{var}[X_{t-k}] .$$

Assuming from the outset that a value $X_{t=0}$ is known (and because it is known, has zero variance) we obtain the well-known property of the random walk

$$\mathrm{var}[X_t] = t\sigma^2 .$$

The variance is thus time dependent; this is a further indication that the random walk is not stationary. Since the variance is linear in the time variable, the standard deviation is proportional to the square root of time. This is the well-known *square root law* for scaling the volatility with time.

Another special case of an AR(1) process is *white noise* which has an expectation equal to zero and constant variance. It is defined by

$$X_t = \varepsilon_t .$$

The random variables $\{\varepsilon_t\}$ are *iid* random variables with variance σ^2. This corresponds to the AR(1) process with $\phi = 0$. The stationarity criterion $|\phi| < 1$ is satisfied and the above results for the stationary AR(1) process can be applied with $\phi = 0$, for example, $\mathrm{cov}(X_{t+h}, X_t) = 0$ and $\mathrm{var}(X_t) = \sigma^2$.

32.1.2 Univariate GARCH(p, q) Processes

The conditional variance of the AR(p) processes introduced above was always a constant function of time; in each case it was equal to the variance of ε_t. This, however, is not usually the case for financial time series. Take, for example, the returns of the FTSE data set in Fig. 32.1. It is clear to see that the variance of the data sequence is not constant as a function of time. On the contrary, the process goes through both calm and quite volatile periods. It is much more probable that large price swings will occur close to other large price swings than to small swings. This behavior is typical of financial time series and is referred to as *volatility clustering* or simply *clustering*. A process which is capable of modeling such behavior is the *GARCH(p, q) process*, which will

be introduced below. The decisive difference between GARCH and AR(p) processes is that not only past values of X_t are used in the construction of a GARCH process, but past values of the *variance* enter into the construction as well. The GARCH(p, q) process is defined as

$$X_t = \sqrt{H_t}\,\varepsilon_t \quad \text{with} \quad H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i\, H_{t-i} + \sum_{j=1}^{q} \alpha_j\, X_{t-j}^2, \quad \varepsilon_t \sim N(0, 1) \,,$$

(32.13)

where the $\{\varepsilon_t\}$ are *iid* standard normally distributed. The $\{\varepsilon_t\}$ are independent of X_t. Therefore, the time series $\{X_t\}$ is nothing else than white noise $\{\varepsilon_t\}$ with a time-dependent variance which is determined by the coefficients $\{H_t\}$. These H_t take into consideration the past values of the time series *and* the variance. If the $\{X_t\}$ are large (distant from the equilibrium value which is in this case zero as $E[\varepsilon_t] = 0$), then so is $\{H_t\}$. For small values $\{X_t\}$ the opposite holds. In this way, clustering can be modeled. The order q indicates how many past values of the time series $\{X_t\}$ influence the current value H_t. Correspondingly, p is the number of past values of the variance itself which affects the current value of H_t. In order to ensure that the variance is positive, the parameters must satisfy the following conditions:

$$\alpha_0 \geq 0 \tag{32.14}$$

$$\beta_1 \geq 0$$

$$\sum_{j=0}^{k} \alpha_{j+1}\beta_1^{k-j} \geq 0, \quad k = 0, \ldots, q-1 \,.$$

This implies that $\alpha_1 \geq 0$ always holds, the other α_i however, may be negative. Furthermore, the time series $\{X_t\}$ should be (weakly) stationary to prevent it from "drifting away". The following condition is sufficient to guarantee this stationarity:

$$\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j < 1 \,. \tag{32.15}$$

The two most important properties of this process pertain to the conditional expectation and the conditional variance

$$E[X_t|X_{t-1}, \ldots, X_1] = 0 \quad \text{and} \tag{32.16}$$

$$\text{var}[X_t|X_{t-1}, \ldots, X_1] = H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2$$

The first equation holds because $E[\varepsilon_t] = 0$, the second because $\text{var}[\varepsilon_t] = 1$. The H_t are thus the conditional variances of the process. The conditional expectation (under the condition that all X up to time $t - 1$ are known) of H_t is simply H_t itself since no stochastic variable ε appears in Eq. 32.13 where H_t is defined, and thus

$$E[H_t|X_{t-1}, \ldots, X_1] = H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2 . \tag{32.17}$$

H is thus always known one time step in advance of X. This may seem trivial but will be quite useful in Sect. 33.2 when making volatility forecasts.

The *un*conditional variance is by definition

$$\text{var}[X_t] = E[X_t^2] - E[X_t]^2$$
$$= E[H_t \varepsilon_t^2] - E[\sqrt{H_t} \varepsilon_t]^2$$
$$= E[H_t]E[\varepsilon_t^2] - (E[\sqrt{H_t}]E[\varepsilon_t])^2 ,$$

where in the last step we have made use of the fact that $\{\varepsilon_t\}$ are uncorrelated with $\{H_t\}$. Furthermore, since the $\{\varepsilon_t\}$ are *iid* $N(0, 1)$ distributed

$$E[\varepsilon_t] = 0 \quad \text{and} \quad E[\varepsilon_t^2] = E[\varepsilon_t^2] - 0 = E[\varepsilon_t^2] - (E[\varepsilon_t])^2 = \text{var}[\varepsilon_t] = 1$$

and therefore

$$\text{var}[X_t] = E[H_t] = \alpha_0 + \sum_{i=1}^{p} \beta_i E[H_{t-i}] + \sum_{j=1}^{q} \alpha_j E[X_{t-j}^2] .$$

Just as the ε_t, the X_t have zero expectation, which also implies that $E[X_t^2] = \text{var}[X_t]$. Under consideration of this relation and the stationarity (constant

variance), all of the expectations involving squared terms in the above equation can be written as the variance of X_t:

$$E[X_{t-j}^2] = \text{var}[X_{t-j}] = \text{var}[X_t]$$

$$E[H_{t-i}] = \text{var}[X_{t-j}] = \text{var}[X_t] \, .$$

This leads to the following equation for the unconditional variance:

$$\text{var}[X_t] = \alpha_0 + \sum_{i=1}^{p} \beta_i \text{var}[X_t] + \sum_{j=1}^{q} \alpha_j \text{var}[X_t] \Longleftrightarrow$$

$$\text{var}[X_t] = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j} =: \widetilde{\alpha}_0 \, . \tag{32.18}$$

This unconditional variance can of course also be estimated from the observed data, i.e., as usual through the computation of the empirical variance estimator over a large number of realizations of $\{X_t\}$.

The GARCH(p, q) process can be expressed in terms of the unconditional variance $\widetilde{\alpha}_0$ as follows:

$$H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2$$

$$= \alpha_0 \frac{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j} + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2$$

$$= \widetilde{\alpha}_0 - \widetilde{\alpha}_0 \sum_{i=1}^{q} \alpha_i - \widetilde{\alpha}_0 \sum_{j=1}^{p} \beta_j + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2$$

$$= \widetilde{\alpha}_0 + \sum_{i=1}^{p} \beta_i (H_{t-i} - \widetilde{\alpha}_0) + \sum_{j=1}^{q} \alpha_j (X_{t-j}^2 - \widetilde{\alpha}_0) \, . \tag{32.19}$$

The conditional variance H_t can thus be interpreted as the unconditional variance $\widetilde{\alpha}_0$ plus the sum of the distances from this unconditional variance. If all α_j and β_i are greater than zero (which is always the case for a GARCH(1,1) process), this form of the conditional variance has another interpretation:

The β_i terms cause a kind of *persistence* of the variance which serves to model the *volatility clustering* phenomenon: the greater H_{t-i} becomes in comparison to the long-term expectation $\widetilde{\alpha}_0$ (the unconditional variance), the greater the positive contribution of these terms to H_t; the H_t tend to get even larger. Conversely, for values of H_{t-i} which are smaller than $\widetilde{\alpha}_0$ the contribution of these terms become negative and thus H_t will tend to get even smaller.

The terms involving α_i describe the *reaction* of the volatility to the process itself. Values X_{t-j}^2 larger than $\widetilde{\alpha}_0$ favor a growth in the variance; values X_{t-j}^2 smaller than $\widetilde{\alpha}_0$ result in a negative contribution and thus favor a decline in the variance. If the process itself describes a price *change*, as is common in the financial world, this is precisely the effect that strong price changes tend to induce growth in volatility.

Overall, these properties lead us to expect that GARCH models are indeed an appropriate choice for modeling certain phenomena observed in the financial markets (in particular, volatility clustering and the reaction of the volatility to price changes). In practice, we often set $p = 1$ and even $q = 1$. It has been shown that significantly better results are not achieved with larger values of p and q and thus the number of parameters to be estimated would be unnecessarily increased.

32.1.3 Simulation of GARCH Processes

One of the examples to be found in the Excel workbook GARCH.xlsx from the download section [50] is the simulation of a *GARCH(1,1) process*. The first simulated value X_1 of the time series is obtained, according to Eq. 32.13, from a realization of a standard normal random variable followed by multiplication of this number by $\sqrt{H_1}$. Subsequently, H_2 is computed from the values now known at time $t = 1$. Then, a realization X_2 is generated from a standard normal random variable and multiplied by $\sqrt{H_2}$. This procedure is repeated until the end of the time series is reached.

In order to generate a GARCH(p, q) process, q values of the X process

$$X_{-q+1}, X_{-q+2}, \ldots, X_0$$

and p values of the conditional variance

$$H_{-p+1}, H_{-p+2}, \ldots, H_0$$
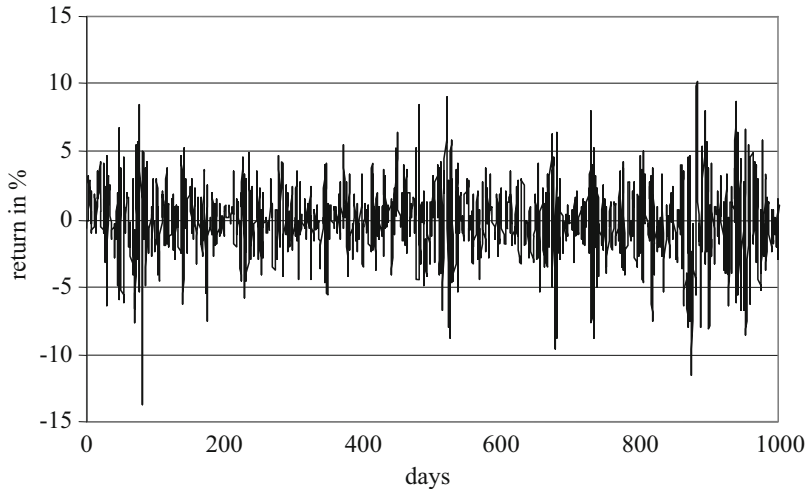
Fig. 32.2 Simulated GARCH(1,1) process. The first 100 values have not been used. Clustering can clearly be observed

must be given in order to be able to compute the first conditional variance H_1 as indicated in Eq. 32.13. The choice of these initial values is not unique but the orders of magnitude of the time series values and the variances should at least be correct. The unconditional expectation $E[X_t]$ and the unconditional variance $var[X_t]$ are therefore good candidates for this choice. The first values of the generated time series should then be rejected (often, 50 values are sufficient), since they still include the above "initial conditions". After taking several steps, realizations of the desired GARCH process can be generated. Figure 32.2 illustrates a simulated GARCH process.

Such simulated time series can be implemented to test optimization methods which have the objective of "finding" parameters from the simulated data series which have been previously used for the simulation. After all, if a data set is given (real or simulated), the parameters of a model have to be determined. Methods for doing this are the subject of the next section.

32.2 Calibration of Time Series Models

All of the time series models introduced above include parameters which may be varied for the purpose of fitting the model "optimally" to the time series data. We represent these parameters as a parameter vector θ. For an AR(p) process, the free parameters are the ϕ_i and σ^2 while the GARCH(p, q) has

the free parameters α_i and β_i. Thus

$$\theta = (\phi_1, \phi_2, \ldots, \phi_p, \sigma^2) \qquad \text{for AR}(p)$$
$$\theta = (\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \beta_2, \ldots \beta_p) \text{ for GARCH}(p, q) \ .$$

A widely used estimation procedure for the determination of unknown parameters in statistics is the *maximum likelihood estimator.* This procedure selects the parameter values which maximize the likelihood of the model being correct. These are just the parameter values which maximize the *probability* (called the *likelihood*) that the values observed will be realized by the assumed model. Using the model, the probability is expressed as a function of the parameters θ. Then this probability function is maximized by varying the parameter values. The parameter values for which the probability function attains a maximum corresponds to a "best fit" of the model to the given data sequence. They are the most probable parameter values given the information available (i.e., given the available time series). This procedure will now be performed explicitly for both an AR(p) and a GARCH(p,q) process.

32.2.1 Parameter Estimation for AR(p) Processes

The likelihood for the AR(p) process is obtained as follows: from Eqs. 32.6 and 32.7 we can see that if we assume an AR(p) process with a parameter vector

$$\theta = (\phi_1, \phi_2, \ldots, \phi_p, \sigma^2)$$

then X_t has the normal distribution

$$\mathrm{N}(\sum_{i=1}^{p} \phi_i X_{t-i}, \sigma^2)$$

The conditional probability for one single observed value of X_t (also called the *conditional likelihood* of X_t) is thus

$$L_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2 \right\} \ .$$

The total likelihood for all T measured data points is in consequence of the independence of ε_t simply a product of all conditional likelihoods:

$$L_\theta(X_1, X_2, \ldots, X_T) = \prod_{t=1}^{T} L_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p})$$

$$= \frac{1}{(2\pi)^{T/2} \sigma^T} \prod_{t=1}^{T} \exp\left\{ -\frac{1}{2\sigma^2} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2 \right\}.$$

Observe that for the likelihoods of the first data points X_t where $t < p + 1$, a further p data points $\{X_0, X_{-1}, \ldots, X_{-p+1}\}$ are required in advance. The extent of the data sequence needed is thus a data set encompassing $T + p$ data points.

Maximizing this likelihood through the variation of the parameters $\phi_1, \phi_2, \ldots, \phi_p$ and σ^2, we obtain the parameters $\{\phi_1, \phi_2, \ldots, \phi_p, \sigma^2\}$ which, under the given model assumptions,[7] actually maximizes the (model) probability that the observed realization $\{X_t\}$ will actually appear. It is, however, simpler to maximize the *logarithm* of the likelihood (because of the size of the terms involved and the fact that sums are more easily dealt with than products). Since the logarithm function is strictly monotone increasing, the maximum of the likelihood function is attained for the same parameter values as the maximum of the logarithm of the likelihood function. The *log-likelihood* function for the AR(p) process is given by

$$\mathcal{L}_\theta = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2.$$

$\phi_1, \phi_2, \ldots, \phi_p$ appear only in the last expression (the sum), which appears with a negative sign in the log-likelihood function. The values of $\phi_1, \phi_2, \ldots, \phi_p$ which maximize the log-likelihood function therefore minimize the expression

$$\sum_{t=1}^{T} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2 \tag{32.20}$$

---

[7] The model assumption is that the time series was generated by an AR(p) process.

This, however, is just a sum of the quadratic deviations. The desired parameter estimates $\{\widehat{\phi}_1, \widehat{\phi}_2, \ldots, \widehat{\phi}_p\}$ are thus the solution to a least squares problem. The $\widehat{\phi}_t$ can thus be determined independently from the variance σ^2. The estimation of the variance is obtained from simple calculus by taking the derivative of the log-likelihood function with respect to σ^2 and setting the resulting value equal to zero (after substituting the optimal ϕ_i, namely the $\widehat{\phi}_i$):

$$\frac{\partial \mathcal{L}_\theta}{\partial \sigma^2} = -\frac{T}{2} \frac{\partial \ln(2\pi\sigma^2)}{\partial \sigma^2} - \frac{\partial}{\partial \sigma^2} \left(\frac{1}{2\sigma^2} \sum_{t=1}^{T} \left[X_t - \sum_{i=1}^{p} \widehat{\phi}_i X_{t-i} \right]^2 \right)$$

$$= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^{T} \left[X_t - \sum_{i=1}^{p} \widehat{\phi}_i X_{t-i} \right]^2 \overset{!}{=} 0 .$$

The optimal estimate for σ^2 becomes

$$\widehat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \left[X_t - \sum_{i=1}^{p} \widehat{\phi}_i X_{t-i} \right]^2 . \tag{32.21}$$

For example, the maximum likelihood estimator for ϕ_1 in the AR(1) process in Eq. 32.11, obtained by minimizing the expression in 32.20, can be determined through the following computation:

$$0 = \frac{\partial}{\partial \phi_1} \sum_{t=1}^{T} [X_t - \phi_1 X_{t-1}]^2 = 2 \sum_{t=1}^{T} (X_t - \phi_1 X_{t-1})(-X_{t-1})$$

$$= 2\phi_1 \sum_{j=1}^{T} X_{j-1}^2 - 2 \sum_{t=1}^{T} X_{t-1} X_t \implies$$

$$\widehat{\phi}_1 = \frac{\sum_{t=1}^{T} X_{t-1} X_t}{\sum_{j=1}^{T} X_{j-1}^2} .$$

Substituting this into Eq. 32.21 yields the maximum likelihood estimator for σ^2

$$\widehat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \left[X_t - \widehat{\phi}_1 X_{t-1} \right]^2$$

$$= \frac{1}{T} \sum_{t=1}^{T} \left[X_t - X_{t-1} \frac{\sum_{i=1}^{T} X_{i-1} X_i}{\sum_{j=1}^{T} X_{j-1}^2} \right]^2 .$$

32.2.2   Parameter Estimation for GARCH(p, q) Processes

The likelihood for the GARCH(p, q) process is obtained as follows: from Eqs. 32.13 and 32.16 we see that

$$X_t | \{X_{t-1}, \ldots, X_{t-q}, H_{t-1}, \ldots, H_{t-p}\} \sim \mathrm{N}(0, H_t) .$$

This implies that, given the information $\{X_{t-1}, \ldots, X_{t-q}, H_{t-1}, \ldots, H_{t-p}\}$, X_t is normally distributed according to $\mathrm{N}(0, H_t)$. The conditional likelihood for one single observation X_t is then

$$L_\theta(X_t | \{X_{t-1}, \ldots, X_{t-q}, H_{t-1}, \ldots, H_{t-p}\}) = \frac{1}{\sqrt{2\pi H_t}} e^{-X_t^2/2H_t}$$

where

$$H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2$$

and with a parameter vector

$$\theta = (\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p) .$$

The overall likelihood of all observations together is, in consequence of the independence of $\{\varepsilon_t\}$, merely the product

$$L_\theta = \prod_{t=1}^{T} L_\theta(X_t | \{X_{t-q}, \ldots, X_{t-1}, H_{t-p}, \ldots, H_{t-1}\}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi H_t}} e^{-X_t^2/2H_t} .$$

Observe that for the likelihood of the first data point X_1 further data points $\{X_0,\ X_{-1}, \ldots, X_{-q+1}, H_0, H_{-1}, \ldots, H_{-p+1}\}$ are required in advance. The total required data sequence $\{X_t\}$ thus encompasses $T + q$ data points. If $T + q$ observations of X_t are available, the first are required as information in advance, the remaining T are included in the likelihood function as observed data. In addition the values $\{H_0, H_{-1}, \ldots, H_{-p+1}\}$ are required as information in advance. In choosing the size of T it is necessary to make a compromise between the exactness of the estimator (T is chosen to be as large as possible) and the time scale with which the market mechanisms change (T is chosen to be as small as possible).

Maximizing this likelihood function by allowing the parameter values in θ to vary, we obtain the parameters which, under the model assumption (a GARCH(p, q) process), maximize the probability of a realization of the market values $\{X_t\}$ observed. It is again easier to work with the log-likelihood function in determining this maximum. Since the log function is strictly monotone increasing, the maximum of the likelihood and the log-likelihood function is attained at the same parameter point. The log-likelihood for the GARCH(p, q) process is given by

$$
\begin{aligned}
\mathcal{L}_\theta &= \sum_{t=1}^{T} \ln L_\theta \left(X_t | \{X_{t-q}, \ldots, X_{t-1}, H_{t-p}, \ldots, H_{t-1}\} \right) \\
&= \sum_{t=1}^{T} \ln \left(\frac{1}{\sqrt{2\pi H_t}} e^{-X_t^2/2H_t} \right) \\
&= -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \ln(H_t) - \frac{1}{2} \sum_{t=1}^{T} \frac{X_t^2}{H_t}
\end{aligned}
\tag{32.22}
$$

where

$$
H_t = H_t(\theta) = \alpha_0 + \sum_{j=1}^{p} \beta_j H_{t-j} + \sum_{k=1}^{q} \alpha_k X_{t-k}^2 \,.
$$

This is the function which must now be maximized through the variation of the parameter vector θ. The space of valid parameters θ is limited by the constraints stated in Eqs. 32.14 and 32.15. This represents an additional difficulty for the optimization. The optimization is quite difficult because (as opposed to the AR(1) process) maximizing the likelihood function cannot be

computed analytically but must be accomplished by means of a *numerical optimization* procedure. As the function to be maximized has multiple local maxima, a complex "likelihood surface" further complicates the optimization process since *local* optimization methods, such as gradient methods, are unsuitable if the initial value is not well chosen, i.e., if it does not lie close to the global maximum. A suitable algorithm for finding a *global* maximum in such a situation is *simulated annealing*.

32.2.3   Simulated Annealing

Simulated annealing is a numerical algorithm used to find a *global* minimum or maximum of a given function. Its construction is motivated by an effect observed in physics, namely cooling. The cooling of a physical body results in its moving through decreasing energy states traveling a path ending in a state of *minimum energy*. The simulated annealing algorithm attempts to imitate this process. The function whose minimum is to be found thus corresponds to the energy of the physical body.

As a physical body cools, the temperature T declines resulting in a steady loss of energy. The body is composed of billions of atoms which all make a contribution to its total energy. This being the case, there are a multitude of possible energy states with a multitude of *local* energy minima. If the temperature declines very *slowly*, the body surprisingly finds its *global* minimum (for example, the atoms in the body may assume a characteristic lattice configuration). A simple approach to this can be taken from *thermodynamics*: the probability of a body being in a state with energy E when the temperature of the body is T is proportional to the *Boltzmann factor,* $\exp(-E/kT)$:

$$P(E) \sim \exp\left(-\frac{E}{kT}\right) \ ,$$

where k is a thermodynamic constant, the *Boltzmann constant*. It follows that a higher energy state can be attained at a certain temperature though the probability of such an event declines with a decline in temperature. In this way, "unfavorable" energy states *can* be attained and thus the system *can* escape from local energy minima. However, if the temperature drops too quickly, the body remains in a so-called meta-stable state and cannot reach its global energy minimum.[8] It is therefore of utmost importance to cool the body *slowly*.

---

[8]Physicist speak in such cases of "frustrated" systems. An example of such a frustrated system is glass.

This strategy observed in nature is now to be simulated on a computer. In order to replicate the natural scenario, a configuration space (the domain of possible values of the pertinent parameters θ) must be defined. This might be a connected set but could also consist of discrete values (*combinatorial optimization*). In addition, a mechanism is required governing the transition from one configuration to another. And finally, we need a scheme for the cooling process controlling the decline in "temperature" T ($T_0 \rightarrow T_1 \rightarrow \cdots \rightarrow T_n \rightarrow \cdots$). The last two points mentioned are of particular importance; the change-of-configuration mechanism determine how efficient the configuration space is sampled while the second of the above requirements serves to realize the "slow cooling".

For each temperature the parameter sequence forms a *Markov chain*. Each new test point θ is accepted with the probability[9]

$$P = \min \left\{ e^{-[f(\theta_p) - f(\theta_{p-1})]/T}, 1 \right\}$$

where θ_{p-1} represents the previously accepted parameter configuration. The function f is the function to be minimized for each specific problem and is, for example, the (negative) log-likelihood function from Eq. 32.22. This function corresponds to the energy function in physics.

After having traveled a certain number of steps in the Markov chain, the temperature declines according to some mechanism which could for instance be as simple as

$$T_n = \alpha T_{n-1} \quad (0 < \alpha < 1) \ .$$

A new Markov chain is then started. The starting point for the new chain is the end point of the previous chain. In a concrete optimization, the temperature is naturally not to be understood in the physical sense; it is merely an abstract parameter directing the course of the optimization by controlling the transition probability in the Markov chain. However, we choose to retain the designations temperature or cooling scheme as a reminder of the procedure's origin. Figure 32.3 shows a schematic representation of the algorithm.

Simulated annealing is demonstrated in the Excel workbook GARCH.XLSX by means of a VBA program. The algorithm in the workbook is used to fit the

---

[9]The minimum function is only required since a probability can be at most equal to one.

$\theta_0 \quad \xrightarrow{\ P\ } \theta_1 \qquad \xrightarrow{\ P\ } \theta_2 \qquad \xrightarrow{\ P\ } \cdots \xrightarrow{\ P\ } \theta_n \qquad$ temperature T_1

cooling

$\theta_n \quad \xrightarrow{\ P\ } \theta_{n+1} \qquad \xrightarrow{\ P\ } \theta_{n+2} \qquad \xrightarrow{\ P\ } \cdots \xrightarrow{\ P\ } \theta_{2n} \qquad$ temperature T_2

cooling

$\theta_{2n} \quad \xrightarrow{\ P\ } \theta_{2n+1} \qquad \xrightarrow{\ P\ } \theta_{2n+2} \qquad \xrightarrow{\ P\ } \cdots \xrightarrow{\ P\ } \theta_{3n} \qquad$ temperature T_3

$\theta_{(m-1)n} \xrightarrow{\ P\ } \theta_{(m-1)n+1} \xrightarrow{\ P\ } \theta_{(m-1)n+2} \xrightarrow{\ P\ } \cdots \xrightarrow{\ P\ } \theta_{mn} \qquad$ temperature T_m
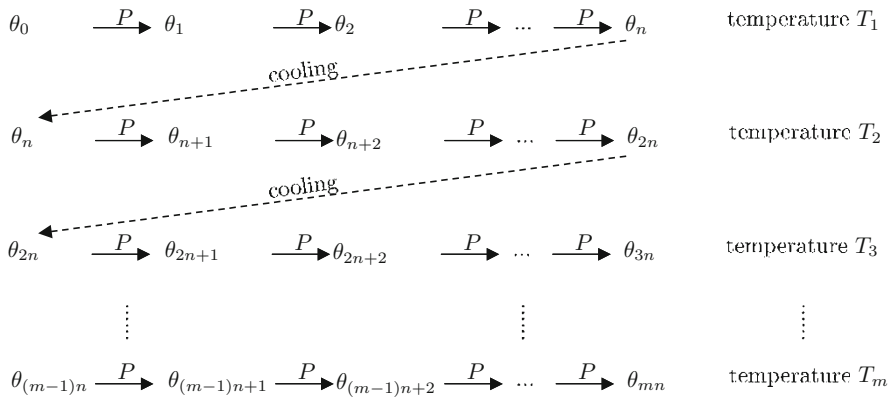
Fig. 32.3 Simulated annealing using m Markov chains with n steps in each chain. If the cooling is slow enough and m and n are large enough, then θ_{mn} is a good approximation of the parameter vector necessary to achieve the global minimum of the function f

parameters of a GARCH(1,1) process making use of the first 400 points of a given (simulated) data set. No emphasis is placed on the speed of computation since our object is to demonstrate the fundamental principles as clearly as possible.

33

Forecasting with Time Series Models

Having selected a model and fitted its parameters to a given times series, the model can then be used to estimate new data of the time series. If such data are estimated for a time period following the final data value X_T of the given time series, we speak of a *prediction* or *forecast*. The estimation of data lying between given data points is called *interpolation*. The question now arises as to how a model such as those given in Eqs. 32.6 or 32.13 could be used to obtain an "optimal" estimate. To answer this question the *forecasting error*

$$X_{T+k} - \widehat{X}_{T+k} , \quad k \in \mathbb{N}$$

between the estimated values $\widehat{X}_{T+k}$ and the actual observed time series values X_{T+k} can be used if the last value used in the calibration of the model was X_T. The best forecast is that which minimizes the *mean square error* (*MSE* for short). The MSE is defined as the expectation of the squared forecasting error

$$\text{MQF} := \text{E}[(X_{T+k} - \widehat{X}_{T+k})^2] . \tag{33.1}$$

This expression is the mathematical formulation of the intuitive concept of the "distance" between the estimated and the actual values which is to be minimized on average (more cannot be expected when dealing with random variables). Minimizing this mean square error yields the result that the best forecasting estimate (called the *optimal forecast*) is given by the conditional expectation

$$\widehat{X}_{T+k} = \text{E}[X_{T+k}|X_T, \ldots, X_2, X_1] . \tag{33.2}$$

This is the expectation of X_{T+k}, conditional on all available information about the time series up to and including T.

In practice, however, the concrete computation of this conditional expectation is generally very difficult since the joint distribution of the random variables must be known. Therefore, we often limit our consideration to the *linear forecast*

$$\widehat{X}_{T+k} = u_1 X_1 + u_2 X_2 + \cdots + u_T X_T \qquad (33.3)$$

with appropriate coefficients u_i. This linear forecast is, in contrast to the optimal forecast, often more easily interpreted. For the special case that the $\{X_t\}$ are normally distributed, the linear forecast and the optimal forecast agree. The *best* linear forecast can be characterized by the fact that the forecasting error $X_{T+k} - \widehat{X}_{T+k}$ and the $X_1, X_2, \ldots, X_T$ are uncorrelated. The intuitive interpretation is that the $X_1, X_2, \ldots, X_T$ cannot provide any additional information for the forecast and the error is thus purely random.

33.1 Forecasting with Autoregressive Models

This forecasting procedure will now be applied to an AR(p) process, Eq. 32.6. The optimal one-step forecast is, according to Eq. 33.2, given directly by the conditional expectation in Eq. 32.7

$$\widehat{X}_{T+1}^{\text{optimal}} = \mathrm{E}[X_{T+1}|X_T, \ldots, X_1] = \sum_{i=1}^{p} \phi_i X_{T+1-i} \ . \qquad (33.4)$$

This has the form indicated in Eq. 33.3. The optimal one-step forecast is thus the *best linear* one-step forecast. Equation 32.6 shows that the forecasting error $X_{T+1} - \widehat{X}_{T+1}$ is precisely ε_{T+1} and independent of $X_1, X_2, \ldots, X_T$. The MSE for the one-step forecast is given by Eq. 33.1 with $k = 1$. Thus,

$$\mathrm{MSE} = \mathrm{E}[(X_{T+1} - \widehat{X}_{T+1})^2] = \mathrm{E}[\varepsilon_{T+1}^2] = \sigma^2 \ .$$

The optimal two-step forecast is the conditional expectation of X_{T+2} on the basis of knowledge of $X_T, \ldots, X_1$, as can be seen from Eq. 33.2:

$$\widehat{X}_{T+2}^{\text{optimal}} = \mathrm{E}[X_{T+2}|X_T, \ldots, X_1] \ .$$

This is *not* the equivalent to the conditional expectation of X_{T+2} on the basis of knowledge of $X_{T+1}, \ldots, X_1$. Hence, Eq. 32.7 cannot be applied directly. An additional difficulty arises due to the fact that X_{T+1} is unknown. The *optimal* two-step forecast cannot be calculated. We therefore proceed by computing the *linear* two-step forecast. The best linear two-step forecast is obtained by actually calculating the conditional expectation of X_{T+2} as if all the $X_{T+1}, \ldots, X_1$ were known and replacing the (unknown) value X_{T+1} by its best estimate $\widehat{X}_{T+1}$ (which was calculated in the previous step):

$$\widehat{X}_{T+2}^{\text{linear}} = \mathrm{E}[X_{T+2}|\widehat{X}_{T+1}, X_T, \ldots, X_1] \, .$$

Now Eq. 32.7 can be applied to this conditional expectation and utilizing Eq. 33.4 we obtain

$$\widehat{X}_{T+2}^{\text{linear}} = \phi_1 \widehat{X}_{T+1} + \sum_{j=2}^{p} \phi_j X_{T+2-j} = \phi_1 \sum_{i=1}^{p} \phi_i X_{T+1-i} + \sum_{j=2}^{p} \phi_j X_{T+1-(j-1)}$$

$$= \phi_1 \sum_{i=1}^{p} \phi_i X_{T+1-i} + \sum_{i=1}^{p-1} \phi_{i+1} X_{T+1-i}$$

$$= \phi_1 \phi_p X_{T+1-p} + \sum_{i=1}^{p-1} [\phi_1 \phi_i + \phi_{i+1}] X_{T+1-i} \, .$$

The linear two-step forecast then has the form indicated in Eq. 33.3. The forecasting error is found to be

$$X_{T+2} - \widehat{X}_{T+2} = \underbrace{\sum_{i=1}^{p} \phi_i X_{T+2-i} + \varepsilon_{T+2}}_{X_{T+2}, \text{ see Eqn. 32.6}} - \phi_1 \widehat{X}_{T+1} - \sum_{j=2}^{p} \phi_j X_{T+2-j}$$

$$= \varepsilon_{T+2} + \phi_1 (X_{T+1} - \widehat{X}_{T+1})$$

$$= \varepsilon_{T+2} + \phi_1 \varepsilon_{T+1} \, .$$

Thus, the forecasting error is a sum of two normally distributed random variables and therefore itself normally distributed. The MSE can now be

computed as follows:

$$\text{MSE} = \text{E}[(X_{T+2} - \widehat{X}_{T+2})^2] = \text{E}[(\varepsilon_{T+2} + \phi_1 \varepsilon_{T+1})^2]$$

$$= \text{E}[\varepsilon_{T+2}^2] + \phi_1^2 \text{E}[\varepsilon_{T+1}^2] + 2\phi_1 \text{E}[\varepsilon_{T+2}\, \varepsilon_{T+1}]$$

$$= \text{var}[\varepsilon_{T+2}^2] + \phi_1^2 \, \text{var}[\varepsilon_{T+1}^2] + 2\phi_1 0$$

$$= \sigma^2(1 + \phi_1^2) \,.$$

This implies that the forecasting error $X_{T+2} - \widehat{X}_{T+2}$ of the two-step forecast is normally distributed with variance $\sigma^2(1 + \phi_1^2)$, i.e., $\text{N}(0, \sigma^2(1 + \phi_1^2))$.

Proceeding analogously, the best linear h-step forecast is obtained by taking the conditional expectation of X_{T+h} as if all X_t were known up to X_{T+h-1} and then replacing the yet unknown values of X_t for $T < t < h$ with their best estimators calculated inductively in previous steps as described above:

$$\widehat{X}_{T+h}^{\text{linear}} = \text{E}[X_{T+h} | \widehat{X}_{T+h-1}, \widehat{X}_{T+h-2}, \ldots, \widehat{X}_{T+1}, X_T, \ldots, X_1] \,.$$

Equation 32.7 is then applied to these conditional expectations resulting in

$$\widehat{X}_{T+h} = \sum_{i=1}^{\min(h-1,\, p)} \phi_i \widehat{X}_{T+h-i} + \sum_{j=h}^{p} \phi_j X_{T+h-j} \,.$$

The forecasting error of the h-step forecast is

$$X_{T+h} - \widehat{X}_{T+h} = \underbrace{\sum_{i=1}^{p} \phi_i X_{T+h-i} + \varepsilon_{T+h}}_{X_{T+h},\ \text{see Eqn. 32.6}} - \sum_{i=1}^{\min(h-1,p)} \phi_i \widehat{X}_{T+h-i} - \sum_{j=h}^{p} \phi_j X_{T+h-j}$$

$$= \varepsilon_{T+h} + \sum_{i=1}^{\min(h-1,p)} \phi_i X_{T+h-i} + \sum_{i=h}^{p} \phi_i X_{T+h-i}$$

$$- \sum_{i=1}^{\min(h-1,p)} \phi_i \widehat{X}_{T+h-i} - \sum_{j=h}^{p} \phi_j X_{T+h-j}$$

$$= \varepsilon_{T+h} + \sum_{i=1}^{\min(h-1,p)} \phi_i \left(X_{T+h-i} - \widehat{X}_{T+h-i} \right) \,.$$

This is a recursion expressing the h-step forecasting error in terms of the forecasting errors for fewer than h steps. From this it can be shown that the h-step forecasting error is distributed as $N(0, \sigma^2(1 + \phi_1^2 + \cdots + \phi_{h-1}^2))$.

The unknown coefficients $\phi_1, \phi_2, \ldots, \phi_p$ are estimated from the time series as shown in Sect. 32.2.1. The ϕ_i in the forecast equation are simply replaced with $\widehat{\phi}_i$.

33.2 Volatility Forecasts with GARCH(p, q) Processes

GARCH models of the form indicated in Eq. 32.13 are not suitable for the prediction of the actual values X_i of a time series since the random variable in Eq. 32.13 appears as a *product* (rather than a sum as in Eq. 32.6). In consequence, the conditional expectations of the X_i are identically zero. However, GARCH models are well adapted for forecasting the (conditional) *variance* of time series values. According to Eq. 33.2, the conditional expectation is in general the optimal forecast. We are therefore looking for the *conditional expectation of the conditional variance.*

33.2.1 Forecast Over Several Time Steps

The One-Step Forecast

Equation 32.16 shows that the conditional variance of X_T is equal to H_T if all X_t for $t \leq T - 1$ are known. Its conditional expectation is then the conditional expectation of H_T. Based on Eq. 32.17, the conditional expectation of H_T is simply H_T itself if the X values are known up to the time $T - 1$. Hence, the *optimal* one-step forecast for the conditional variance is[1]

$$\widehat{\mathrm{var}}_{T+1}^{\text{optimal}} = E[\mathrm{var}_{T+1} | X_T, \ldots, X_1]$$
$$= E[H_{T+1} | X_T, \ldots, X_1]$$

---

[1] In order to keep the notation as simple as possible, we will adopt the convention of denoting the conditional variance by $\mathrm{var}_{T+h} := \mathrm{var}[X_{T+h} | X_T, \ldots, X_1]$, likewise for its estimators $\widehat{\mathrm{var}}_{T+h} := \widehat{\mathrm{var}}[X_{T+h} | X_T, \ldots, X_1]$ for any $h > 0$. When these abbreviations for the conditional variances are used, it is always to be understood that the values $X_T, \ldots, X_1$ are known.

$$= H_{T+1}$$

$$= \alpha_0 + \sum_{j=1}^{q} \alpha_j X_{T+1-j}^2 + \sum_{i=1}^{p} \beta_i H_{T+1-i} . \qquad (33.5)$$

To clarify the argument used in the derivation of this result, each of the equalities in Eq. 33.5 will receive somewhat more scrutiny. The first equation is obtained from the general forecast equation, Eq. 33.2. The second follows from Eq. 32.16. The third holds as a result of Eq. 32.17 while the fourth equation is derived from Eq. 32.13 used in the construction of the GARCH process.

The Two-Step Forecast

The two-step forecast is somewhat more complicated. The optimal two-step forecast is, according to Eq. 33.2, the conditional expectation of var_{T+2} under the condition that $X_T, \ldots, X_1$ are known:

$$\widehat{\text{var}}_{T+2}^{\text{optimal}} = \text{E}[\text{var}_{T+2} | X_T, \ldots, X_1] .$$

Again, this is *not* equal to the conditional expectation of var_{T+2} under the condition that $X_{T+1}, \ldots, X_1$ are known. Equation 32.16 cannot be applied directly. Indeed, the *optimal* two-step forecast cannot be computed. We calculate instead, analogously to the linear forecast of the AR(p) process illustrated in Sect. 33.1, the best possible two-step forecast by replacing the expectation of var_{T+2} conditional upon $X_T, \ldots, X_1$ with the conditional expectation of var_{T+2} as if the $X_{T+1}, \ldots, X_1$ were all known:

$$\widehat{\text{var}}_{T+2} = \text{E}[\text{var}_{T+2} | X_{T+1}, \ldots, X_1] .$$

Now Eq. 32.16 can be applied to obtain

$$\widehat{\text{var}}_{T+2} = \text{E}[\text{var}_{T+2} | X_{T+1}, \ldots, X_1] = H_{T+2} .$$

Remember however, that X_{T+1} is *not* known and therefore H_{T+2} appearing here is not known at time T. The best we can do is to replace H_{T+2} by its optimal estimator which, according to Eq. 33.2, is given by its conditional

expectation

$$\widehat{\mathrm{var}}_{T+2} = \mathrm{E}[H_{T+2}|X_T, \ldots, X_1] \, . \tag{33.6}$$

Inside *this* expectation, we now replace H_{T+2} in accordance with the construction in 32.13:

$$\widehat{\mathrm{var}}_{T+2} = \mathrm{E}[H_{T+2}|X_T, \ldots, X_1]$$

$$= \alpha_0 + \sum_{j=1}^{q} \alpha_j \mathrm{E}[X_{T+2-j}^2|X_T, \ldots, X_1] + \sum_{i=1}^{p} \beta_i \mathrm{E}[H_{T+2-i}|X_T, \ldots, X_1]$$

$$= \alpha_0 + \alpha_1 \mathrm{E}[X_{T+1}^2|X_T, \ldots, X_1] + \sum_{j=2}^{q} \alpha_j X_{T+2-j}^2 + \sum_{i=1}^{p} \beta_i H_{T+2-i} \, . \tag{33.7}$$

In the last step, we have exploited the fact that all X_t are known for all times $t \leq T$ and, according to Eq. 32.17, all of the H_t for the times $t \leq T+1$. The expectation of the known quantities can be replaced by the quantities themselves. Only one unknown quantity remains, namely X_{T+1}^2. Because of Eq. 32.16, the conditional expectation of all of the X_{T+h} for all $h > 0$ is always equal to zero. The expectation of X_{T+h}^2 can therefore be replaced by the variance of X_{T+h}:

$$\mathrm{E}[X_{T+h}^2|X_T, \ldots, X_1] = \mathrm{E}[X_{T+h}^2|X_T, \ldots, X_1] - \underbrace{(\mathrm{E}[X_{T+h}|X_T, \ldots, X_1])^2}_{0}$$

$$= \mathrm{var}[X_{T+h}|X_T, \ldots, X_1] \quad \text{for every } h > 0 \, . \tag{33.8}$$

For $h = 1$ this implies

$$\mathrm{E}[X_{T+1}^2|X_T, \ldots, X_1] = \mathrm{var}[X_{T+1}|X_T, \ldots, X_1] = H_{T+1} \, ,$$

where Eq. 32.16 has again been used in the last step. The two-step forecast then becomes

$$\widehat{\mathrm{var}}_{T+2} = \alpha_0 + \alpha_1 H_{T+1} + \sum_{j=2}^{q} \alpha_j X_{T+2-j}^2 + \sum_{i=1}^{p} \beta_i H_{T+2-i} \, . \tag{33.9}$$

The Three-Step Forecast

For the two-step forecast, only the value of X_{T+1} in Eq. 33.7 was unknown, the necessary H values were known up to time $T + 1$. This it no longer the case in the three-step forecast. In this case, some of the H values are also unknown. Because of this additional difficulty, it is advisable to demonstrate the computation of a three-step forecast before generalizing to arbitrarily many steps.

The three-step forecast now proceeds analogous to the two-step forecast: the optimal forecast is, as indicated in Eq. 33.2, the conditional expectation of X_{T+3} under the condition that $X_T, \ldots, X_1$ are known.

$$\widehat{\mathrm{var}}_{T+3}^{\mathrm{optimal}} = \mathrm{E}[\mathrm{var}_{T+3} | X_T, \ldots, X_1] \, .$$

Again, Eq. 32.16 cannot be directly applied since the X are only known up to X_T and not up to X_{T+2}. The best possible three-step forecast is thus, analogous to Eq. 33.6

$$\widehat{\mathrm{var}}_{T+3} = \mathrm{E}[H_{T+3} | X_T, \ldots, X_1] \, . \tag{33.10}$$

In this expectation we now replace H_{T+3} with its expression constructed in Eq. 32.13 to obtain

$$\widehat{\mathrm{var}}_{T+3} = \mathrm{E}[H_{T+3} | X_T, \ldots, X_1]$$

$$= \alpha_0 + \sum_{i=1}^{p} \beta_i \, \mathrm{E}[H_{T+3-i} | X_T, \ldots, X_1] + \sum_{j=1}^{q} \alpha_j \, \mathrm{E}[X_{T+3-j}^2 | X_T, \ldots, X_1]$$

$$= \alpha_0 + \sum_{i=2}^{p} \beta_i \, H_{T+3-i} + \beta_1 \, \mathrm{E}[H_{T+2} | X_T, \ldots, X_1]$$

$$+ \alpha_1 \mathrm{E}[X_{T+2}^2 | X_T, \ldots, X_1] + \alpha_2 \mathrm{E}[X_{T+1}^2 | X_T, \ldots, X_1] + \sum_{j=3}^{q} \alpha_j X_{T+3-j}^2 \, .$$

$$\tag{33.11}$$

In the last step, the expectations of the known values were again replaced by the values themselves (all X_t for times $t \leq T$ and all H_t for times $t \leq T + 1$). Only three unknown values remain, namely X_{T+1}^2, X_{T+2}^2 and H_{T+2}. For the

In this equation, we have to replace three different types of conditional expectations:

- X^2_{T+h-j} is the squared process itself. Since the squared process has no influence on the best forecast, we have to use the latest known value X^2_{T+h-j} for all $j < h$. For $j \geq h$, we have no other choice than to use the latest estimator for the variance $\widehat{\text{var}}_{T+h-j}$.
- H_{T+h-i} is the unobservable volatility process. As long as it lies in the past, i.e., for $i \geq h$, we can use the value of H_{T+h-i}. If, however, $i < h$ holds, we have to replace H_{T+h-i} by the best estimator $\widehat{\text{var}}_{T+h-i}$.

Putting everything together, we get the following general recursive formula for the calculation of the forecast $\widehat{\text{var}}_{T+h}$ for the variance in a GARCH(p, q) model:

$$
\widehat{\text{var}}_{T+h} = \alpha_0 + \sum_{j=1}^{q} \alpha_j A_j + \sum_{i=1}^{p} \beta_i B_i ,
$$

where

$$
A_j = \begin{cases} X^2_{T+h-j} & \text{for } j \geq h \\ \widehat{\text{var}}_{T+h-j} & \text{for } j < h \end{cases}
$$

and

$$
B_i = \begin{cases} H_{T+h-i} & \text{for } i \geq h \\ \widehat{\text{var}}_{T+h-i} & \text{for } i < h . \end{cases}
$$

The recursion starts with the well-known GARCH(p, q) variance equation for $h = 1$

$$
\widehat{\text{var}}_{T+1} = H_{T+1} = \alpha_0 + \sum_{j=1}^{q} \alpha_j X^2_{T+1-j} + \sum_{i=1}^{p} \beta_i H_{T+1-i} .
$$

In the following step for $h = 2$, we have to differentiate between two cases. If q or p are equal to 1, then the corresponding sum over $\alpha_j X^2_{T+2-j}$

$$
= \sum_{j=1}^{\min(h-1,\,q)} \alpha_j E[X^2_{T+h-j}|X_T,\ldots,X_1] + \sum_{j=h}^{q} \alpha_j X^2_{T+h-j} + \alpha_0
$$

$$
+ \sum_{i=1}^{\min(h-2,\,p)} \beta_i E[H_{T+h-i}|X_T,\ldots,X_1] + \sum_{i=h-1}^{p} \beta_i H_{T+h-i} \, .
$$

In the last step, the expectation of the known values have again been replaced by the values themselves (all X_t for times $t \le T$ and all H_t for times $t \le T+1$). The remaining expectations of the Hs are replaced according to Eq. 33.13 through the respective variance estimators. We again use Eq. 33.8 for the conditional expectation of X^2 and write

$$
E[X^2_{T+1}|X_T,\ldots,X_1] = \mathrm{var}[X_{T+1}|X_T,\ldots,X_1] = H_{T+1}
$$

$$
E[X^2_{T+k}|X_T,\ldots,X_1] = \mathrm{var}[X_{T+k}|X_T,\ldots,X_1] = \widehat{\mathrm{var}}_{T+k} \quad \text{for } k > 1 \, .
$$

Substituting all of these relations for the conditional expectations finally yields the general h-step forecast of the conditional volatility in the GARCH(p, q) model:

$$
\widehat{\mathrm{var}}_{T+h} = \sum_{j=1}^{\min(h-2,\,q)} \alpha_j \widehat{\mathrm{var}}_{T+h-j} + \sum_{j=h}^{q} \alpha_j X^2_{T+h-j} + \alpha_{h-1} H_{T+1} + \alpha_0
$$

$$
\text{(33.14)}
$$

$$
+ \sum_{i=1}^{\min(h-2,\,p)} \beta_i \widehat{\mathrm{var}}_{T+h-i} + \sum_{i=h-1}^{p} \beta_i H_{T+h-i} \, .
$$

Together with the start value, Eq. 33.5, in the form of $\widehat{\mathrm{var}}_{T+1} = H_{T+1}$ the h-step forecast can be computed recursively for all h.

From Eqs. 33.13 and 33.2, the estimator for the variance is simultaneously the estimator for H, and thus

$$
\widehat{H}_{T+h} = \widehat{\mathrm{var}}_{T+h} \quad \text{for all } h > 0 \, .
$$

33.2.2 Forecast for the Total Variance

In the financial world, the time series $\{X_T\}$ is usually taken to represent a relative price change (yield), see for example Fig. 32.1 of the FTSE data. We

are therefore also interested in the variance of the return over an *entire* time period of length h, i.e., in the variance of the sum $\sum_{j=1}^{h} X_{T+j}$. In forecasts such as in Eq. 33.14, we are only dealing with a prediction of the conditional variance *after h* steps and not with a prediction of the variance *over* the entire term of h steps (from X_T to X_{T+h}). In other words, Eq. 33.14 is forecasting the conditional variance of X_{T+h} alone, and not predicting the variance of the sum $\sum_{j=1}^{h} X_{T+j}$. The variance of the total return $\sum_{j=1}^{h} X_{T+j}$ for independent (in particular uncorrelated) returns is simply the sum of the variances as can be seen in Eq. A.17. Since the X_t of the process in Eq. 32.13 are uncorrelated (because the ε_t are *iid*), the estimator for the total variance of the GARCH process over h steps is simply

$$\widehat{\text{var}}[\sum_{j=1}^{h} X_{T+j}|X_T, \dots, X_1] = \sum_{j=1}^{h} \widehat{\text{var}}[X_{T+j}|X_T, \dots, X_1] = \sum_{j=1}^{h} \widehat{\text{var}}_{T+j} .$$

(33.15)

Even in the case of weak autocorrelations between the returns in a given time series, this result holds in good approximation.

33.2.3 Volatility Term Structure

The variance of the total return over a term from T until $T + h$ is a function of this term. The square root of the (annualized) variance of the total return as a function of the term is called the *volatility term structure*. This plays an important role in pricing options since for an option with a lifetime of h, the volatility associated with this term is the relevant parameter value. From the estimator for the variance of the total return over the pertinent term, we obtain the estimator of the volatility structure as

$$\sigma(T, T + h) = \sqrt{\frac{1}{h}\widehat{\text{var}}\left[\sum_{i=1}^{h} X_{T+i}\Big|X_T, \dots, X_1\right]} = \sqrt{\frac{1}{h}\sum_{j=1}^{h} \widehat{\text{var}}_{T+j}} .$$

(33.16)

33.3 Volatility Forecasts with GARCH (1,1) Processes

For the GARCH(1, 1) process ($q = 1$, $p = 1$), all of the above estimators can be computed explicitly and the recursion equation 33.14 can be carried out. The start value of the recursion is simply

$$\widehat{\text{var}}_{T+1} = H_{T+1} = \alpha_0 + \alpha_1 X_T^2 + \beta_1 H_T \ . \tag{33.17}$$

as can be seen from Eq. 33.5. The two-step forecast as given by Eq. 33.9 simplifies to

$$\widehat{\text{var}}_{T+2} = \alpha_0 + \kappa H_{T+1}$$

where we defined the abbreviation

$$\kappa := \alpha_1 + \beta_1 \ .$$

For $h > 2$ and $p = q = 1$, the upper limits in the sums over the variance estimators in the general recursion equation 33.14 are simply

$$\min(h - 2, \ q) = \min(h - 2, \ p) = 1 \quad \text{for } h > 2 \ .$$

Neither of the other sums makes any contribution since the lower limit in these sums is greater than the upper limit. The term $\alpha_{h-1} H_{T+1}$ likewise does not exist since $q = 1$ implies that only α_0 and α_1 exist. However, $h - 1$ is greater than 1 for $h > 2$. All things considered, the h-step forecast in Eq. 33.14 reduces to

$$\widehat{\text{var}}_{T+h} = \alpha_0 + \kappa \ \widehat{\text{var}}_{T+h-1}$$

where $\kappa = \alpha_1 + \beta_1$. This recursion relation has a closed form expression in the form of a geometric series:

$$\begin{aligned}
\widehat{\text{var}}_{T+h} &= \alpha_0 + \kappa \widehat{\text{var}}_{T+h-1} \\
&= \alpha_0 + \kappa(\alpha_0 + \kappa \widehat{\text{var}}_{T+h-2}) = \alpha_0(1 + \kappa) + \kappa^2 \widehat{\text{var}}_{T+h-2} \\
&= \alpha_0(1 + \kappa) + \kappa^2(\alpha_0 + \kappa \widehat{\text{var}}_{T+h-3}) = (\alpha_0(1 + \kappa + \kappa^2) + \kappa^3 \widehat{\text{var}}_{T+h-3}) \\
&\cdots
\end{aligned}$$

$$= \alpha_0 \underbrace{\sum_{i=1}^{h-1} \kappa^{i-1}}_{\text{geometric series}} + \kappa^{h-1} \underbrace{\widehat{\text{var}}_{T+1}}_{H_{T+1}}$$

$$= \alpha_0 \left(\frac{1 - \kappa^{h-1}}{1 - \kappa} \right) + \kappa^{h-1} H_{T+1}$$

$$= \widetilde{\alpha}_0 + \kappa^{h-1} (H_{T+1} - \widetilde{\alpha}_0)$$

$$= \widetilde{\alpha}_0 + \kappa^{h-1} \left(\alpha_0 + \beta_1 H_T + \alpha_1 X_T^2 - \widetilde{\alpha}_0 \right), \qquad h > 1, \qquad (33.18)$$

where for $\widehat{\text{var}}_{T+1}$ the start-value of the recursion H_{T+1} was used and the geometric series was calculated according to Eq. 15.10. Here

$$\widetilde{\alpha}_0 := \frac{\alpha_0}{1 - \alpha_1 - \beta_1}$$

again denotes the *un*conditional variance from Eq. 32.18. The GARCH(1,1) prediction for the conditional variance after h steps is therefore equal to the unconditional variance plus the difference between the one-step forecast and the unconditional variance dampened by the factor κ^{h-1}. The stationarity condition requiring that $\alpha_1 + \beta_1 < 1$ implies that for $h \to \infty$ (a long prediction period) the GARCH prediction converges towards the unconditional variance.

The variance of the total return $\sum_{i=1}^{h} X_{T+i}$ over a term of length h as the sum of the conditional forecasts is obtained for the $textGARCH(1,1)$ process as indicated in Eq. 33.15:

$$\widehat{\text{var}}[\sum_{i=1}^{h} X_{T+i} | X_T, \dots, X_1] = \sum_{i=1}^{h} \widehat{\text{var}}_{T+i}$$

$$= \sum_{i=1}^{h} \left[\widetilde{\alpha}_0 + \kappa^{i-1} \left(\alpha_0 + \beta_1 H_T + \alpha_1 X_T^2 - \widetilde{\alpha}_0 \right) \right]$$

$$= h\widetilde{\alpha}_0 + \left(\frac{1 - \kappa^h}{1 - \kappa} \right) \left(\alpha_0 + \beta_1 H_T + \alpha_1 X_T^2 - \widetilde{\alpha}_0 \right),$$

$$(33.19)$$

where Eq. 15.10 for the geometric series is used again in the last step. The volatility term structure resulting from the 33.16 GARCH(1, 1) process is thus

$$\sigma(T, T + h) = \sqrt{\frac{1}{h}\widehat{\text{var}}[\sum_{i=1}^{h} X_{T+i}|X_T, \ldots, X_1]}$$

$$= \sqrt{\widetilde{\alpha}_0 + \frac{1}{h}\left(\frac{1 - \kappa^h}{1 - \kappa}\right)\left(\alpha_0 + \beta_1 H_T + \alpha_1 X_T^2 - \widetilde{\alpha}_0\right)} \ .$$

which approaches the unconditional variance $\widetilde{\alpha}_0$ like $1/\sqrt{h}$ for large h.

33.4 Volatility Forecasts with Moving Averages

In addition to the relatively modern GARCH models, older methods such as *moving averages* exist in the market, which, despite their obvious shortcomings, are still widely used, thanks to their simplicity. Before entering into a discussion of volatility forecasts via moving averages and comparing them with those of the GARCH models, we will first introduce the two most important varieties, the *simple* moving average, abbreviated here as *MA* and the *exponentially weighted moving average*, abbreviated as *EWMA*.

The (simple) moving average measures the conditional variance (of a time series with zero mean) simply as the sum of evenly weighted squared time series values over a time window of width b. The form for the MA corresponding to Eq. 32.16 is simply

$$\text{Var}[X_t|X_{t-1}, \ldots, X_1] = \frac{1}{b}\sum_{k=1}^{b} X_{t-k}^2 \ .$$

The well-known *phantom structures* arise from this equation because every swing in the X_t^2 is felt fully for b periods and then suddenly disappears completely when the term causing the perturbation no longer contributes to the average. An improvement would be to consider *weighted* sums where time series values further in the past are weighted less than values closer to the present. This can be realized, for example, by the exponentially weighted

moving average EWMA. The conditional variance in the EWMA is

$$\text{Var}[X_t|X_{t-1}, \ldots, X_1] = \frac{1}{\sum_{j=1}^{b} \lambda^{j-1}} \sum_{k=1}^{b} \lambda^{k-1} X_{t-k}^2$$

$$= \frac{X_{t-1}^2 + \lambda^1 X_{t-2}^2 + \cdots + \lambda^{b-1} X_{t-b}^2}{1 + \lambda^1 + \lambda^2 + \cdots + \lambda^{b-1}}.$$

For $\lambda < 1$ the values lying further back contribute less. The values commonly assigned to the parameter λ lie between 0.8 and 0.98. Naturally, the simple MA can be interpreted as a special case of the EWMA with $\lambda = 1$. The conditional variance of the EWMA is very similar to that of a GARCH(1,1) process since the recursion for H_T in Eq. 32.16 can be performed explicitly for $p = q = 1$ and the conditional variance of the GARCH(1,1) process becomes

$$\text{Var}[X_t|X_{t-1}, \ldots, X_1] = H_t$$

$$= \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 H_{t-1}$$

$$= \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \left[\alpha_0 + \beta_1 H_{t-2} + \alpha_1 X_{t-2}^2 \right]$$

$$= \alpha_0(1 + \beta_1) + \alpha_1(X_{t-1}^2 + \beta_1 X_{t-2}^2) + \beta_1^2 \left[\alpha_0 + \beta_1 H_{t-3} + \alpha_1 X_{t-3}^2 \right]$$

$$= \alpha_0(1 + \beta_1 + \beta_1^2) + \alpha_1(X_{T-1}^2 + \beta_1 X_{T-2}^2 + \beta_1^2 X_{T-3}^2) + \beta_1^3 H_{t-3}$$

$$\cdots$$

$$= \alpha_0 \sum_{k=1}^{b} \beta_1^{k-1} + \alpha_1 \sum_{k}^{b} \beta_1^{k-1} X_{T-k}^2 + \beta_1^b H_{T-b}.$$

If we now choose the parameters $\alpha_0 = 0$, $\beta_1 = \lambda$ and $\alpha_1 = \lambda(\sum_{j=1}^{b} \lambda^j)^{-1}$ then this conditional variance after b steps (apart from remainder term $\beta_1^b H_{T-b}$ which contains the influence of factors lying still further in the past) is exactly the same expression as for the EWMA. The difference between the GARCH(1, 1) and EWMA models first appears clearly in variance forecasts over more than one time step.

The conditional variances presented above can be interpreted as a one-step forecast for the conditional variance. The forecast over h steps delivers nothing new for the moving averages since both the MA and EWMA are static and fail to take the time structure into consideration. They start with the basic assumption that prices are lognormally distributed with a constant volatility.

This implies that

$$\widehat{\text{var}}[X_{T+h}|X_T, \ldots, X_1] = \widehat{\text{var}}[X_{T+1}|X_T, \ldots, X_1] = \text{var}[X_{T+1}|X_T, \ldots, X_1] \,.$$

holds for the h-step forecast of the MA as well as for the EWMA. As mentioned after Eq. 33.14, the conditional variance *after h* steps is being forecasted and *not* the variance of the total return *over* a term of h steps. The prediction for the variance of the total return as the sum over the conditional one-step forecasts is for moving averages (MA and EWMA) simply

$$\widehat{\text{var}}[\sum_{i=1}^{h} X_{T+i}|X_T, \ldots, X_1] = \sum_{i=1}^{h} \underbrace{\widehat{\text{var}}[X_{T+i}|X_T, \ldots, X_1]}_{\widehat{\text{var}}[X_{T+1}|X_T, \ldots, X_1]} = h \ \text{var}[X_{T+1}|X_T, \ldots, X_1] \,.$$

This is again the famous *square root law* for the growth of the standard deviation over time. The variance simply increases linearly over time and the standard deviation is therefore proportional to the square root of time. This leads to a static prediction of the volatility, and extrapolating, for example, daily to yearly volatilities in this way can easily result in an overestimation of the volatilities. The *volatility term structure* for the moving average is then, as expected, a constant:

$$\sigma(T, T+h) = \sqrt{\frac{1}{h}\widehat{\text{var}}[\sum_{i=1}^{h} X_{T+i}|X_T, \ldots, X_1]} = \sqrt{\text{var}[X_{T+1}|X_T, \ldots, X_1]} \,.$$

In the Excel workbook GARCH.xlsx , the one-step forecast of a GARCH$(1, 1)$ process, an MA with $b = 80$ and an EWMA with $b = 80$ and $\lambda = 0,95$ are presented. Furthermore, the ten-step forecast of the GARCH$(1, 1)$ process is shown. Since the time series we are dealing with is a simulated GARCH process, the "true" volatility is known (it is the H_t from the simulated series) and direct comparison can be made with each of the various estimates. As is clearly illustrated in GARCH.xlsx, the one-step GARCH$(1, 1)$ forecast (H_t with the parameters $\widehat{\alpha}_0$, $\widehat{\alpha}_1$ and $\widehat{\beta}_1$ fitted by *simulated annealing*) produces estimates which are quite close to the true volatility. The computation of the GARCH volatility term structure is presented in GARCH.xlsx as well.

34

Check for updates

Principal Component Analysis

34.1 The General Procedure

In addition to the autoregressive models described above, which are used for instance in the form of GARCH models when modeling volatility, a further technique of time series analysis, called *principal component analysis* (abbreviated as *PCA*), is widely applied in the financial world. This technique is employed in the analysis of term structure evolutions, for instance. A first approach to describe the *stochastic* dynamic of an interest rate term structure could be to define a risk factor for each vertex of the curve, i.e. the zero rate or the forward rate at that vertex. However, because of the large number of vertices, this approach would be quite calculation intensive. Instead, it is possible to reduce the number of stochastic variables to just a few (1 or 2, sometimes more) driving factors without loss of too much information. This approach has its justification in principal component analysis. Principal component analysis is a statistical technique which extracts the statistical components from the time series which are most relevant for the dynamics of the process in order of their importance. Applying this method to interest rates (i.e. the vertices of the interest rate curve) shows that often more than 90% of the term structure's dynamics can be ascribed to the one or two most important components.

Two other well-known statistical methods used in time series analysis are *factor analysis* [32] and *cointegration* [86]. Principal component analysis sets itself apart from these other methods by the ease with which the results

can be interpreted. An extensive description of principal component analysis can be found in [116], for example.

Since the time dependence of the data is not analyzed in PCA, the remarks made in Chap. 32 hold for this type of time series analysis as well: principal component analysis is a suitable approach for *stationary* time series only. Time series exhibiting non-stationary behavior like trends, etc. should be first freed of their non-stationary components by, for example, estimating trends and subsequently subtracting them from the time series or by taking differences of the data with respect to time, in other words, by performing a pre-treatment as described in Chap. 35. A reasonable principal component analysis of non-stationary time series is not to be achieved without first removing these potential non-stationary components.

In principal component analysis, the time series of not just one but *several* stochastic process are considered which may be strongly correlated. Principal component analysis is therefore an example of *multivariate* time series analysis. In financial applications, each process represents a risk factor. The situation is thus similar to that in Eq. 21.21 in Sect. 21.5 where n risk factors were considered as well. In the case of PCA, it is not a matter of the stochastic differential equations of the risk factors, but of their historic evolution. For each risk factor, there exists an associated time series of data for times t_k, $k = 1, \ldots T$. For example, the data of the relative changes (as in Eq. 31.1)

$$X_i(t_k) = \ln \left(\frac{S_i(t_k + \delta t)}{S_i(t_k)} \right) \approx \frac{S_i(t_k + \delta t) - S(t_k)}{S_i(t_k)}$$

$$i = 1, \ldots n, \quad k = 1, \ldots T \ .$$

These time series are assumed to be stationary, i.e., pre-treatment of the data has been performed already and the non-stationary components have been removed, if this has proved to be necessary. As a typical example of such a group of n risk factors, we can consider the relative changes of the interest rates at the vertices of an interest rate curve. The variances and pair wise correlations of these n risk factors can be arranged in a covariance matrix $\delta \Sigma$ as in Eq. 21.22. The entries of the covariance matrix can be determined from historical data as described in Sect. 31.1.

The dimension of this matrix is obviously equal to the number of risk factors (equivalently, the number of time series) and is thus equal to n. The central idea of principal component analysis is now to reduce the dimension of the problem on the basis of its statistical structure. The reduction is achieved by transforming the variables X_i into new uncorrelated variables Y_i appearing in

order of the size of their variance. This results, circumstances permitting, in a significant simplification of the subsequent analysis. It is in many cases possible to neglect statistical dependences of higher order in these new coordinates and to consider the time series of the new (transformed) variables as independent. As will be shown in the construction of the transformations, the new variables generally have different variances. In this case, it is possible to neglect those variables with relatively small variance.

This transformation of the random variables $X_1, \ldots, X_n$ into the new variables $Y_1, \ldots, Y_n$ is linear and thus we can write $Y_k = \sum_{i=1}^{n} \alpha_{ki} X_i$, or equivalently, in the vector notation used in Sect. 21.5

$$\mathbf{Y} = \boldsymbol{\alpha}\mathbf{X} \quad \text{mit} \quad \mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \tag{34.1}$$

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{\alpha}^1\right)^T \\ \vdots \\ \left(\boldsymbol{\alpha}^n\right)^T \end{pmatrix}.$$

In the last step, the rows of the transformation matrix have been written in terms of the vectors $\boldsymbol{\alpha}$ defined as follows:

$$\left(\boldsymbol{\alpha}^k\right)^T := \begin{pmatrix} \alpha_{k1} & \cdots & \alpha_{kn} \end{pmatrix} \implies \boldsymbol{\alpha}^k = \begin{pmatrix} \alpha_{k1} \\ \vdots \\ \alpha_{kn} \end{pmatrix}.$$

This implies that the components α_i^k of these vectors and the components α_{ki} of the matrix are related as follows:

$$\alpha_i^k = \alpha_{ki} = \left(\boldsymbol{\alpha}^k\right)^T_i \quad \forall\, k, i = 1, \ldots n .$$

The transformation Eq. 34.1 is thus given by

$$Y_k = \left(\boldsymbol{\alpha}^k\right)^T \mathbf{X} = \sum_{i=1}^{n} \alpha_{ki} X_i , \quad k = 1, \ldots n . \tag{34.2}$$

The coefficients of the matrix $\boldsymbol{\alpha}$ are now chosen such that the transformed variables Y_k possess the following properties:

We require every "transformation vector" $\boldsymbol{\alpha}_k$ (each row of the matrix $\boldsymbol{\alpha}$) to have a norm of 1, i.e.,

$$\left(\boldsymbol{\alpha}^k\right)^T \boldsymbol{\alpha}^k = \sum_{i=1}^{n} \alpha_i^k \alpha_i^k = \sum_{i=1}^{n} (\alpha_{ki})^2 \overset{!}{=} 1 \quad \forall k = 1, \ldots n \; . \tag{34.3}$$

We now select the components α_i^1 of $\boldsymbol{\alpha}^1$ in such a way that the variance of the first transformed variable $Y_1 = \left(\boldsymbol{\alpha}^1\right)^T \mathbf{X}$ is as large as possible (maximal) while satisfying Eq. 34.3. This is an optimization problem subject to the constraint $\left(\boldsymbol{\alpha}^1\right)^T \boldsymbol{\alpha}_1 = 1$. It will be shown explicitly in the material below how problems of this type are treated.

Having determined $\boldsymbol{\alpha}^1$ (and consequently Y_1), we proceed by determining the components α_i^2 of $\boldsymbol{\alpha}^2$ such that the variance of the second transformed variable $Y_2 = \left(\boldsymbol{\alpha}^2\right)^T \mathbf{X}$ is as large as possible (maximal), again subject to the condition that Eq. 34.3 holds. *In addition*, Y_2 must be uncorrelated with the vector Y_1 already determined. This additional condition can be expressed as

$$\mathrm{cov}(Y_2, Y_1) \overset{!}{=} 0 \; .$$

Using Eq. 34.2, we can write this covariance as

$$\mathrm{cov}(Y_2, Y_1) = \mathrm{cov}\left(\left(\boldsymbol{\alpha}^2\right)^T \mathbf{X}, \left(\boldsymbol{\alpha}^1\right)^T \mathbf{X}\right)$$

$$= \mathrm{cov}\left(\sum_{i=1}^{n} \alpha_i^2 X_i, \sum_{j=1}^{n} \alpha_j^1 X_j\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{2i} \alpha_{1j} \underbrace{\mathrm{cov}(X_i, X_j)}_{\delta \Sigma_{ij}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\boldsymbol{\alpha}^2\right)_i^T \delta \Sigma_{ij} \alpha_j^1$$

$$= \left(\boldsymbol{\alpha}^2\right)^T \delta \boldsymbol{\Sigma} \boldsymbol{\alpha}^1 \; .$$

This implies that, in addition to the constraint expressed in Eq. 34.3, $\boldsymbol{\alpha}^2$ must satisfy the condition

$$\left(\boldsymbol{\alpha}^2\right)^T \delta \boldsymbol{\Sigma} \boldsymbol{\alpha}^1 \overset{!}{=} 0 \tag{34.4}$$

Thus, the determination of $\boldsymbol{\alpha}^2$ (and thus of Y_2) involves an optimization problem with *two* constraints.

The remaining variables $Y_3, \ldots, Y_n$ are determined successively subject to analogous conditions: in the kth step, $\boldsymbol{\alpha}^k$ is chosen such that the variance of the kth transformed variable $Y_k = \left(\boldsymbol{\alpha}^k\right)^T \mathbf{X}$ is as large as possible (maximal) subject to the condition that Eq. 34.3 as well as the *additional $k-1$ conditions* hold, namely that Y_k be uncorrelated with *all* of the previously determined Y_i, $i = 1, \ldots k - 1$, i.e.,

$$\mathrm{cov}(Y_k, Y_i) \overset{!}{=} 0 \quad \text{for all} \quad i < k .$$

The formulation of these $k - 1$ conditions on $\boldsymbol{\alpha}^i$ is analogous to Eq. 34.4

$$\left(\boldsymbol{\alpha}^k\right)^T \delta \boldsymbol{\Sigma}\, \boldsymbol{\alpha}^i \overset{!}{=} 0 \quad \text{for all} \quad i = 1, \ldots k - 1 . \tag{34.5}$$

Overall, k conditions must be satisfied in the kth step. This decreases the maximal possible variance of Y_k from step to step, since with each step, the maximum is taken over a smaller class of vectors. It is thus not surprising that Y_1 has the greatest variance among the Y_k and that the variances of the Y_k decrease rapidly with increasing k. Therefore, the first Y_k contribute most to the total variance of all Y_k. These transformed variables Y_k are therefore referred to as the *principal components* and the vectors $\boldsymbol{\alpha}^k$ defined in the above construction as the *principal axes* of the system.

To provide the reader with a concrete example, we compute the first principal component Y_1 here. In order to do so, the variance of Y_1 conditional upon the satisfaction of Eq. 34.3 is maximized. According to Eq. A.12, the variance of Y_1 is

$$\mathrm{var}(Y_1) = \mathrm{cov}\left(\left(\boldsymbol{\alpha}^1\right)^T \mathbf{X}, \left(\boldsymbol{\alpha}^1\right)^T \mathbf{X}\right) = \mathrm{cov}\left(\sum_{i=1}^{n} \alpha_i^1 X_i, \sum_{j=1}^{n} \alpha_j^1 X_j\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{1i}\alpha_{1j}\underbrace{\mathrm{cov}(X_i, X_j)}_{\delta \Sigma_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\boldsymbol{\alpha}^1\right)_i^T \delta \Sigma_{ij}\, \alpha_j^1$$

$$= \left(\boldsymbol{\alpha}^1\right)^T \delta \boldsymbol{\Sigma}\, \boldsymbol{\alpha}^1 . \tag{34.6}$$

This is now maximized subject to the constraint $\left(\boldsymbol{\alpha}^1\right)^T \boldsymbol{\alpha}_1 = 1$. We can take constraints into account in an optimization problem using the method of *Lagrange multipliers*.[1] The method has already been discussed in Sect. 26.2.1. It requires the construction of the *Lagrange function* $\mathcal{L}$. This function is equal to the function to be maximized less a "zero" multiplied by a *Lagrange multiplier* λ. This zero is written in the form of the constraint to be satisfied, which in our case here is $0 = \left(\boldsymbol{\alpha}^1\right)^T \boldsymbol{\alpha}_1 - 1$. The Lagrange function is thus given by

$$\mathcal{L}(\lambda_1) = \underbrace{\left(\boldsymbol{\alpha}^1\right)^T \delta \boldsymbol{\Sigma} \, \boldsymbol{\alpha}^1}_{\text{zu maximieren}} - \lambda_1 \underbrace{\left[\left(\boldsymbol{\alpha}^1\right)^T \boldsymbol{\alpha}^1 - 1\right]}_{\text{Nebenbedingung}} \qquad (34.7)$$

$$= \sum_{i,j=1}^{n} \alpha_{1i} \delta \Sigma_{ij} \alpha_{1j} - \lambda_1 \sum_{i=1}^{n} (\alpha_{1i})^2 + \lambda_1 \ .$$

In order to find the optimal values α_{1i} subject to this constraint, we differentiate $\mathcal{L}$ with respect to the parameter α_{1i} and set the resulting expression equal to zero; in other words, we locate the maximum of the Lagrange function.[2]

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \alpha_{1i}} = 2 \sum_{j}^{n} \delta \Sigma_{ij} \alpha_{1j} - 2\lambda_1 \alpha_{1i} \quad \forall \, i = 1, \ldots n$$

from which it follows

$$\sum_{j}^{n} \delta \Sigma_{ij} \alpha_{1j} - \lambda_1 \alpha_{1i} = 0 \quad \forall \, i = 1, \ldots n$$

or in the compact matrix notation:

$$(\delta \boldsymbol{\Sigma} - \lambda_1 \mathbf{1}) \boldsymbol{\alpha}^1 = 0 \ , \qquad (34.8)$$

where, as in Eq. 21.37, $\mathbf{1}$ denotes the n-dimensional identity matrix. This condition implies, however, that the covariance matrix applied to the vector $\boldsymbol{\alpha}^1$

---

[1]An introduction in the technique of Lagrange multipliers for solving extreme value problems with boundary conditions can be found, for example, in [34].

[2]Since the Lagrange function differs only by *zero* from the value to be maximized (the variance), this equals the maximal value we are looking for. Of course, this is only true, if the difference is indeed equal to zero, i.e., if the boundary condition is fulfilled. This is the short form explanation of this method.

has no other effect than to multiply $\boldsymbol{\alpha}^1$ by a *number*: $\delta\boldsymbol{\Sigma}\boldsymbol{\alpha}^1 = \lambda_1\boldsymbol{\alpha}^1$. Equation 34.8 is therefore the *eigenvalue* equation of the covariance matrix (compare this to Eq. 22.22 in Sect. 22.3.2). The Lagrange multiplier λ_1 must in consequence be an eigenvalue of the covariance matrix and $\boldsymbol{\alpha}^1$ is the associated *eigenvector*. As discussed in detail in Sect. 22.3.2, the eigenvalue equation in 34.8 has a non-trivial solution $\boldsymbol{\alpha}^1 \neq \mathbf{0}$ if and only if the matrix $(\delta\boldsymbol{\Sigma} - \lambda_1\mathbf{1})$ is singular. For this to be the case, its determinant must be equal to zero

$$\det\left(\delta\boldsymbol{\Sigma} - \lambda_1\mathbf{1}\right) = 0 \ .$$

The eigenvalue λ_1 is the solution to this determinant equation. Having determined λ_1, it can be substituted into Eq. 34.8 to calculate the eigenvector $\boldsymbol{\alpha}^1$.

The eigenvalue λ_1 has another important intuitive interpretation which can be seen immediately if we multiply both sides of Eq. 34.8 on the left by $\left(\boldsymbol{\alpha}^1\right)^T$:

$$\lambda_1\boldsymbol{\alpha}^1 = \delta\boldsymbol{\Sigma}\boldsymbol{\alpha}^1$$
$$\left(\boldsymbol{\alpha}^1\right)^T \lambda_1\boldsymbol{\alpha}^1 = \left(\boldsymbol{\alpha}^1\right)^T \delta\boldsymbol{\Sigma}\boldsymbol{\alpha}^1$$
$$\lambda_1\underbrace{\left(\boldsymbol{\alpha}^1\right)^T \boldsymbol{\alpha}^1}_{1} = \underbrace{\left(\boldsymbol{\alpha}^1\right)^T \delta\boldsymbol{\Sigma}\boldsymbol{\alpha}^1}_{\text{Var}(Y_1)} \ .$$

Therefore, λ_1 is the variance of the principal component Y_1.

Analogously (i.e., maximization of the variances of the new variables Y_k), we proceed with the remaining principal components. As already mentioned above, one more constraint (see Eq. 34.5) must be taken into account with each further step. Analogous to Eq. 34.7, for each constraint a zero (in the form of the constraint multiplied by a Lagrange multipliers $\lambda_i, i = 1, \ldots k$) is subtracted from the variance of Y_k and the resulting Lagrange function $\mathcal{L}$ is maximized. In the computation of the kth principal component, k Lagrange multipliers appear in the expression for the Lagrange function. The complexity of the corresponding computation increases substantially with each new constraint. We thus refrain from explicitly performing the computation here. We wish to remark, however, on the following properties holding for all principal components as was shown for Y_1:

The Lagrange multipliers λ_k can be shown to be the eigenvalues of the covariance matrix $\delta\boldsymbol{\Sigma} = \text{cov}(X, X)$ appearing in decreasing order. The row vectors of the transformation matrix $\boldsymbol{\alpha}$, i.e., the $\boldsymbol{\alpha}^k$, are the associated eigenvectors.

This result paves the way for the implementation of the principal component analysis: we do not take the optimization problem as the starting point for our analysis, but rather we compute the eigenvalues and eigenvectors of the covariance matrix directly. In a second step, the eigenvalues λ_k and their associated eigenvectors $\boldsymbol{\alpha}^k$ are ordered according to the size of the eigenvalues, thereby constructing the transformation matrix from the ordered row vectors $\boldsymbol{\alpha}^k$ as defined in Eq. 34.1:

$$\boldsymbol{\alpha} = \begin{pmatrix} \left(\boldsymbol{\alpha}^1\right)^T \\ \vdots \\ \left(\boldsymbol{\alpha}^n\right) \end{pmatrix} .$$

Now, the original data X_i in the time series of the n risk factors are transformed into the new variables Y_i by applying the matrix $\boldsymbol{\alpha}$; this is done for each observation in the time series

$$Y_j(t_k) = \sum_{i=1}^{n} \alpha_{ji} X_i(t_k) \quad \text{for all} \quad j = 1, \ldots n \quad \text{and all} \quad k = 1, \ldots T .$$

Finally, those components Y_k of the transformed data are neglected if the associated eigenvalues (and thus the variances[3] of the new random variables Y_k) are small. It is often the case that the number of time series in the new variables Y_k necessary for further investigation can in this way be reduced to just two or three.

In many software packages, for example SAS, Matlab, or IDL, principal component analysis is provided as one of the few multivariate time series analysis procedures or at least modules exist allowing for its construction. The steps listed above are automated in these packages; only the original time series are needed as input. The covariance matrix is estimated from this data. The output is then obtained in the form of the eigenvalues of the covariance matrix, the associated eigenvectors, i.e., the transformation matrix, and the time series of the principal components. The reduction in dimension is achieved in that the data is reduced by $n - l - 1$ variables $Y_{l+1}, \ldots, Y_n$. Here, l is chosen conditional upon the relative size of the eigenvalues.

---

[3]The kth eigenvalue is in fact equal to the variance of the new random variable Y_k.

34.2 Principal Component Analysis of the German Term Structure

In this section, we conduct a principal component analysis on real interest rate data. The aim of our investigation is to determine the typical drivers of the interest rate term structure and the relative contribution of each individual driver to the total dynamics of the interest rate curve. The results of this section are relevant in several different respects. Firstly, the investigation introduced here will provide a plausible explanation for why it is for many problems sufficient to model the entire interest rate curve in very low dimensional spaces as is done in modern one, two, and three-factor term structure models although the term structure is clearly constructed with a larger number of vertices. Secondly, *stress scenarios* for typical and statistically independent movements in the interest rate curve can be identified.

The time series investigated are the yields for ten vertices with terms between one and ten years for the German term structure in monthly time steps spanning over a period of ten years. The data were subjected to a principal component analysis as described in the above section *without* having first been pre-treated by taking time-differences.[4] We find that the variances λ_i of the principal components Y_i, arranged in decreasing order, decrease *very* quickly. The proportion of the variances λ_1 and λ_2 of the first and second principal components with respect to the total variance $\sum_{i=1}^{10} \lambda_i$ is

$$\frac{\lambda_1}{\sum_{i=1}^{10} \lambda_i} \approx 96\% \,, \qquad \frac{\lambda_2}{\sum_{i=1}^{10} \lambda_i} \approx 3\% \,,$$

respectively. The proportion of the third and the fourth principal components is only

$$\frac{\lambda_3}{\sum_{i=1}^{10} \lambda_i} \approx 1\% \,, \qquad \frac{\lambda_4}{\sum_{i=1}^{10} \lambda_i} \ll 1\% \,,$$

---

[4]It has already been pointed out that principle component analysis assumes that the data in the time series are stationary. For the following investigation, this assumption is made keeping in mind that the results of the investigation should convey only a qualitative impression of term structure dynamics. Using the data directly (i.e. without a pre-treatment such as taking time differences, etc.) will simplify the *interpretation* of the results substantially.
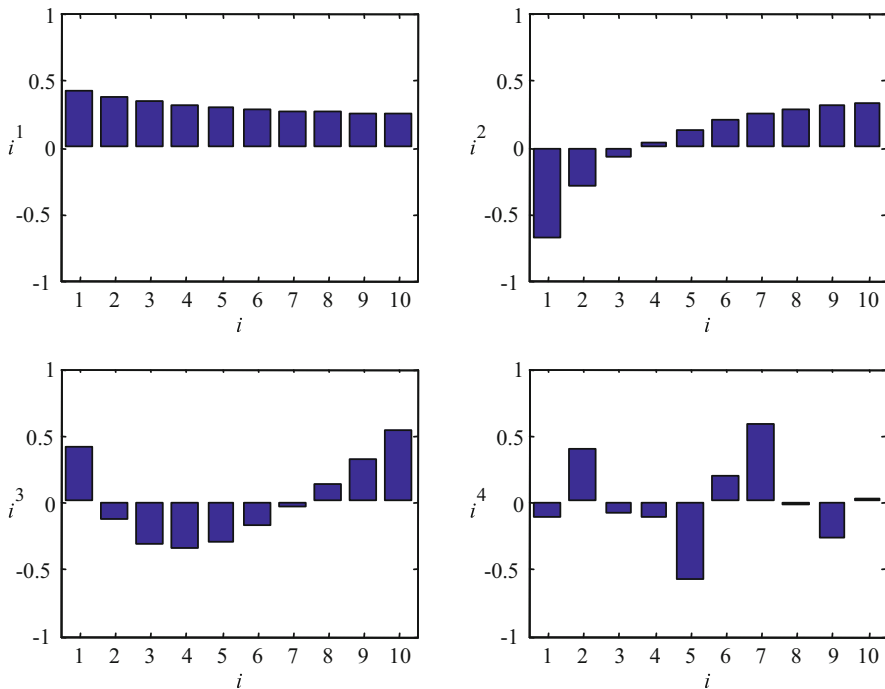
Fig. 34.1 The components of the first four eigenvectors $\alpha^1, \ldots, \alpha^4$. For each of these eigenvectors the components $\alpha_1^i, \ldots, \alpha_{10}^i$ are shown. The simple structure facilitates an intuitive interpretation of the associated principal components $Y_1, \ldots, Y_4$

respectively. The variances of the further principal components disappear almost completely, lying well below one tenth of one percent. Obviously, the dynamics of the term structure can be described by just a very *few* variables. Indeed, this result provides an excellent motivation for modeling the term structure in spaces of small dimension (or even in *one*-dimensional spaces as in Chap. 14). In Fig. 34.1, the first four of ten eigenvectors are presented. These eigenvectors can be interpreted quite easily.

The first eigenvector $\boldsymbol{\alpha}^1$ weights the interest rates of all terms approximately equally. The eigenvector normalization allows the first principal component Y_1 to be interpreted as the mean interest rate level. The fluctuations of this principal component Y_1 thus represent *parallel shifts of the entire term structure*.

The second eigenvector $\boldsymbol{\alpha}^2$ weights the short term interest rates negatively and the long term rates positively. Considering the inverse transformation of the principal components to the original interest rate vectors, we can draw conclusions as to the interpretation of the second principal component Y_2: adding the second component to the first has the effect of adjusting the mean

interest rate level by a mean slope. A change in this principal component Y_2 thus changes the mean *slope* of the term structure, or in other words, effects a *rotation of the term structure.*

The third eigenvector $\boldsymbol{\alpha}^3$ can be interpreted analogously. This vector weights the short and long-term rates positively, interest rates for the terms of intermediate length negatively. The addition of the associated principal component Y_3 to the first two thus effects a change in the mean *curvature of the term structure.* The fourth principal axis $\boldsymbol{\alpha}^4$ shows a periodic change in sign. With the associated principal component, periodic structures in the term structure can be represented such as those described in [159], for example.

Several practical conclusions for the analysis of scenarios commonly used in risk management can be drawn from the principal component analysis. Many of the common scenarios used to model a change in the term structure can be described in terms of the above decomposition. The most frequently used scenario is the *parallel shift,* which involves an increase or decrease in the entire term structure by a constant number of basis points. A further scenario, called the *twist*, involves a change in the slope of the term structure. This scenario is commonly realized through the addition (or subtraction) of, for example, m basis points to the interest rate corresponding to a term of m years; this is done for all terms in the term structure. Yet another scenario found in risk management is called *hump*. This scenario describes an increase in the short and long term rates and decrease in those for terms of intermediate length or vice versa. These three scenarios, the parallel shift, the twist and the hump, very often deduced on the basis of subjective experience, in fact correspond exactly to the first three principal components of the term structure. The scenarios mentioned here thus represent, from the statistical point of view, the most significant movements in the term structure. From the construction of the principal components, we can assume that these movements are approximately independent of one another and thus a simple representation of the dynamics of the term structure exists.

Finally, the analysis leads to the following results: first, practically all typical fluctuations in the term structure can be described by a combination of the above three scenarios. Second, making use of the eigenvalues, i.e., the variances of the principal components, confidence levels can be determined for the associated random variables. For example, information on the probability of the actual occurrence of one of these scenarios could be computed. We could proceed one step further: since the covariance matrix for the principal

components is diagonal, the value at risk can be differentiated with respect to movements in the individual principal components. The threat of VaR losses can be traced back to different types of movements in the term structure and interpreted accordingly. In addition, the values at risk from the three named stress scenarios can be taken to be uncorrelated and the value at risk can be computed by simply taking the sum of their squares.

Pre-Treatment of Time Series and Assessment of Models

35.1 Pre-Treatment of Time Series

The *pre-treatment* for the transformation of a given data set into a *stationary* time series has been mentioned several times in the preceding sections and will receive detailed treatment in this section. The basis for pre-treating a time series is its decomposition into a *trend* component g_t, a *seasonal* component s_t, and a random component[1] Z_t:

$$X_t = g_t + s_t + Z_t \ . \tag{35.1}$$

Z_t then represents the stationary time series with $\mathrm{E}[Z_t] = 0$. The *trend* g_t is a deterministic function of the time variable t, which represents a long-term development, for example a polynomial or an exponential function.[2] A weaker trend can sometimes be more readily recognized after a compression of the time axis. The *season* s_t represents a periodic component with a period p:

$$s_t = s_{t+p} \ . \tag{35.2}$$

Examples for such a component are the yearly cycle which overlays the inflation index or the weekly cycle overlaying stock price quotes. It follows

---

[1]Notation: capital letters here denote random variables, small letters denote deterministic functions.

[2]Or, for example, a so-called *logistic curve* $c_0/(1 + c_1 \exp(-c_2 t))$. Such curves address the possibility of saturation.

that the sum $\sum_{i=1}^{p} s_{t+i}$ of p successive values is a constant. This constant can be incorporated into the trend g_t so that, without loss of generality, the sum can be assumed to be equal to zero:

$$\sum_{i=1}^{p} s_{t+i} = 0 .$$

35.1.1 Differencing

If the trend and the seasonal component are not of interest, they can be eliminated by taking differences of certain values in the time series. Equation 35.2 implies that a seasonal component with period p can be eliminated by taking the pth difference Δ_p:

$$\begin{aligned}
\widetilde{X}_t := \Delta_p X_t &= X_t - X_{t-p} \\
&= g_t - g_{t-p} + Z_t - Z_{t-p} \\
&=: \widetilde{g}_t + \widetilde{Z}_t
\end{aligned}$$

with the new trend $\widetilde{g}_t = g_t - g_{t-p}$ and a new stationary time series $\widetilde{Z}_t = Z_t - Z_{t-p}$. Under the assumption that the remaining trend can be represented by a polynomial of degree k, this trend can be eliminated by taking first order differences in the above function k times. If, for example, the trend is a second degree polynomial, i.e., if

$$\widetilde{g}_t = a + bt + ct^2$$

then differencing twice eliminates this trend: taking a first order difference once yields

$$\Delta \widetilde{X}_t := \widetilde{X}_t - \widetilde{X}_{t-1} = b - c + 2ct + \widetilde{Z}_t - \widetilde{Z}_{t-1}$$

Differencing a second time results in

$$\Delta^2 \widetilde{X}_t = \Delta \widetilde{X}_t - \Delta \widetilde{X}_{t-1} = \widetilde{X}_t - 2\widetilde{X}_{t-1} + \widetilde{X}_{t-2} = 2c + \Delta^2 \widetilde{Z}_t$$

This time series now has neither a trend nor a seasonal component. After a subsequent elimination of the expectation $2c$, the new time series $\widetilde{\widetilde{Z}}_t := \Delta^2 \widetilde{X}_t - 2c$ is stationary and a time series model can be fitted.

In this way, the seasonal and trend components can be eliminated. The original data can, of course, be reconstructed through the inverse transformation. If we are interested in estimating the trend or the seasonal component themselves, we should begin by estimating the trend. This occurs either through one of the above mentioned parametric functions, (the parameters are then determined using a least squares estimator) or through moving averages as described in the next subsection. This estimated trend $\widehat{g}_t$ is then subtracted from the time series and the resulting series $X_t - \widehat{g}_t$ is used to obtain an estimator $\widehat{s}_t$ for the seasonal component. Because of the periodicity, the sum $\sum_{i=1}^{p} \widehat{s}_{t+i}$ is constant in time. This constant is then ascribed to the expectation of the time series so that $\sum_{i=1}^{p} \widehat{s}_{t+i} = 0$ holds. The difference $Z_t = X_t - \widehat{g}_t - \widehat{s}_t$ can then be viewed as a stationary time series and a time series model can be fitted.

35.1.2 Filters

One possibility of estimating the trend and eliminating the season is the application of *moving averages*. Starting from the time series $\{X_t\}$ a new process Y_t is constructed as

$$Y_t = \sum_{i=-m_1}^{m_2} q_i X_{t+i}$$

Frequently $m_1 = m_2$ (*symmetric average*) is chosen. Moving averages are an example of *filters.* Filters admit some information from the old time series into the new one while removing other information; the data sequence $\{X_t\}$ is filtered. Moving averages are an example of *linear* filters.

The weights q_i employed in the averaging are selected so that the trend is conserved, the seasonal component eliminated and the variance of the rest of the data sequence is minimized. An example taken from empirical economic research is the moving average

$$Y_t = \frac{1}{8}X_{t-2} + \frac{1}{4}X_{t-1} + \frac{1}{4}X_t + \frac{1}{4}X_{t+1} + \frac{1}{8}X_{t+2}$$

which allows a linear trend to "flow through", filters out a period-four seasonal component, (quarterly data) and reduces the variance; this can be seen in the following computation: let X_t be of the form in Eq. 35.1 which includes a linear trend with slope a (i.e., $g_{t+i} = g_t + a\,i$) and a seasonal component of

period four, i.e., $\sum_{i=1}^{4} s_{t+i} = 0$. Then

$$
Y_t = \sum_{i=-2}^{2} q_i X_{t+i} = \sum_{i=-2}^{2} q_i g_{t+i} + \sum_{i=-2}^{2} q_i s_{t+i} + \sum_{i=-2}^{2} q_i Z_{t+i} .
$$

The trend of Y_t remains the same as that of X_t:

$$
\sum_{i=-2}^{2} q_i g_{t+i} = \sum_{i=-2}^{2} q_i (g_t + ai)
$$

$$
= \frac{1}{8}(g_t - 2i) + \frac{1}{4}(g_t - i) + \frac{1}{4}g_t + \frac{1}{4}(g_t + i) + \frac{1}{8}(g_t + 2i)
$$

$$
= g_t - \frac{i}{4} - \frac{i}{4} + \frac{i}{4} + \frac{i}{4} = g_t .
$$

The seasonal component is eliminated:

$$
\sum_{i=-2}^{2} q_i s_{t+i} = \frac{1}{8}s_{t-2} + \frac{1}{4}s_{t-1} + \frac{1}{4}s_t + \frac{1}{4}s_{t+1} + \frac{1}{8}\underbrace{s_{t+2}}_{=s_{t-2}} \quad \text{since} \quad s_{t+4} = s_t
$$

$$
= \frac{1}{4}\underbrace{(s_{t-2} + s_{t-1} + s_t + s_{t+1})}_{=0} \quad \text{since} \quad \sum_{i=1}^{p} s_{t+i} = 0
$$

$$
= 0
$$

and the variance is (under the assumption that the stationary time series Z_t is weakly or not at all autocorrelated) substantially reduced:

$$
\text{Var}\left[\sum_{i=-2}^{2} q_i Z_{t+i}\right] = \sum_{i,j=-2}^{2} q_i q_j \, \text{cov}\left[Z_{t+i}, Z_{t+j}\right]
$$

$$
= \left(\frac{1}{64} + 3\frac{1}{16} + \frac{1}{64}\right) \text{Var}[Z_t] + \sum_{i \neq j} q_i q_j \underbrace{\text{cov}\left[Z_{t+i}, Z_{t+j}\right]}_{\approx 0}
$$

$$
\approx \frac{7}{32} \text{var}[Z_t] .
$$

A further example of a filter is the Spencer 15-point moving average with coefficients

$$\frac{1}{320}[-3, -6, -5, 3, 21, 46, 67, 74, 67, 46, 21, 3, -5, -6, -3] \, .$$

This filter even allows all polynomials up to and including polynomials of order three to flow through.

A complete filter can only be accomplished "within" the time series, namely for those data for which $t = m_1 + 1, \ldots, T - m_2$, where the length of the data sequence, i.e., the number of data points, is denoted by T. One possibility of applying the boundary points of the data sequence is to simply extend the moving average past the end of the known values, replacing the unknown values with the last value X_T (at the end of the sequence) and with the first value X_1 (at the beginning of the sequence). Alternatively, we could use smaller supports for the filters at the ends than in the middle.

35.1.3 Scaling

If the *variance* of the data sequence shows an increasing or decreasing behavior as a function of time, a variance stabilizing transformation, such as the *Box-Cox scaling* defined by

$$T_\lambda(X_t) = \frac{X_t^\lambda - 1}{\lambda}, \quad \text{for} \quad X_t \geq 0, \lambda > 0$$

$$T_0(X_t) = \ln(X_t), \quad \text{for} \quad X_t > 0, \lambda = 0$$

can be implemented. The second expression applies to the limiting case $\lambda \to 0$. If the data sequence contains negative values, we may add a positive constant to each of the data points. The parameter values $\lambda = 0$ or $\lambda = 1/2$ are most commonly used. Intuitively, the transformation implies that, depending on the choice of λ, small values can be magnified (or reduced) or vice versa.

As an example, we consider the transformation $T(X) = \ln(X)$. This transformation yields an approximately constant variance, when the standard deviation of the time series changes linearly with the expectation, i.e., when $\text{std}(X_t) = a\mu_t$. This can be shown as follows: the Taylor series expansion up to first order of the transformed time series about the expectation μ_t yields for

the standard deviation

$$\text{std}(T(X_t)) = \sqrt{\text{var}(T(X_t))} \approx \sqrt{\text{var}(T(\mu_t) + T'(\mu_t)(X_t - \mu_t))}$$
$$= T'(\mu_t) \, \text{std}(X_t) = T'(\mu_t) \, a\mu_t \, .$$

This expression is in fact constant when the transformation $T(X) = \ln(X)$ is chosen since $T'(X) = X^{-1}$.

35.2 Measuring the Goodness of Time Series Models

Having found a model for a stationary time series and fitted its parameters to the time series data, we immediately ask "how well does the model describe the time series?", in other words, how good is the fit? This question is answered by considering the deviations (also called the *residues)* of the values in the time series actually observed in comparison with those produced by the model. If the model is good, all of the information contained in the time series should also be contained in the fitted model; the residues should be of a purely random nature. For an AR(p) process as in Eq. 32.6, the residues are

$$\widehat{\varepsilon}_t = X_t - \sum_{i=1}^{p} \widehat{\phi}_i X_{t-i} \, . \tag{35.3}$$

If the fit is good, the residues should be according to N($0, \sigma^2$) normally distributed *iid* random variables with variance σ^2.

For a GARCH(p, q) process of the form 32.13, the situation is somewhat different since the random variables are *multiplicative* and not additive. We can, however, define the residues as the *quotient* of the observed data and the values obtained from the fitted time series model.

$$\widehat{\varepsilon}_t = X_t / \sqrt{\widehat{H}_t} \, . \tag{35.4}$$

Here, if the fit is good, the residues should be *iid* standard normal random variables.

In order to test whether the residues form a sequence of *iid* random variables $\varepsilon_t \sim \text{N}(0, \sigma^2)$, all statistical procedures for testing whether a sample of observations are the realizations of normally distributed *iid* random variables may be applied. Some of these tests will be introduced below. These tests are

applied to the GARCH(1, 1) process in the Excel workbook GARCH.xlsx from the download section [50].

35.2.1 Hypothesis Tests

Sign Test

We begin with a very simple test. For a sample of size T of *iid* random variables, the number V of points for which $\hat{\varepsilon}_t > \hat{\varepsilon}_{t-1}$ holds, has the following expectation and variance

$$E[V] = \frac{1}{2}(T - 1) \quad \text{and} \quad \text{var}[V] = \frac{1}{12}(T + 1) \,.$$

For large values of T, this number V is approximately $N(E[V], \text{var}[V])$ distributed. The quotient $\frac{V - E[V]}{\sqrt{\text{var}[V]}}$ is thus $N(0, 1)$ distributed. The hypothesis of a normal distribution is *rejected* with a confidence of $1 - \alpha$ (with α typically 5% or 10%) if

$$\frac{|V - E[V]|}{\sqrt{\text{var}[V]}} = \frac{|V - (T - 1)/2|}{\sqrt{(T + 1)/12}} > N^{-1}(1 - \alpha/2)$$

holds, where here $N^{-1}(1 - \alpha/2)$ denotes the percentile for the probability $1 - \alpha/2$ of the $N(0, 1)$ distribution.[3] Conversely, from the measurements of V, the confidence with which this test will reject the hypothesis can be computed. From the above equation, we obtain for α

$$N\left(\frac{|V - (T - 1)/2|}{\sqrt{(T + 1)/12}}\right) = 1 - \alpha/2 \,,$$

where N stands for the cumulative standard normal distribution. The confidence with which the hypothesis will be rejected thus becomes

$$\text{Confidence} = 1 - \alpha = 2N\left(\frac{|V - (T - 1)/2|}{\sqrt{(T + 1)/12}}\right) - 1 \,.$$

---

[3]The percentile associated with $1 - \alpha/2$ is sought since the test under consideration is two-sided (we consider the absolute value of the normalized random variable). In total, however, the confidence level of the test is $1 - \alpha$ since $P\left(\frac{|V - E[V]|}{\sqrt{\text{var}[V]}} \leq N^{-1}(1 - \alpha/2)\right) = 1 - \alpha$.

However, this test does not reject the hypothesis of an independent and identically distributed random sample in the case of cyclical data because approximately half of the residues are then increasing. In practice, this does not present a problem since the periodicity can be easily spotted by considering a plot of the residues.

Inflection Points Test

This test has a similar structure as in the sign test. For *iid* random variables, the number of inflection points U for which $\hat{\varepsilon}_{t-1} < \hat{\varepsilon}_t$ and $\hat{\varepsilon}_t > \hat{\varepsilon}_{t+1}$ or else $\hat{\varepsilon}_{t-1} > \hat{\varepsilon}_t$ and $\hat{\varepsilon}_t < \hat{\varepsilon}_{t+1}$ holds, must have the following expectation and variance

$$E[U] = \frac{2}{3}(T-2) \quad \text{and} \quad \text{var}[U] = \frac{16T - 29}{90} .$$

For large T, this number U is approximately $N(E[U], \text{var}[U])$ distributed. Thus, analogous to the sign test, the hypothesis of *iid* random variables will be *rejected* with a confidence $1 - \alpha$ if

$$\frac{|U - E[U]|}{\sqrt{\text{var}[U]}} = \frac{|U - 2(T-2)/3|}{\sqrt{\text{var}[(16T - 29)/90]}} > N^{-1}(1 - \alpha/2) .$$

Analogous to the sign test, the confidence with which the hypothesis will be rejected can be calculated from U as

$$\text{Confidence} = 1 - \alpha = 2N\left(\frac{|U - 2(T-2)/3|}{\sqrt{\text{var}[(16T - 29)/90]}}\right) - 1 .$$

Percentile Test and Kuiper Statistic

We start by selecting a two-sided confidence level α. Two-sided means that $\alpha/2$ (percent) of the residues may fall below the lower boundary of the confidence interval and $\alpha/2$ above the upper boundary. The confidence interval itself is determined from the conditional variance σ^2 of the residues $\hat{\varepsilon}_t$ (from Eq. 35.4 $\sigma^2 = 1$ for GARCH processes, for example) under the assumption that residues are normally distributed, i.e., $\hat{\varepsilon}_t/\sigma$ are assumed to have a standard normal distribution. We now count the residues $\hat{\varepsilon}_t$ lying outside this confidence interval. For these residues we either have $N(\hat{\varepsilon}_t/\sigma) > 1 - \alpha/2$ or

$N(\hat{\varepsilon}_t/\sigma) < \alpha/2$. If the goodness of the fit is high, then the residues should indeed be $N(0, \sigma^2)$ distributed and therefore the proportion (relative to the total number of residues) of the residues lying outside the confidence interval should be approximately equal to the confidence α. This should hold for every arbitrary α between 0 and 1.

We therefore consider the quotients

$$F(\alpha) = \frac{\text{number of residues where } N(\hat{\varepsilon}_t/\sigma) > 1 - \alpha/2 \text{ or } N(\hat{\varepsilon}_t/\sigma) < \alpha/2}{\text{number of all residues}} \quad \text{for} \quad 0 \le \alpha \le 1 .$$

(35.5)

For a good fit, this should be close to α, i.e., the variable $F(\alpha) - \alpha$ should be close to zero for all $0 \le \alpha \le 1$. The deviations $F(\alpha) - \alpha$ should be uniformly distributed over the entire spectrum of α (this implies that the assumption of a normal distribution is correct) and they should be independent (this implies in particular no autocorrelations). In risk management, the performance at the "tails" of the distribution, i.e., for $0 \le \alpha \le 0.05$ and $0.95 \le \alpha \le 1.0$, is of particular importance.

The *Kuiper statistic* defined as

$$K(F(\alpha), \alpha) = \max_{0 \le \alpha \le 1} \{F(\alpha) - \alpha\} + \max_{0 \le \alpha \le 1} \{\alpha - F(\alpha)\}$$

(35.6)

adds the largest deviation of $F(\alpha)$ from α from above and from below. The Kuiper statistic selects the "worst" deviations over the entire range of α and aggregates them into a single number which can then be compared with the corresponding number of other models for the purpose of evaluating the goodness of each of the models.

Estimation of the Autocorrelation Function

The estimators $\hat{\varrho}(h)$ for the autocorrelations defined in Eq. 31.13 are approximately normally distributed for an independent identically distributed sequence of T random variables with finite variance. For large values of T, this distribution is approximately $N(0, 1/T)$. Thus, for approximately $1 - \alpha$ percent of the autocorrelations between the residues,

$$-\frac{N^{-1}(1 - \alpha/2)}{\sqrt{T}} < \hat{\varrho}(h) < \frac{N^{-1}(1 - \alpha/2)}{\sqrt{T}}$$

should hold if the fit is good. Here again, N denotes the cumulative standard normal distribution. More "outliers" would indicate that the assumption of independent, identically distributed random variables is violated by the residues.

QQ Plot

In a *QQ plot*, also referred to as a *percentile plot* or a *quantile plot*, the percentiles of two distributions are plotted against one another. We will denote the two cumulative distribution in question by Φ_1 and Φ_2. For the QQ plot, we consider the question: given the probability $P = \Phi_1(X)$, which argument Y must be substituted into the distribution Φ_2 in order to generate the same probability P. In other words: which Y has to be chosen so that $\Phi_2(Y) = \Phi_1(X)$. Both of the arguments are then the percentiles of the two distributions Φ_1 and Φ_2 for the same probability value P. These two percentiles are plotted against one another. We plot Y against X explicitly by defining the function $Y = f(X)$ as

$$f(X) = \Phi_2^{-1}(\Phi_1(X)) \ .$$

If, for example, Φ_1 and Φ_2 are two *normal* distributions, the QQ plot is also referred to as a *normal plot*. For two normal distribution, i.e., for $\Phi_1 = N(\mu_1, \sigma_1^2)$ and $\Phi_2 = N(\mu_2, \sigma_2^2)$, the condition for the percentiles X and Y is

$$\frac{1}{\sqrt{2\pi\sigma_1^2}} \int_{-\infty}^{X} e^{-\frac{(u-\mu_1)^2}{2\sigma_1^2}} \, du = \Phi_1(X) = \Phi_2(Y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \int_{-\infty}^{Y} e^{-\frac{(u-\mu_2)^2}{2\sigma_2^2}} \, du \ .$$

With the substitution $w = \frac{u-\mu_i}{\sigma_i} \implies du = \sigma_i dw$ on both sides, i.e., for $i = 1, 2$, this equation becomes

$$\int_{-\infty}^{\frac{X-\mu_1}{\sigma_1}} e^{-w^2/2} \, dw = \int_{-\infty}^{\frac{Y-\mu_2}{\sigma_2}} e^{-w^2/2} \, dw$$

These two integrals differ only in their upper limit. If they are to be equal, it suffices for their upper limits to be equal, yielding the function for the

percentile plot

$$\frac{X - \mu_1}{\sigma_1} = \frac{Y - \mu_2}{\sigma_2} \Rightarrow$$

$$f(X) = Y = \frac{\sigma_2}{\sigma_1}(X - \mu_1) + \mu_2 . \tag{35.7}$$

Thus, if both random variables have a normal distribution, the QQ plot is simply a straight line whose slope is given by the ratio of the standard deviations, $\frac{\sigma_2}{\sigma_1}$, and whose offset, $\mu_2 - \frac{\sigma_2}{\sigma_1}\mu_1$, is determined by the expectation of the distributions as well as the ratio of the standard deviations. Usually, one of the distributions, Φ_2 say, is obtained from empirical sample values (*empirical probability distribution*) and it is tested whether this distribution agrees with another, theoretical distribution Φ_1. To do this, the T sample values $\hat{\varepsilon}_i$ are arranged in increasing order resulting in the sequence $\hat{\varepsilon}_{(1)}, \hat{\varepsilon}_{(2)}, \ldots, \hat{\varepsilon}_{(T)}$ with $\hat{\varepsilon}_{(1)} < \hat{\varepsilon}_{(2)} < \ldots < \hat{\varepsilon}_{(T)}$. To distinguish the ordered from the original sample sequence, the indices of the *ordered* sequence appear in parentheses. The empirical distribution is obtained from the following deliberations: the proportion of all sample values which are less than or equal to $\hat{\varepsilon}_{(i)}$ is i/T. For large T, this ratio is a good approximation of the cumulative empirical distribution, i.e., i/T approximates the *empirical probability* that a random number is less than or equal to $\hat{\varepsilon}_{(i)}$

$$\Phi_2(\hat{\varepsilon}_{(i)}) = P_i \approx i/T ,$$

This implies that the percentile of the empirical distribution associated with the probability $P_i \approx i/T$ is

$$Y_i = \Phi_2^{-1}(P_i) = \hat{\varepsilon}_{(i)} .$$

The associated percentile of the *theoretical* distribution Φ_1 is then[4]

$$X_i = \Phi_1^{-1}(P_i) \approx \Phi_1^{-1}(i/T) \quad \text{for all} \quad i < T . \tag{35.8}$$

---

[4]Note that as long as we have only finitely many observations, i.e. for $T < \infty$, the *empirical* probability Φ_2 for any $\hat{\varepsilon}_i$ being $\leq \hat{\varepsilon}_{(T)}$ is of course 1 since $\hat{\varepsilon}_{(T)}$ is *by definition* the largest value ever observed. The *theoretical* cumulative probability $\Phi_1(x)$, on the other hand, approaches 1 only for x approaching the upper limit of the range of the possible arguments of Φ_1, usually for $x \to \infty$. Or in the language of Eq. 35.8:

$$\text{``}X_T \approx \Phi_1^{-1}(T/T) = \Phi_1^{-1}(1) = \infty\text{''}$$

Therefore we have to exclude the largest value $\hat{\varepsilon}_{(T)}$ from the QQ plot when testing an empirical distribution.

The QQ plot is now obtained by plotting the ordered pairs (X_i, Y_i), i.e., by graphing $\hat{\varepsilon}_{(i)}$ against $\Phi_1^{-1}(i/T)$ for all $i < T$. In this way it can be checked, for example, whether the sample values are normally distributed by testing how well a QQ plot of the sample values against those obtained from a normal distribution delivers a straight line.[5] It is convenient to chose the *standard normal distribution*, $\Phi_1 = N(0, 1)$, as the theoretical distribution in this case. Equation 35.7 then reduces to $Y_i = \mu_2 + \sigma_2 X_i$ for $i = 1, \ldots, T - 1$ and the standard deviation and the expectation of the empirical distribution can be read off directly as the slope and offset of the QQ plot.

A normal plot such as this reveals much more than a common two-sample test since the entire distribution can be tested and deviations from the normal distribution such as non-linearities, skewness, or dissimilar distribution boundaries can be identified. In addition to this visual examination, the *correlation coefficient* between $Y_i = \hat{\varepsilon}_{(i)}$ and $X_i = \Phi_1^{-1}(P_i)$

$$R^2 = \frac{\left(\sum_{i=1}^{T}(Y_i - \overline{Y}) X_i\right)^2}{\sum_{i=1}^{T}(Y_i - \overline{Y})^2 \sum_{i=1}^{T} X_i^2} \quad \text{with} \quad \overline{Y} = \sum_{i=1}^{T} Y_i / T$$

serves as an analytic variable for testing the linearity.[6] For a straight line, $R^2 = 1$ must hold.[7]

35.2.2 Goodness of Fit vs. Goodness of Forecast

All of the above criteria can be applied to not only test the goodness of a *fit* but also the goodness of a *forecast*. The forecast must occur *out of sample*, i.e., a portion of the given data set is used to obtain an estimator (i.e., for fitting the model), while the remaining data is predicted using the fitted model. The goodness is then evaluated on the basis of the "accuracy" of the prediction. This "accuracy" must be quantified for each respective problem.

---

[5]It can be shown that using $P_i = \frac{i-3/8}{T-3/4+1}$ for the probabilities in a normal plot is slightly better than using $P_i = i/T$. This, however, is only a marginal improvement over the much more intuitive value $P_i = i/T$.

[6]Here, the correlation coefficient for the special case that the X_i have mean zero, $\overline{X} = 0$, is presented, which is the case for a QQ plot against the standard normal distribution. In general, the correlation coefficient also has the form given in the equation, but with X_i replaced by $(X_i - \overline{X})$.

[7]For a formal test, we need confidence intervals under the hypothesis of normally distributed residuals. For example, for $T = 200$, $P(R^2 < 0.987) = 0.05$ (5% interval) and $P(R^2 < 0.989) = 0.1$ (10% interval).

Testing the goodness of a forecast by using the entire sample to fit the model and then "predicting" values *in* the sample is obviously optimistically biased and does not yield a realistic assessment of the actual forecasting ability of the model. Time series models are often constructed for the primary purpose of making forecasts. Thus, the goodness of a model should be judged according to its performance in making predictions. In the Excel workbook GARCH.XLSX, the GARCH(1, 1) model was fitted to the first 400 time series values via maximum likelihood estimation using simulated annealing. The goodness tests presented in GARCH.XLS refer to the data points 401 to 1001 and are therefore *out of sample forecasting tests.*

However, there are by all means tests available which only make sense *within* that part of the sample which was used for fitting the model (*in sample* tests). In such cases, we are only interested in the goodness of fit achieved by the model, not in its forecasting power. If, for example, the best model is to be identified from among a whole class of different models it is of course natural to ask which model we prefer. The goodness of fit, measured in terms of the smallest possible residues or maximum possible likelihood, can not be the only deciding factor because a model with many parameters might be well fitted but can also be easily overfitted. Moreover, models with few parameters are generally preferable as they are easier to interpret.

The *Akaike information criterion* is a compromise between larger likelihood and fewer parameters. It is given by

$$\text{AIC} = -2\mathcal{L} + \frac{2(Q+1)T}{T-Q-2}, \qquad (35.9)$$

where Q denotes the number of parameters of the model to be evaluated and $\mathcal{L}$ the log-likelihood of the fitted process. For the GARCH(p, q) process the order is $Q = p+q+1$, for an AR(p)-process we have $Q = p$. The model with the smallest AIC is preferable. The Akaike Criterion represents a compromise between a large likelihood and a small number of parameters; models with many parameters are penalized.

35.2.3 Examples: Goodness of GARCH Models

The comparison of a GARCH(1, 1) process with moving averages by means of measures of goodness is also explicitly demonstrated in the Excel workbook GARCH.XLSX. In addition, we will demonstrate these methods using the above mentioned FTSE data set presented in Fig. 32.1. The first 2,000 values of the
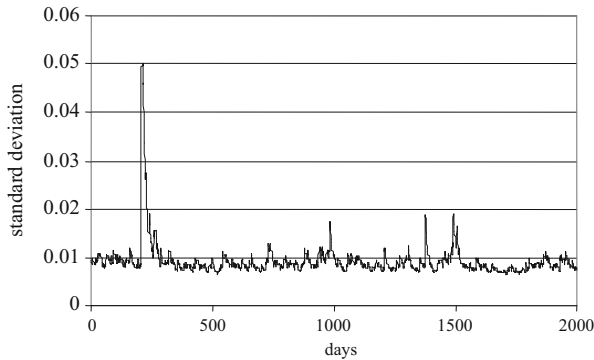
Fig. 35.1 Conditional standard deviation of the GARCH$(1, 1)$ process fitted to the FTSE time series. The crash in October 1987 is clear to see

data set are used in the parameter estimation. Simulated annealing algorithms were employed to obtain the optimized parameter values $\alpha_0 = 0.000$, $\alpha_1 = 0.089$ and $\beta_1 = 0.849$. Figure 35.1 shows the conditional standard deviations $h_t = \alpha_0 + \beta_1 h_{t-1} + \alpha_1 X_{t-1}^2$ of the GARCH process, where the corresponding values for X_{t-1} were taken from the FTSE data set. The Akaike criterion for the goodness-of-fit of the GARCH process is computed as -6519.87. The last 934 values serve to test the one-step forecast for the variance performed by employing Eq. 33.17. Here, we use a procedure (which has not been yet been discussed in previous sections) to directly test the estimated conditional variance. We proceed as follows: beginning with a value X_{t-1} of the FTSE time series (for example with $t - 1 = 2500$), we calculate $h_t = \alpha_0 + \beta_1 h_{t-1} + \alpha_1 X_{t-1}^2$ for the next day ($t = 2501$) with the parameters estimated using the first 2,000 values. This h_t is then used to compute the confidence intervals of the normal distribution $N(0, h_t^2)$ corresponding to a confidence level of 99% and 95%, respectively, to establish whether the next value $X_t = X_{2501}$ of the FTSE time series lies within this confidence interval. Subsequently, the next h_{2502} is computed from h_{2501} and the value X_{2501} from the FTSE time series data. The corresponding confidence interval is then calculated and it is established whether the actual value X_{2502} lies within this confidence interval, and so on for all 934 out-of-sample values. The number of outliers is counted and compared with the expected number of outliers. For a test of the ten day forecast, Eq. 33.18 is used to predict the conditional variance for the daily yield of the FTSE ten days into the future. For example, the conditional variance h_{2510} is estimated using the value X_{2500} from the

Table 35.1 Confidence level $\alpha = 99\%$

1-day forecast	10-day forecast
GARCH: 8 (9)	GARCH: 77 (67)
MA: 21 (9)	MA: 145 (67)
EWMA: 18 (9)	EWMA: 117 (67)

Table 35.2 Confidence level $\alpha = 95\%$

1-day forecast	10-day forecast
GARCH: 21 (46)	GARCH: 158 (337)
MA: 69 (46)	MA: 517 (337)
EWMA: 45 (46)	EWMA: 391 (337)

FTSE time series and the associated value h_{2500}. Using h_{2510}, the confidence interval for the 99% and 95% confidence levels of the normal distribution $N(0, h_{2510}^2)$ are computed and it is subsequently established whether the FTSE time series value X_{2510} lies within this confidence interval, and so on for all out-of-sample values. Again, the number of outliers is counted and compared with the expected number of outliers.[8] The results of these tests are presented in Tables 35.1 and 35.2.

The same tests were also performed for moving averages. The time window consists of 250 days (one year). The number of time series values lying outside of the confidence interval (outliers) is given for the uniformly weighted moving average (MA) and for the exponentially weighted moving average (EWMA). The weighting parameter for the EWMA is $\lambda = 0.94$. The values in parentheses indicate the theoretically expected number of outliers. For the 1-day predictions, the GARCH forecast is always superior. For the 10-day forecast, we observe mixed results: at 5% confidence, the GARCH process underestimates the theoretical number of outliers by almost one half, at 1% confidence, the number of outliers is very close to the theoretical value. The converse holds for the EWMA. The MA shows the worst performance. Fitting a GARCH(1, 2) process leads to almost identical results.

---

[8]The forecast of the variance of the *total* return (for example, the 10-day total return) can be tested analogously. The estimator $\text{var}[\sum_{i=1}^{10} X_{2500+i}|X_{2500}, \ldots, X_1]$ is computed according to 33.19. The result is used to determine the confidence interval and then it is checked whether $\sum_{i=1}^{10} X_{2500+i}$ lies within this confidence interval. The X_{2500+i} are taken from the FTSE time series. Again, the number of outliers is compared with the expected number of outliers.
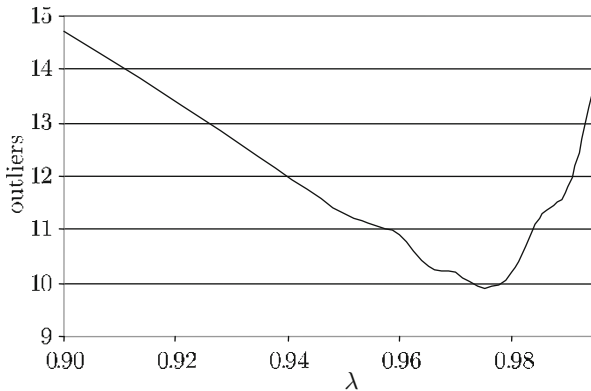
Fig. 35.2 Number of outliers outside a 99% confidence interval as a function of λ for a 10 day EWMA volatility forecast. The theoretical value is 67 outliers

Of all the investigated time series, the GARCH process is superior to all others in making a 1-day forecast. For the 10-day forecasts, the evaluation is not as uniform. For a similar fit to the S&P500 index over the same time span as for the FTSE, the 10 day prediction is better at 5% confidence than at 1%; the opposite to what we observed for the FTSE. For longer prediction periods, it is therefore important to conduct a separate analysis for each time series and repeat this regularly at certain times in order to asses the goodness of the methods and identify changes as early as possible.

Figure 35.2 shows the number of outliers for the EWMA as a function of λ. In this way, it can be established which λ is the best, based on the goodness criterion of interest.

A further variant of analysis is the Kuiper statistic K. This statistic will be illustrated using data taken from the S&P500 index over the same time window as for the FTSE index in the above example (see Figs. 35.3 and 35.4). Again, a GARCH(1, 1) process was fitted to the first 2,000 time series values of the S&P500 returns, the last 934 were reserved for out-of-sample testing of the forecasting. For MA and EWMA, the time window is again 250 days. The Kuiper statistic K is least favorable for the GARCH process. However, in the areas so important in risk management, namely to the far left and far right of the distribution corresponding to small and large percentiles, respectively, the GARCH process shows the best performance. The best results for the 1-day forecasts are those from the GARCH model. The deviation in the middle of the distribution indicates that the assumed conditional normal distribution
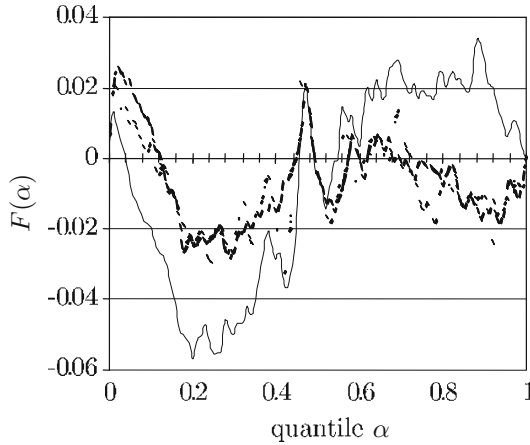
Fig. 35.3 $F(\alpha)$ as defined in Eq. 35.5 for GARCH (solid line), MA (dotted line) and EWMA (dash-dotted line). The Kuiper statistics, Eq. 35.6, are K(GARCH)=0.093, K(MA)=0.058, K(EWMA)=0.055 for $\lambda = 0.8$
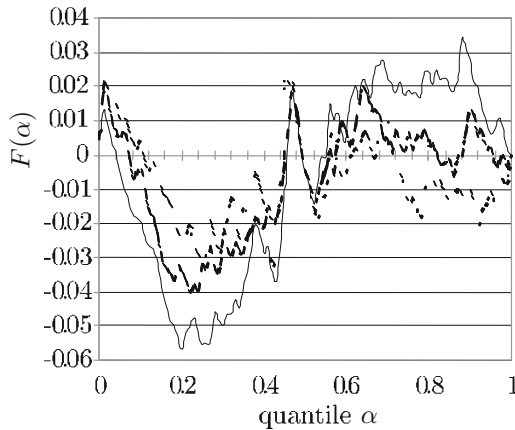
Fig. 35.4 $F(\alpha)$ as in Fig. 35.3. The only thing changed is that λ is now 0.94 for the EWMA. The Kuiper statistic for this new EWMA is K(EWMA)=0.064 for $\lambda = 0.94$

does not properly reflect the skewness in the actual distribution of the returns. For $\lambda = 0.8$, the value of K for the EWMA is the smallest but we observe a "reverse" in the performance of the EWMA at the tails of the distribution: here its performance is poorest. For $\lambda = 0.94$, the value of K is higher but the fit at the tails is better, for example between 0 and 0.05, it is better than MA.

Finally, Fig. 35.5 shows daily standard deviations for the S&P500 data set after the 400th day in the out-of-sample region (i.e., the region not used for
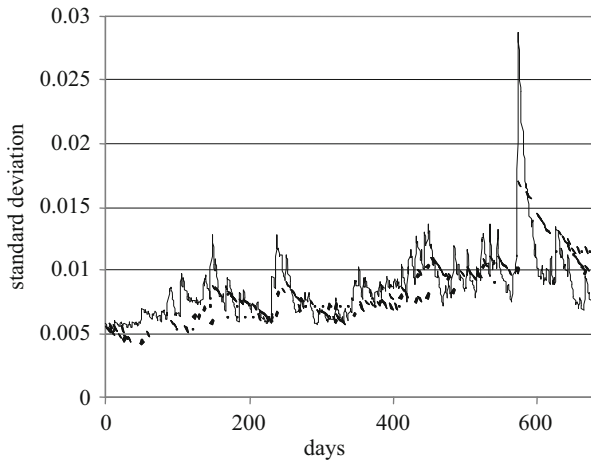
Fig. 35.5 Daily standard deviations of the S&P500 returns from GARCH (solid line), MA (dotted line) and EWMA (dash-dotted line). The window for the moving averages spans 250 time series values. The EWMA damping factor is $\lambda = 0.975$

fitting the parameters of the GARCH model). The increasing smoothness in the sequence GARCH, EWMA and MA is clear to see. In addition, we can observe the "phantom patterns" for the EWMA and more pronouncedly for the MA after the peak shortly before the 600th day.

Correction to: Derivatives and Internal Models

Correction to:
H.-P. Deutsch, M. W. Beinker, *Derivatives and Internal Models*,
Finance and Capital Markets Series,
https://doi.org/10.1007/978-3-030-22899-6

The book was inadvertently published without the below copyright comment. This copyright statement has been updated in this new version.

Hans-Peter Deutsch/Mark Beinker: Derivate und Interne Modelle. Modernes Risikomanagement. 5., überarbeitete und erweiterte Auflage (ISBN: 978-3-7910-3312-9) published by Schäffer-Poeschel Verlag für Wirtschaft, Steuern und Recht GmbH Stuttgart, Germany. Copyright © 2014.

A

Probability and Statistics

The basics of probability and statistics are presented in this appendix, providing the reader with a reference source for the main body of this book without interrupting the flow of argumentation there with unnecessary excursions. Only the basic definitions and results have been included in this appendix. More complicated concepts (such as stochastic processes, Ito Calculus, Girsanov Theorem, etc.) are discussed directly in the main body of the book as the need arises.

A.1 Probability, Expectation and Variance

A variable whose values is dependent on *random events*, is referred to as a *random variable* or a *random number*. An example of a random variable is the number observed when throwing a dice. The distribution of this random number is an example of a *discrete* distribution. A random number is discretely distributed if it can take on only discrete values such as whole numbers, for instance. A random variable has a *continuous* distribution if it can assume arbitrary values in some interval (which can by all means be the infinite interval from $-\infty$ to ∞). Intuitively, the values which can be assumed by the random variable lie together *densely*, i.e., arbitrarily close to one another.

An example of a discrete distribution on the interval from 0 to ∞ ("infinity") might be a random variable which could take on the values 0.00, 0.01, 0.02, 0.03, …, 99.98, 99.99, 100.00, 100.01, 100.02, …, etc., like a Bund future with a tick-size of 0.01% corresponding to a value change of 10 euros

per tick on a nominal of 100,000 euros. In contrast, a continuously distributed random variable can take on arbitrary values in the pertinent interval, for example, $\sqrt{5}$, π or $7/3$.

The *probability* $P(x < a)$ that a random variable x will be less than some arbitrary number a is the sum of the probabilities of all events in which x takes on values less than a. For continuous distributions, for which the possible values of x lie "infinitely dense" in some interval, the sum takes on the form of an integral:

$$P(x < a) = \sum_{i \text{ where } x_i < a} p(x_i) \longrightarrow \int_{-\infty}^{a} p(x)dx \ . \tag{A.1}$$

The function p is called the *probability density* of the random variable x. It is often referred to as the *probability distribution*, *distribution density* or simply *the distribution*. In this text, the abbreviation pdf for *probability density function* will frequently be used in reference to the function p.

The function P is called the *cumulative probability*. It is often referred to as the *cumulative probability distribution* or simply *cumulative distribution* as well. We will frequently use the abbreviation cpf for *cumulative probability function* in reference to P.

It is certain that a random number will take on *some* value. The probability of something at all happening is thus equal to one. This property is called the *normalization to one* and holds for all probability distributions:

$$1 = \sum_{i} p(x_i) \longrightarrow \int_{-\infty}^{\infty} p(x)dx = 1 \ . \tag{A.2}$$

The *expectation* of x is computed by taking the weighted sum of all possible values taken on by the random variable where the weights are the corresponding probabilities belonging to those values

$$E[x] = \sum_{i} x_i p(x_i) \longrightarrow \int_{-\infty}^{\infty} xp(x)dx \ . \tag{A.3}$$

A function f of a random variable x is again a random variable. The expectation of this function is calculated analogously by taking the weighted

sum of the values of the function $f(x)$ evaluated at all possible values of x. The weights are again the probabilities belonging to the values of x:

$$E[f(x)] = \sum_i f(x_i)p(x_i) \longrightarrow \int_{-\infty}^{\infty} f(x)p(x)dx . \tag{A.4}$$

A special case is particularly interesting: setting $f(x)$ equal to the square of the deviation of x from its expected value, i.e., $f(x) = (x - E[x])^2$, measures how strongly x fluctuates around its expected value. This measure is called the *variance*.

$$\text{var}[x] = E\left[(x - E[x])^2\right] = \sum_i (x_i - E[x])^2\, p(x_i) \longrightarrow \int_{-\infty}^{\infty} (x - E[x])^2\, p(x)dx .$$
$$\tag{A.5}$$

The square root of the variance is called the *standard deviation*, abbreviated as std.

$$\text{std}[x] := \sqrt{\text{var}[x]} = \sqrt{E\left[(x - E[x])^2\right]} . \tag{A.6}$$

For both discrete and continuous distributions, there exists a simple connection between the variance and the expectation: the variance is equal to the difference between the *expectation of the square* of the random variable and the *square of the expectation* of the same random variable:

$$\text{var}[x] = E\left[x^2\right] - E[x]^2 . \tag{A.7}$$

The derivation is presented here for discrete distributions; the proof for the continuous case is completely analogous:

$$E\left[(x - E[x])^2\right] = \sum_i (x_i - E[x])^2\, p(x_i)$$

$$= \underbrace{\sum_i x_i^2 p(x_i)}_{E[x^2]} - 2E[x]\underbrace{\sum_i x_i p(x_i)}_{E[x]} + E[x]^2 \underbrace{\sum_i p(x_i)}_{1}$$

$$= E\left[x^2\right] - 2E[x]^2 + E[x]^2 = E\left[x^2\right] - E[x]^2 .$$

A.2 Multivariate Distributions, Covariance, Correlation and Beta

Two random variables x and y which are not statistically independent (for example, the price of a Siemens share and the DAX) are said to be *correlated*. The probability $P(x < a, y < b)$ that a random variable x will be less than some value a and *simultaneously* that the second random variable y will be less than a value b equals the sum of the probabilities of all events in which $x < a$ and $y < b$. For continuous distributions, i.e., for "infinitely dense" values of the random variables, the sums converge to integrals:

$$P(x < a, y < b) = \sum_{x_i < a} \sum_{y_j < b} p(x_i, y_j) \longrightarrow \int_{-\infty}^{a} dx \int_{-\infty}^{b} dy \, p(x, y) .$$

$$(A.8)$$

The function p is in this case the *joint probability density* for a pair of random variables x and y. Since such density functions refer to more than one (in this case two) random variables, they are referred to as *multivariate probability densities* or *multivariate probability distributions*, *multivariate distribution densities* or simply *multivariate distributions*.

Just as for a single random variable, the expectation of an arbitrary function $f(x, y)$ is calculated by taking the weighted sum of the values of the function $f(x, y)$ evaluated at all possible values of x and y. The weights are the joint probabilities belonging to the value pairs of (x, y):

$$E[f(x, y)] = \sum_i \sum_j p(x_i, y_j) f(x_i, y_j) \longrightarrow \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \, p(x, y) f(x, y) .$$

$$(A.9)$$

A special case is particularly interesting: setting the arbitrary function $f(x, y)$ equal to the product of the deviation of the random variables from their respective expectations, i.e., $f(x, y) = (x - E[x])(y - E[y])$, yields a measure for the fluctuation of each random variable from its expectation as well as the degree of statistical dependence between the two random variables; or in other words, a measure for the variance as well as for the correlation. This expectation is called the *covariance* between x and y.

$$\text{cov}\,[x,y] = \text{E}\,[(x - \text{E}\,[x])\,(y - \text{E}\,[y])] \tag{A.10}$$

$$= \sum_{i,j} (x_i - \text{E}\,[x])\,(y_j - \text{E}\,[y])\,p(x_i, y_j)$$

$$\longrightarrow \int\int dx\,dy\,p(x,y)\,(x - \text{E}\,[x])\,(x - \text{E}\,[x])\ .$$

For both discrete and continuous distributions, there exists a natural extension (which is just as easily shown) of the relation in Eq. A.7 between the variance and the expectations: the covariance is the difference between the expectation of the product and the product of the expectations of the random variables under consideration:

$$\text{cov}\,[x,y] = \text{E}\,[xy] - \text{E}\,[x]\,\text{E}\,[y]\ . \tag{A.11}$$

The symmetry property of the covariance can be seen immediately from this equation; the covariance of x with y is the same as the covariance of y with x. The covariance of a random variable with itself is its variance.

$$\text{cov}\,[x,y] = \text{cov}\,[y,x]\ , \quad \text{cov}[x,x] = \text{var}\,[x]\ . \tag{A.12}$$

The covariance has an additional useful property, it is bilinear: let a, b, c, d be constants and x, y, u, z random variables. Then

$$\text{cov}[ax + by, cu + dz] = ac\ \text{cov}[x, u] + ad\ \text{cov}[x, z] \tag{A.13}$$

$$+ bc\ \text{cov}[y, u] + bd\ \text{cov}[y, z]\ .$$

As often mention in the main body of this book, the covariances between n random variables can be represented in the form of an n by n matrix, called the *covariance matrix*. Because of the symmetry in Eq. A.12, the information in the matrix entries appearing above the diagonal is the same as that below the diagonal; the entries in the diagonal itself are the variances of each of the n respective variables.

Dividing the covariance by the standard deviations of the two respective random variables yields the *correlation coefficient* ρ, also called the *correlation between x and y*. This value always lies between -1 and $+1$:

$$\rho(x, y) := \frac{\text{cov}\,[x, y]}{\sqrt{\text{Var}[x]\,\text{Var}[y]}} = \frac{\text{E}\,[xy] - \text{E}\,[x]\,\text{E}\,[y]}{\sqrt{(\text{E}\,[x^2] - \text{E}\,[x]^2)(\text{E}\,[y^2] - \text{E}\,[y]^2)}} \,.$$

(A.14)

The symmetry property of the correlation can be seen immediately from this equation: The correlation of x with y is the same as the correlation of y with x. The correlation of a random variable with itself equals one.

$$\rho(x, y) = \rho(y, x)\,, \quad \rho(x, x) = 1\,.$$

(A.15)

Two random numbers x_1 and x_2 are called *uncorrelated* if the correlation between the two is zero, i.e., if

$$\rho(x_i, x_j) = \delta_{ij} = \begin{cases} 1 \text{ for } i = j \\ 0 \text{ for } i \neq j \end{cases} \quad \text{mit } i, j \in \{1, 2\}\,,$$

(A.16)

where δ_{ij} denotes the well known *Kronecker delta*.

The *variance of a sum* of random variables is just the sum of the *covariances* of these random variables (in a similar ways as the expectation of a sum equals the sum of the expectations):

$$\text{E}\left[\sum_i x_i\right] = \sum_i \text{E}\,[x]$$

$$\text{Var}\left[\sum_i x_i\right] = \sum_{i,j} \text{cov}[x_i, x_j] = \sum_{i,j} \rho\,(x_i, x_j)\,\sqrt{\text{var}[x_i]\,\text{Var}[x_j]}\,.$$

(A.17)

The first of these equations, i.e., the linearity of the expectation, follows directly from the definition, Eq. A.3. This linearity of the expectation is the only thing needed to show the second equation, i.e., the equation for the

variance of a sum:

$$\text{var}\left[\sum_i x_i\right]$$

$$= \text{E}\left[\left(\sum_i x_i - \text{E}\left[\sum_j x_j\right]\right)^2\right] = \text{E}\left[\left(\sum_i x_i - \sum_j \text{E}\left[x_j\right]\right)^2\right]$$

$$= \text{E}\left[\left(\sum_i (x_i - \text{E}\left[x_i\right])\right)^2\right] = \text{E}\left[\sum_{i,j} (x_i - \text{E}\left[x_i\right])\left(x_j - \text{E}\left[x_j\right]\right)\right]$$

$$= \sum_{i,j} \text{E}\left[(x_i - \text{E}\left[x_i\right])\left(x_j - \text{E}\left[x_j\right]\right)\right] = \sum_{i,j} \text{cov}[x_i, x_j] \ .$$

The first step is just the definition of the variance, Eq. A.5. The linearity of the expectation is used in the second step for the inner expectation and in the fifth step for the outer expectation.

A direct application of Eq. A.17 is the fact that the variance of a sum of *uncorrelated* random numbers (see Eq. A.16) is just the sum of the individual variances:

$$\text{var}\left[\sum_i x_i\right] = \sum_{i,j} \underbrace{\rho\left(x_i, x_j\right)}_{\delta_{ij}} \sqrt{\text{var}[x_i]\,\text{var}[x_j]} = \sum_i \text{var}[x_i] \ . \qquad \text{(A.18)}$$

Thus, the standard deviation, Eq. A.6, of such a sum of *uncorrelated* random numbers is obtained by adding the squares of each individual standard deviation and then taking the square root of this sum:

$$\text{std}\left[\sum_i x_i\right] = \sqrt{\sum_i \text{var}[x_i]} = \sqrt{\sum_i \text{std}[x_i]^2} \ . \qquad \text{(A.19)}$$

Besides the above two symmetric quantities (covariance and correlation), a further *a*symmetric quantity is quite useful as well. This is the *beta* of y with respect to x defined as the covariance of x and y divided by the variance of x:

$$\beta(x, y) = \frac{\text{cov}[x, y]}{\text{var}[y]} = \sqrt{\frac{\text{var}[x]}{\text{var}[y]}} \rho(x, y) = \frac{\text{E}[xy] - \text{E}[x]\text{E}[y]}{\text{E}[y^2] - \text{E}[y]^2} .$$

(A.20)

A symmetry property can be seen immediately in this definition: the beta of a random variable with itself is indeed equal to one, the beta of x with respect to y is however not the same as the beta of y with respect to x. The conversion can be accomplished as follows:

$$\beta(y, x) = \frac{\text{var}[y]}{\text{var}[x]} \beta(x, y) , \quad \beta(x, x) = 1 .$$

(A.21)

Note that all of the above equations (in particular the very useful Eq. A.17 and the properties A.18 and A.19 of uncorrelated variables) have been derived directly from first principles and are therefore valid no matter what probability distribution the random variables may have.

A.3   Moments and Characteristic Functions

The expectation and the variance are examples of the *moments* of a distribution. In general, the n-th moment of the distribution of the random variable x is defined as the expectation of the nth power of the random variable:

$$\text{E}[x^n] = \sum_i x_i^n p(x_i) \longrightarrow \int_{-\infty}^{\infty} x^n p(x) dx .$$

(A.22)

The *central moments* μ_i of a distribution are defined as the "expectation of the powers of the difference between a random variable and its expectation":

$$\mu_j := \text{E}[(x - \text{E}[x])^j] .$$

(A.23)

The first central moment ($j = 1$) is thus by definition equal to zero. The *expectation* of x is the first moment, the *variance* the *second central* moment

of the distribution. With the third central moment $E[(x - E[x])^3]$ we can calculate the *skewness*, a measure for the asymmetry of the density. With the fourth central moment $E[(x - E[x])^4]$ we can calculate the *curtosis*, a measure for the weight of the distribution at the tail ends of its range. The exact definitions of the skewness and curtosis are

$$\text{Skewness} := \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{E[(x - E[x])^3]}{E[(x - E[x])^2]^{3/2}}$$

$$\text{Curtosis} := \frac{\mu_4}{\mu_2^2} = \frac{E[(x - E[x])^4]}{E[(x - E[x])^2]^2} \,. \tag{A.24}$$

A.3.1 Moment Generating Functions

The moment generating function is a very useful tool for the explicit computation of moments. The *moment generating function* (in short *MGF*) of a random variable x with density function $\text{pdf}(x)$ is defined as the expectation of e^{sx} for an arbitrary real value s

$$G_x(s) = E[e^{sx}] = \int_{-\infty}^{\infty} e^{sx} \,\text{pdf}(x)dx \,, \tag{A.25}$$

if this integral exists. This corresponds to the *Laplace transformation* of the pdf. Expanding the exponential function e^{sx} in its Taylor series, we see that the coefficient of s^n is determined by the nth moment of the distribution:

$$G_x(s) = \int_{-\infty}^{\infty} \underbrace{\sum_{n=0}^{\infty} \frac{1}{n!} s^n x^n}_{e^{sx}} \text{pdf}(x)dx$$

$$= \sum_{n=0}^{\infty} \frac{s^n}{n!} \int_{-\infty}^{\infty} x^n \,\text{pdf}(x)dx$$

$$= \sum_{n=0}^{\infty} \frac{s^n}{n!} E[x^n] \,. \tag{A.26}$$

Differentiating the moment generating function with respect to s at the point $s = 0$ yields all moments of the distribution (and thus the name):

$$\left. \frac{\partial^n G_x(s)}{\partial s^n} \right|_{s=0} = E[x^n] . \tag{A.27}$$

This extraordinarily useful fact can be shown as follows:

$$\begin{aligned}
\left. \frac{\partial^n G_x(s)}{\partial s^n} \right|_{s=0} &= \left. \frac{\partial^n}{\partial s^n} \sum_{i=0}^{\infty} \frac{1}{i!} s^i E[x^i] \right|_{s=0} \\
&= \left. \frac{1}{i!} \frac{\partial^{n-1}}{\partial s^{n-1}} \sum_{i=0}^{\infty} E[x^i] \frac{\partial s^i}{\partial s} \right|_{s=0} \\
&= \left. \frac{1}{i!} \frac{\partial^{n-1}}{\partial s^{n-1}} \sum_{i=1}^{\infty} E[x^i] i \, s^{i-1} \right|_{s=0} \\
&= \left. \frac{1}{i!} \frac{\partial^{n-2}}{\partial s^{n-2}} \sum_{i=2}^{\infty} E[x^i] i (i-1) \, s^{i-2} \right|_{s=0} \\
&= \cdots \\
&= \left. \frac{1}{i!} \sum_{i=n}^{\infty} E[x^i] i (i-1) \cdots (i-n+1) s^{i-n} \right|_{s=0} .
\end{aligned}$$

Notice that the lower limit in the sum increases by one each time a derivative is taken. The first term of the sum is always independent of the differentiating variable. For $i = 0$, for example, $\partial s^i / \partial s = \partial s^0 / \partial s = \partial 1 / \partial s = 0$, and so on. In the last step, all derivatives have been performed. The expression can now be evaluated at $s = 0$. Naturally, $s^{i-n} = 0$ for $s = 0$ and $i > n$. Thus, *only* the first summand where $i = n$ makes a contribution to the sum since in this term we have $s^{i-n} = s^0 = 1$. Setting $i = n$ in this term immediately yields Eq. A.27.

The *central* moments defined in Eq. A.23 can likewise be calculated using the moment generating function: If a random number x has a distribution $pdf(x)$ with expectation $\mu = E[x]$, then the moments of the random number $\tilde{x} := x - \mu$ are exactly equal to the *central* moments of x. But for the moments of $\tilde{x}$ we can use the MGF of the distribution of $\tilde{x}$.

$$G_{\tilde{x}}(s) = \int_{-\infty}^{\infty} e^{s\tilde{x}} \, \mathrm{pdf}(\tilde{x}) d\tilde{x} = \int_{-\infty}^{\infty} e^{s\tilde{x}} \, \mathrm{pdf}(x) d\tilde{x}$$

$$= \int_{-\infty}^{\infty} e^{s(x-\mu)} \, \mathrm{pdf}(x) dx = e^{-s\mu} \int_{-\infty}^{\infty} e^{sx} \, \mathrm{pdf}(x) dx \ .$$

The first step is just definition A.25. For the second step we made use of the fact that if the difference between two random numbers is just a constant they must have the same distribution.[1] In the third step we used that $\tilde{x} = x - \mu$ means for the differentials $d\tilde{x} = dx$. Thus the MGF for the central moments is simply $e^{-s\mu}$ times the MGF for the (ordinary) moments where μ denotes the first (ordinary) moment:

$$G_{x-\mu}(s) = e^{-s\mu} G_x(s) \ . \tag{A.28}$$

With this MGF the central moments defined in Eq. A.23 can be calculated completely analogously to Eq. A.27:

$$\mathrm{E}[(x - \mathrm{E}[x])^n] = \left. \frac{\partial^n}{\partial s^n} \exp\left(-s\mathrm{E}[x]\right) G_x(s) \right|_{s=0} \ . \tag{A.29}$$

The general procedure for calculating central moments is thus: first calculate the expectation using Eq. A.27. Then insert the result into Eq. A.29 for the central moments.

For many distributions an explicit analytical expression for the MGF can be obtained using the integral representation in Eq. A.25. This will be demonstrated below for several important distributions. Having obtained such an expression, the moments can be calculated by simply differentiating this function as indicated in Eq. A.27.

The MGF has another very useful property: If two random variables x and y are independent then

$$G_{x+y}(s) = G_x(s) \, G_y(s) \ . \tag{A.30}$$

The *distribution* of a *sum* of independent random variables is generally very difficult to determine, even when the distributions of the individual random

---

[1]This is trivial: if $\tilde{x} = x - \mu$ then the probability for $\tilde{x} < a - \mu$ is of course the same as the probability for $x < a$.

variables in the sum are known.[2] The *MGF* of such a sum, in contrast, can be calculated quite easily by taking the product of the MGFs of each of the distributions! This is the most useful property of the moment generating function. In Eq. A.30, each of the random variables in the sum can by all means be governed by completely *different* distributions. The only condition which needs to be satisfied for Eq. A.30 to hold is the statistical *independence* of the random variables under consideration. Equation A.30 is quite simple to prove:

$$G_{x+y}(s) \equiv \mathrm{E}[e^{s(x+s)}] = \mathrm{E}[e^{sx}e^{sy}] = \mathrm{E}[e^{sx}]\mathrm{E}[e^{sy}] = G_x(s)\,G_Y(s)\,.$$

This is a consequence of the fact that $\mathrm{E}[f(x)g(y)] = \mathrm{E}[f(x)]\mathrm{E}[g(y)]$ holds for arbitrary functions f, g of independent random variables x, y.

A further property of the MGF in connection with Eq. A.30 is that for all non-stochastic values a, b and random variables x we have

$$G_{ax+b}(s) = e^{sb}G_x(as)\,. \tag{A.31}$$

The proof of this result is also quite simple:

$$G_{ax+b}(s) = \mathrm{E}[e^{s(ax+b)}] = e^{bs}\mathrm{E}[e^{(as)x}] = e^{sb}G_x(as)\,.$$

We have already encountered a special case of this in Eq. A.28.

A.3.2 Characteristic Functions

Similar to the moment generating function, the *characteristic function* of a random variable x with probability density function $\mathrm{pdf}(x)$ is defined as the expectation of e^{isx} for an arbitrary real value s

$$\Phi_x(s) := \mathrm{E}[e^{isx}] = \int_{-\infty}^{\infty} e^{isx}\,\mathrm{pdf}(x)dx\,. \tag{A.32}$$

Here, i denotes the imaginary number $i = \sqrt{-1}$. This is just the *Fourier transformation* of the pdf.

---

[2]Only in a few special cases, for example when each of the random variables is normally distributed, can the distribution of the sum be easily specified.

The *Fourier transformation* of the cpf, the *cumulative* distribution function, is sometimes used as well. The results derived below hold for these functions as well. Thus, we will formulate the characteristic function more generally by writing

$$\Phi_x(s) = \int_{-\infty}^{\infty} e^{isx} f(x) dx ,$$

where the function f can be taken to be either $f(x) = \text{pdf}(x)$ or $f(x) = \text{cpf}(x)$.

The advantage of the characteristic function is that its inverse function, the *inverse Fourier transformation* always exists:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx} \Phi_x(s) ds . \tag{A.33}$$

Thus, if Φ_x is known, then the *distribution* (pdf or cpf) can be computed directly (and not only its moments as was the case with the moment generating function). The validity of Eq. A.33 can be shown quite easily:

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx} \Phi_x(s) ds = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx} \int_{-\infty}^{\infty} e^{isx'} f(x') dx' ds$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-is(x-x')} f(x') dx' ds$$

$$= \int_{-\infty}^{\infty} \delta(x - x') f(x') dx'$$

$$= f(x) ,$$

where the *Dirac delta function* was used in the above derivation:

$$\delta(x - x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-is(x-x')} ds .$$

This is not a function in the strict sense but a so-called *distribution* with the defining property

$$\int_{-\infty}^{\infty} \delta(x - x') f(x') dx' := f(x) .$$

It is precisely this property of the delta function that yields the invertibility of the Fourier transformation.

Analogously to the moment generating function, Eq. A.30 holds for the characteristic function as well, i.e.,

$$\Phi_{x+y}(s) = \Phi_x(s)\,\Phi_Y(s) \tag{A.34}$$

for independent random variables x and y. Likewise, for non-stochastic values a, b and a random variable x

$$\Phi_{ax+b}(s) = e^{ibs}\,\Phi_x(as) \tag{A.35}$$

holds. The proof is completely analogous to that of Eq. A.31

A.4   A Collection of Important Distributions

A.4.1 The Uniform Distribution

The simplest of all distributions is the *uniform distribution*. This distribution has both a continuous and a discrete version. A random variable is said to be *uniformly distributed* if its distribution density is constant. The normalizing equation A.2 implies immediately that the distribution density for a discrete random variable with n possible outcomes (or for a continuously distributed random variable taking on values in an interval $[a, b]$) is given by:

$$1 = \sum_{i=1}^{n} \underbrace{p(x_i)}_{\text{constant}} = np \implies p = \frac{1}{n}$$

$$1 = \int_{-\infty}^{\infty} \underbrace{p(x)}_{\text{constant}} dx = p \int_a^b dx = p(b-a) \implies p = \frac{1}{b-a}. \tag{A.36}$$

The expectation and the variance of the continuous form of the uniform distribution are

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{b-a} \int_a^b xdx = \frac{1}{b-a} \frac{b^2-a^2}{2} = \frac{a+b}{2}$$

$$\text{var}[x] = \int_{-\infty}^{\infty} (x - E[x])^2 \, p(x)dx = \frac{1}{b-a} \int_a^b \left(x - \frac{a+b}{2}\right)^2 dx = \frac{(b-a)^2}{12}.$$

$$(A.37)$$

Most random number generators generate uniformly distributed random numbers between 0 and 1. For $a = 0$ and $b = 1$, the expectation and the variance of the uniform distribution is given by $1/2$ and $1/12$, respectively.

The moment generating function of the uniform distribution is by definition A.25

$$G_x(s) \equiv \int_{-\infty}^{\infty} e^{sx} p(x)dx = \frac{1}{b-a} \int_a^b e^{sx}dx = \frac{1}{b-a} \left[\frac{1}{s}e^{sx}\right]_{x=a}^{x=b},$$

Thus

$$G_x(s) = \frac{e^{bs} - e^{as}}{s(b-a)}.$$

$$(A.38)$$

Naive differentiation of G_x with respect to s does not lead us directly to the desired result since the factor s appears in the denominator forbidding us from setting $s = 0$. The function can, however, be written in such a way that it is well defined for this value since as s approaches zero the numerator converges towards zero faster than the denominator (since the exponential function converges towards 1 faster than s converges to zero). This can be seen by expanding the exponential function in its Taylor series as in Eq. A.38:

$$\frac{1}{s}e^{bs} = \frac{1}{s} \sum_{n=0}^{\infty} \frac{1}{n!} b^n s^n = \sum_{n=0}^{\infty} \frac{1}{n!} b^n s^{n-1} = \frac{1}{s} + \sum_{i=0}^{\infty} \frac{b^{i+1}}{(i+1)} \frac{s^i}{i!}$$

and analogously for e^{as}/s. Taking the difference of these two expressions, the $1/s$ term disappears and we are left with an equivalent form of the MGF in Eq. A.38. This is given by

$$G_x(s) = \sum_{n=0}^{\infty} \frac{s^n}{n!} \frac{b^{n+1} - a^{n+1}}{(n+1)(b-a)} \; .$$

Comparison of the coefficients of s^n with Eq. A.26 yields *all* of the moments immediately

$$\mathrm{E}[x^n] = \frac{b^{n+1} - a^{n+1}}{(n+1)(b-a)} \; .$$

For example, the first moment ($n = 1$) is simply

$$\mathrm{E}[x] = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2}$$

which agrees with the result in Eq. A.37.

The characteristic function, A.32, of the continuous uniform distribution function is given by

$$\Phi_x(s) \equiv \int_{-\infty}^{\infty} e^{isx} \frac{1}{b-a} dx = \frac{e^{ibs} - e^{ias}}{is(b-a)} \; . \tag{A.39}$$

This equation can be found in every table of Fourier-transformed functions since it is just the Fourier transform of a constant. Note that the characteristic function can also be found by replacing s with is in the moment generating function Eq. A.38.

A.4.2 The Binomial Distribution and the Bernoulli Experiment

Suppose that an experiment has only two possible outcomes (for example, heads or tails when tossing a coin or up or down in one time step of a binomial model) and the probability of one of the two outcomes (for example, "heads" or "up") of the experiment is p. Then the normalizing equation A.2 implies that the probability of the alternate outcome ("tails" or "down") is $1 - p$. In mathematics, such an experiment is referred to as a *Bernoulli experiment.*

If p is independent of the number of experimental trials then the probability of the observing exactly j outcomes associated with the probability p in n trials is given by the *binomial distribution*.

$$p_n(j) = \binom{n}{j} p^j (1-p)^{n-j} \ . \tag{A.40}$$

Obviously, this is a discrete distribution taking on $n+1$ possible values. The term $p^j (1-p)^{n-j}$ is the probability that the result of the n Bernoulli trials will occur *in a certain order*, for instance, up, up, down, up, down, down, etc., with precisely j "ups". However, since the number of "up" terms does not depend on the order in which they appear, this probability must be multiplied by the number of all *permutations* having j "ups". The number of these permutations is given by the *binomial coefficient*[3]:

$$\binom{n}{j} := \frac{n!}{j!(n-j)!} = \frac{1 \times 2 \times \cdots \times n}{(1 \times 2 \times \cdots \times j)(1 \times 2 \times \cdots \times n - j)} \ . \tag{A.41}$$

This is a result from the theory of combinations which has long been known.

The probability that "up" will be observed *at least k* times is naturally the sum from k to n of the density defined in Eq. A.40. This yields precisely the binomial probability arising in the recombinant binomial trees for European options:

$$B_{n,p}(j \geq k) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} \ . \tag{A.42}$$

The moment generating function of the binomial distribution is, by the definition in Eq. A.25,

$$G_{B_{n,p}}(s) \equiv \mathrm{E}\left[e^{sx}\right] = \sum_{j=0}^{n} e^{sj} p_n(j) = \sum_{j=0}^{n} e^{sj} \binom{n}{j} p^j (1-p)^{n-j}$$

$$= \sum_{j=0}^{n} \binom{n}{j} \left[pe^s\right]^j \left[1-p\right]^{n-j} \ .$$

---

[3]This reads "n choose j" or "n over j".

Note that this is precisely the function appearing in the *binomial formula* for the nth power of a sum $(a + b)^n$ where $a = pe^s$ and $b = 1 - p$. Thus the moment generating function of the binomial distribution is simply

$$G_{B_{n,p}}(s) = \left(pe^s + 1 - p\right)^n \; . \tag{A.43}$$

According to Eq. A.27 all of the moments can now be calculated directly by differentiating the MGF with respect to s. Doing so yields, for example, the expectation as

$$\mathrm{E}[j] = \left. \frac{\partial G_{B_{n,p}}(s)}{\partial s} \right|_{s=0} = \left. n \left(pe^s + 1 - p\right)^{n-1} pe^s \right|_{s=0} = np \; .$$

The second moment can be computed analogously:

$$\mathrm{E}[j^2] = \left. \frac{\partial^2 G_{B_{n,p}}(s)}{\partial s^2} \right|_{s=0} = np \left. \frac{\partial}{\partial s} \left(pe^s + 1 - p\right)^{n-1} e^s \right|_{s=0}$$

$$= np \, (n-1) \left.\left(pe^s + 1 - p\right)^{n-2} pe^s e^s\right|_{s=0} + np \left.\left(pe^s + 1 - p\right)^{n-1} e^s\right|_{s=0}$$

$$= n(n-1)p^2 + np \; .$$

Thus, according to Eq. A.7, the second *central* moment, the variance of the binomial distribution, becomes

$$\mathrm{Var}\,[j] = \mathrm{E}[j^2] - (\mathrm{E}[j])^2 = np(1-p) \; .$$

Summarizing, the expectation and variance of the binomial distribution are given by

$$\mathrm{E}\,[j] \equiv \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} j = np$$

$$\mathrm{Var}\,[j] \equiv \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} (j - np)^2 = np(1-p) \; . \tag{A.44}$$

Analogously to Eq. A.43, the characteristic function of the binomial distribution, can be found from Eq. A.32 to be

$$\Phi_{B_{n,p}}(s) = \left(pe^{is} + 1 - p \right)^n \ .$$
(A.45)

A.4.3 The Normal Distribution and the Central Limit Theorem

It is well known that the sum of random variables is itself a random variable. One of the most important theorems of mathematical statistics, the *central limit theorem*, makes a broad statement on the sum of random variables. The intuitive content of this theorem can be expressed as follow:

Central Limit

The sum of a large number of independent random variables is approximately normally distributed, regardless of how the individual random variables are distributed, if the contribution of each random variable to the sum is almost negligible.

This theorem is the reason for the extraordinary importance of the normal distribution above all others. For instances, the *mean* of some random variables is defined as the *sum* of these random variables divided by the number of variables. Thus, such means are *always* approximately normally distributed according to the central limit theorem, *regardless of the distribution of the random variables*. The fact that such sums are often *not* normally distributed can only mean that the assumptions in the statement of the theorem are not satisfied. Either the measured variables are not (purely) random variables and/or they are not completely independent (uncorrelated). The most common reason, however, is an insufficient number of available trial results. Because only in the *limit*, i.e., for an *infinite* number of random variables, does the central limit theorem hold exactly and not merely as an approximation (hence the name central *limit* theorem).

The *normal distribution*, also called *Gaussian distribution*, is a continuous distribution which is completely determined by two parameters (denoted by μ and σ). The density has the explicit form

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ .$$
(A.46)

The expectation and the variance of the normal distribution according to the definition in Eq. A.4 and Eq. A.5 are given by

$$E[x] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$$

$$\mathrm{var}[x] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2 \ . \tag{A.47}$$

Hence, the two parameters μ and σ are equal to the expectation and the standard deviation of the distribution, respectively. This is quite practical, since through observing the results of random trials x, the expectation and the variance can be approximated by measuring the mean of x and x^2 and applying Eq. A.7, thus obtaining an approximation of the entire distribution. The density of the distribution can be written as

$$p(x) = \frac{1}{\sqrt{2\pi E\left[(x-E[x])^2\right]}} \exp\left\{-\frac{1}{2}\frac{(x-E[x])^2}{E\left[(x-E[x])^2\right]}\right\} \ . \tag{A.48}$$

The ratio of the square of the deviation of the random variable from its expectation to the *expectation* of this same factor appears in the argument of the exponential function.

The probability $P(x \leq a)$ of the event that a random variable x, distributed according to Eq. A.46 with parameter values μ and σ, will be less than or equal to a given value a is the cumulative probability cpf of the normal distribution evaluated at a, frequently denoted by $N_{\mu,\sigma}(a)$. Thus, according to Eq. A.1 we have

$$N_{\mu,\sigma}(a) = \mathrm{cpf}(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{a} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \ . \tag{A.49}$$

The following notation is also often used to express the fact that a random variable x is normally distributed with expectation μ and variance σ^2

$$x \sim N(\mu, \sigma^2) \ . \tag{A.50}$$

The reader should be careful not to confuse this expression with that for the cumulative distribution function $N_{\mu,\sigma}(a)$ in Eq. A.49.

The integral in the definition of the *cumulative normal distribution* cannot be calculated as a closed form expression. But through a simple change of variable

$$y = \frac{x - \mu}{\sigma} \Longrightarrow dx = \sigma\, dy \tag{A.51}$$

$$y(x = -\infty) = -\infty\ , \quad y(x = a) = \frac{a - \mu}{\sigma}$$

the normal distribution can be transformed into the *standard* normal distribution. This is the normal distribution with expectation 0 and variance 1, i.e., with the density function

$$p(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\ . \tag{A.52}$$

In the notation introduced in Eq. A.50 we write

$$y \sim N(0, 1)\ .$$

Random variables which are standard normally distributed with density function A.52 form the basis of many random walk models applied in this book. The probability $P(x \le a)$ that a standard normally distributed random variable x will be less than or equal to a given number a is called the *cumulative standard normal distribution* $N(a)$:

$$N(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}}\, dx\ . \tag{A.53}$$

This distribution appears for instance in the Black-Scholes formula. The density Eq. A.52 of the standard normal distribution is symmetric about zero and thus $p(x) = p(-x)$. From this we derive a very useful *symmetry* relation holding for the standard normal distribution, namely

$$N(-a) = 1 - N(a)\ . \tag{A.54}$$

This symmetry is so important in practical applications that we will provide the proof here. We first write

$$\mathrm{N}(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx = \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx}_{1} - \frac{1}{\sqrt{2\pi}} \int_{a}^{\infty} e^{-\frac{x^2}{2}} dx \ .$$

The first integral equals 1 because of Eq. A.2. In the second integral, we make the substitution

$$u := -x \Rightarrow du = -dx$$
$$u(x = a) = -a \ , \quad u(x = \infty) = -\infty \ ,$$

thus obtaining

$$\mathrm{N}(a) = 1 + \frac{1}{\sqrt{2\pi}} \int_{-a}^{-\infty} e^{-\frac{x^2}{2}} du = 1 - \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-a} e^{-\frac{x^2}{2}} du}_{\mathrm{N}(-a)} \ ,$$

which proves the symmetry relation in Eq. A.54.

The integral in Eq. A.53 cannot be computed explicitly. However, there exist tables and numerical routines for the computation of the cumulative standard normal distribution. A simple polynomial approximation, exact up to six decimal places,[4] is (see for example [1])

$$\mathrm{N}(x) = 1 - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \sum_{i=1}^{5} a_i y^i \quad \text{for} \quad x \geq 0 \qquad (A.55)$$

---

[4]Depending on the circumstances, a more precise approximation might be necessary. E.g., the Black-Scholes formula depends on the difference of two cumulative normal distributions. If both values differ only slightly, the result may have an accuracy much less than these 6 digits. In such a case, a numerical simulation can occasionally deliver more accurate results than the allegedly "analytical" Black Scholes formula.

where

$$y = \frac{1}{1 + 0.2316419\,x}$$

$$a_1 = 0.319381530\,, \quad a_2 = -0.356563782$$

$$a_3 = 1.781477937\,, \quad a_4 = -1.821255978$$

$$a_5 = 1.330274429\,.$$

This approximation holds only for non-negative values $x \geq 0$. But the values for $x < 0$ can simply be obtained by applying the symmetry relation A.54.

Calculations with the *standard* normal distribution are thus quite simple. A frequently used method is therefore to transform normally distributed random variables into *standard* normal random variables via a transformation as in Eq. A.51. Then all necessary calculations are performed using tools like Eqs. A.55 and A.54. Having completed the calculation, an inverse transformation can be performed to determine the original variables. For example, a 99% confidence interval for a standard normal distribution is

$$0,99 = P(y \leq a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{y^2}{2}} dy \Rightarrow a \approx 2,326\,.$$

This is the upper bound for the *standard* normally distributed random variable y. Performing the inverse transformation back to the variable x according to Eq. A.51 yields the boundary of the confidence interval for the original random variable: the expectation plus 2.326 standard deviations.

The moment generating function of the standard normal distribution is, by definition A.25

$$G_{N(0,1)}(s) = E[e^{sx}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2 - 2sx}{2}\right\} dx$$

$$= \exp\left(\frac{1}{2}s^2\right) \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-s)^2}{2}\right\} dx}_{1}\,,$$

where we have used the method of completing the squares in the last step.[5] The remaining integral is precisely the probability that a normally distributed

---

[5] $x^2 - 2sx = (x - s)^2 - s^2.$

random variable $x \sim N(s, 1)$ will take on any arbitrary value. Equation A.2 implies that this probability is 1. The moment generating function of the *standard* normal distribution is thus simply

$$G_{N(0,1)}(s) = \exp\left(\frac{1}{2}s^2\right) . \tag{A.56}$$

The moment generating function of the normal distribution with expectation μ and variance σ now follows immediately from the transformation A.51 with Eq. A.31

$$G_{N(\mu,\sigma^2)}(s) = e^{\mu s} G_{N(0,1)}(\sigma s) = \exp\left(\mu s + \frac{1}{2}\sigma^2 s^2\right) . \tag{A.57}$$

From this equation, all *moments* of a normal distribution can be calculated by means of Eq. A.27. The *central* moments are given by Eq. A.29:

$$
\begin{aligned}
E[(x - E[x])^n] &= \frac{\partial^n}{\partial s^n} \exp\left(-s\mu\right) \exp\left(\mu s + \frac{1}{2}\sigma^2 s^2\right)\Big|_{s=0} \\
&= \frac{\partial^n}{\partial s^n} \exp\left(\frac{1}{2}\sigma^2 s^2\right)\Big|_{s=0} .
\end{aligned}
$$

The second central moment, i.e., the variance is

$$
\begin{aligned}
E[(x - E[x])^2] &= \frac{\partial^2}{\partial s^2} e^{\sigma^2 s^2/2}\Big|_{s=0} \\
&= \frac{\partial}{\partial s} \sigma^2 s e^{\sigma^2 s^2/2}\Big|_{s=0} \\
&= \sigma^2 e^{\sigma^2 s^2/2} + \left(\sigma^2 s\right)^2 e^{\sigma^2 s^2/2}\Big|_{s=0} \\
&= \left(1 + \sigma^2 s^2\right)\sigma^2 e^{\sigma^2 s^2/2}\Big|_{s=0} \\
&= \sigma^2 .
\end{aligned}
$$

Differentiating the result in the penultimate step yields the third central moment

$$\begin{aligned}
E[(x - E[x])^3] &= \left.\frac{\partial^3}{\partial s^3} e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= \left.\frac{\partial}{\partial s}\left(1 + \sigma^2 s^2\right)\sigma^2 e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= \left. 2s\sigma^4 e^{\sigma^2 s^2/2} + \left(1 + \sigma^2 s^2\right)s\sigma^4 e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= \left.\left(3s + \sigma^2 s^3\right)\sigma^4 e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= 0 \, .
\end{aligned}$$

Again, differentiating the result in the penultimate step yields the next (i.e., fourth) central moment

$$\begin{aligned}
E[(x - E[x])^4] &= \left.\frac{\partial^4}{\partial s^4} e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= \left.\frac{\partial}{\partial s}\left(3s + \sigma^2 s^3\right)\sigma^4 e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= \left.\left(3 + 3\sigma^2 s^2\right)\sigma^4 e^{\sigma^2 s^2/2} + \left(3s + \sigma^2 s^3\right)\sigma^6 s e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= \left.\left(3 + 3\sigma^2 s^2 + 3\sigma^2 s + \sigma^4 s^3\right)\sigma^4 e^{\sigma^2 s^2/2}\right|_{s=0} \\
&= 3\sigma^4 \, .
\end{aligned}$$

We summarize these results for reference:

$$\begin{aligned}
E[x] &= \mu \\
E[(x - E[x])^2] &= \sigma^2 \\
E[(x - E[x])^4] &= 3\sigma^4 \\
E[(x - E[x])^n] &= 0 \quad \text{for all odd } n \, .
\end{aligned} \tag{A.58}$$

From these moments the *skewness* and the *curtosis* of the normal distribution follow directly from their respective definitions, Eq. A.24:

$$\text{kewness} \equiv \frac{E[(x - E[x])^3]}{E[(x - E[x])^2]^{3/2}} = 0$$

$$\text{urtosis} \equiv \frac{E[(x - E[x])^4]}{E[(x - E[x])^2]^2} = 3 \ . \tag{A.59}$$

The characteristic function A.32 of the standard normal density function can be found by replacing s with is in Eq. A.56

$$\Phi_{N(0,1)} = \exp\left(-\frac{1}{2}s^2\right) \ . \tag{A.60}$$

The characteristic function of a normal distribution with expectation μ and variance σ follows immediately from Eq. A.35

$$\Phi_{N(\mu,\sigma^2)}(s) = e^{i\mu s}\Phi_{N(0,1)}(\sigma s) = \exp\left(i\mu s - \frac{1}{2}\sigma^2 s^2\right) \ . \tag{A.61}$$

The Multivariate Normal Distribution

If two random variables $x_i, i = 1, 2$ are both normally distributed with expectation μ_i and covariance[6]

$$\text{cov}\left[x_i, x_j\right] = \sigma_i \rho_{ij} \sigma_j \ ,$$

then their *joint probability distribution*, i.e., the probability that both random numbers simultaneously will have certain values, is given by the *bivariate normal distribution* with the density

$$p(x_1, x_2) = \frac{\exp\left\{-\frac{1}{1-\rho_{12}^2}\frac{1}{2}\sum_{i,j=1,2}\frac{x_i-\mu_i}{\sigma_i}\rho_{ij}\frac{x_j-\mu_j}{\sigma_j}\right\}}{\sqrt{1-\rho_{12}^2}\prod_{i=1,2}\sqrt{2\pi\sigma_i^2}} \ .$$

---

[6]This means that they have variances σ_1^2 and σ_2^2 and the correlation between the two is ρ_{12} (see Eq. A.14).

With this density, the joint (cumulative) probability, as defined in Eq. A.8, for $x_1 < a$ and $x_2 < b$ is

$$P(x_1 < a, x_2 < b) = \frac{\int_{-\infty}^{a} dx_1 \int_{-\infty}^{b} dx_2 \exp\left\{ -\frac{1}{1-\rho_{1,2}^2} \frac{1}{2} \sum_{i,j=1}^{2} \frac{x_i - \mu_i}{\sigma_i} \rho_{ij} \frac{x_j - \mu_j}{\sigma_j} \right\}}{\sqrt{1-\rho_{12}^2} \prod_{i=1}^{2} \sqrt{2\pi\sigma_i^2}} .$$

If the two random variables are *uncorrelated*, i.e., if

$$\rho_{ij} = \delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases},$$

where δ_{ij} denotes the *Kronecker delta*, then the joint distribution of the variables is equal to the product of the distributions of each individual variable:

$$
\begin{aligned}
p(x_1, x_2) &= \frac{\exp\left\{ -\frac{1}{1-0} \frac{1}{2} \sum_{i,j=1,2} \frac{x_i - \mu_i}{\sigma_i} \delta_{ij} \frac{x_j - \mu_j}{\sigma_j} \right\}}{\sqrt{1-0} \prod_{i=1,2} \sqrt{2\pi\sigma_i^2}} \\
&= \frac{\exp\left\{ -\frac{1}{2} \sum_{i=1,2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right\}}{\prod_{i=1,2} \sqrt{2\pi\sigma_i^2}} \\
&= \prod_{i=1,2} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{1}{2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right\} \\
&= \prod_{i=1,2} p(x_i) .
\end{aligned}
\tag{A.62}
$$

If the joint probability density of two random variables x_1 and x_2 fulfills

$$p(x_1, x_2) = p(x_1) p(x_2) , \tag{A.63}$$

then the two random variables are said to be *independent*. As we have just seen, a *necessary* condition for independence is that the two variables are uncorrelated. However, this is generally not a *sufficient* condition, i.e., uncorrelated random variables are not always independent. But—as we have just seen—in

the special case of normally distributed random variables independence and uncorrelation are equivalent.

A.4.4 The Lognormal Distribution

A random variable is *lognormally distributed* if its logarithm is normally distributed. Since the lognormal distribution is defined through the normal distribution, it is also continuous and completely determined by the two parameters μ and σ. The precise definition of the *lognormal distribution* can be stated as follows: let x be a lognormally distributed random variable with parameters μ and σ. Then the probability $H_{\mu,\sigma}(a)$ that x will be less than a given value a is equal to the cumulative distribution function of the normal distribution, Eq. A.49, with the same parameters, evaluated at $\ln(a)$:

$$H_{\mu,\sigma}(a) := N_{\mu,\sigma}(\ln(a)) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\ln(a)} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \ . \tag{A.64}$$

The density function of the lognormal distribution can be derived from the above definition. It has the explicit form

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left\{ -\frac{1}{2} \frac{(\ln(x)-\mu)^2}{\sigma^2} \right\} \ . \tag{A.65}$$

Note the factor $1/x$. The density is thus not merely obtained by substituting x with $\ln(x)$ in Eq. A.46. With this density function we can express $H_{\mu,\sigma}$ explicitly as

$$H_{\mu,\sigma}(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^a \exp\left\{ -\frac{1}{2} \frac{(\ln(x)-\mu)^2}{\sigma^2} \right\} \frac{1}{x} dx \ . \tag{A.66}$$

Note that the lower limit in the integral is zero. The range of such a lognormally distributed random number lies only between zero and infinity.

With the simple change in variable given by $u = \ln(x)$ we can immediately verify that the cumulative distribution function $H_{\mu,\sigma}$ actually satisfies the definition in Eq. A.64:

$$H_{\mu,\sigma}(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^a \exp\left\{-\frac{1}{2}\frac{(\ln(x) - \mu)^2}{\sigma^2}\right\} \frac{1}{x} dx$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\ln(0)=-\infty}^{\ln(a)} \exp\left\{-\frac{1}{2}\frac{(u - \mu)^2}{\sigma^2}\right\} du = N_{\mu,\sigma}(\ln(a)) ,$$

where $u := \ln(x) \Longrightarrow du = \frac{1}{x}dx, \quad x = e^u .$

The expectation and variance of the lognormal distribution can be calculated using Eqs. A.4 and A.5 as

$$E[x] = \int_0^\infty xp(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} dx = e^{\mu + \sigma^2/2}$$

$$\text{var}[x] = \int_0^\infty (x - E[x])^2 p(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty (x - e^{\mu+\sigma^2/2})^2 \frac{1}{x} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} dx$$

$$= e^{2\mu}(e^{2\sigma^2} - e^{\sigma^2}) . \tag{A.67}$$

In general, the moments of the lognormal distribution are given by

$$E[x^n] = \exp\left(n\mu + n^2\sigma^2/2\right) \quad \text{for } n = 1, 2, \ldots \tag{A.68}$$

A.4.5 The Gamma Distribution

The *gamma distribution* is an important distribution because it encompasses a whole class of different distributions (which includes the exponential distribution and the χ^2-distribution, for example). Like the lognormal distribution, the range of the gamma distribution consists solely of the non-negative real numbers. A random variable x has a gamma distribution with parameters λ

and t if it is governed by the following distribution density

$$\text{pdf}(x) = \frac{1}{\Gamma(t)} \lambda^t x^{t-1} e^{-\lambda x} \quad \text{with} \quad \lambda > 0, t > 0 \quad \text{and} \quad x \in [0, \infty[\, ,$$

$$(\text{A.69})$$

where $\Gamma(t)$ denotes the *gamma function*. This function is defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \qquad (\text{A.70})$$

A description of its properties can be found in any mathematical collection of special mathematical functions (see for instance [1]). An important property of the gamma function is the recursion

$$\Gamma(t+1) = t\,\Gamma(t)\,. \qquad (\text{A.71})$$

This allows the gamma function to be interpreted as a generalization of the factorial operation. Two special function values which often serve as the initial values in the above recursion relation are given by

$$\Gamma(1) = 1\,, \quad \Gamma(\tfrac{1}{2}) = \sqrt{\pi}\,. \qquad (\text{A.72})$$

It follows that the gamma function for whole numbers t is in fact nothing other than the factorial:

$$\Gamma(n) = (n-1)! \quad \text{for } n = 1, 2, \ldots$$

Furthermore, this function has a symmetry property which allows it to be evaluated for negative values if the function evaluated at the corresponding positive value is known:

$$\Gamma(-x) = -\frac{\pi}{x \sin(\pi x)} \frac{1}{\Gamma(x)}\,.$$

From this symmetry relation, we see immediately that the gamma function has a simple pole at each negative integer value of x. It is well defined for all positive values. Fortunately, in our consideration of the gamma distribution, Eq. A.69, our needs are restricted to the positive arguments of the gamma function only.

The explicit form of the moment generating function of the gamma distribution can be calculated quite easily. Consider the definition A.25. Then for the gamma distribution, we have

$$G_{\Gamma(t,\lambda)}(s) = \mathrm{E}[e^{sx}] = \frac{\lambda^t}{\Gamma(t)} \int_0^\infty x^{t-1} e^{-(\lambda-s)x} dx .$$

The integral appearing in this expression is quite similar to that in the definition of the gamma function, Eq. A.70. We make the substitution

$$u := (\lambda - s)x \implies du = (\lambda - s)dx,$$
$$u(x = 0) = 0 , \quad u(x = \infty) = \infty \quad \text{for } s < \lambda .$$

The condition $s < \lambda$ is required in order for the upper limit of the integral to remain equal to $+\infty$. From Eq. A.27 we see that the s of interest are in a small neighborhood of zero. Since λ in the distribution Eq. A.69 is strictly greater than zero, the condition $s < \lambda$ is no obstacle for our purposes here. A simple substitution yields

$$
\begin{aligned}
G_{\Gamma(t,\lambda)}(s) &= \frac{\lambda^t}{\Gamma(t)} \int_0^\infty \frac{1}{(\lambda-s)^{t-1}} u^{t-1} e^{-u} \frac{1}{\lambda-s} du \\
&= \frac{\lambda^t}{\Gamma(t)} \frac{1}{(\lambda-s)^t} \underbrace{\int_0^\infty u^{t-1} e^{-u} du}_{\Gamma(t)} .
\end{aligned}
$$

The moment generating function of the gamma distribution thus has the following simple form:

$$G_{\Gamma(t,\lambda)}(s) = \left(\frac{\lambda}{\lambda-s} \right)^t \quad \text{mit} \quad \lambda > 0,\ t > 0,\ s < \lambda . \tag{A.73}$$

Replacing s with is in Eq. A.73 yields the characteristic function given in Eq. A.32 for the gamma distribution

$$\Phi_{\Gamma(t,\lambda)}(s) = \left(\frac{\lambda}{\lambda-is} \right)^t . \tag{A.74}$$

All *moments* can be obtained as indicated in Eq. A.27 directly by differentiating the expression in Eq. A.73 with respect to s. For example, the expectation becomes

$$\mathrm{E}[x] = \left.\frac{\partial G_{\Gamma(t,\lambda)}(s)}{\partial s}\right|_{s=0} = \left.\lambda^t t \frac{1}{(\lambda - s)^{t+1}}\right|_{s=0} = \frac{t}{\lambda} \ . \tag{A.75}$$

The second moment is

$$\mathrm{E}[x^2] = \left.\frac{\partial^2 G_{\Gamma(t,\lambda)}(s)}{\partial s^2}\right|_{s=0} = \left.\lambda^t t(t+1) \frac{1}{(\lambda - s)^{t+2}}\right|_{s=0} = \frac{t(t+1)}{\lambda^2} \ .$$

Proceeding analogously, *all* moments can be explicitly calculated for the gamma distribution:

$$\mathrm{E}[x^n] = \frac{t(t+1)\cdots(t+n-1)}{\lambda^n} \quad \text{for } n = 1, 2, \ldots \tag{A.76}$$

The variance is then

$$\mathrm{Var}[x] = \mathrm{E}[x^2] - \mathrm{E}[x]^2 = \frac{t}{\lambda^2} \ . \tag{A.77}$$

A.4.6 The χ^2-Distribution

As was shown in Sect. A.4.3, a sum of normally distributed random variables is itself normally distributed. A situation frequently encountered (for example, in a value at risk computation or the determination of variances from historical data) involves taking sums of the *squares* of random variables. If the random variables whose squares are added have a standard normal distribution, the distribution of the resulting random variable is easily determined: the sum of n *squared, independent, standard normally distributed random variables*, x_i, ($i = 1, \ldots, n$) has a distribution known as the χ^2-*distribution with n degrees of freedom*

$$x_i \sim \mathrm{N}(0,1) \ , \quad i = 1, \ldots, n \ , \quad x_i \, iid \implies \sum_{i=1}^{n} x_i^2 =: y \sim \chi^2(n) \ . \tag{A.78}$$

It is essential that the random variables x_i in the above definition be independent. The *degree of freedom*, n, of the χ^2-distribution $\chi^2(n)$ can be intuitively thought of as the number of independent (standard normal) random variables which "make up" the random variable with the $\chi^2(n)$-distribution; thus the name "degree of freedom".

It is often the case that the above sum is taken over only one single element. If a random variable x has a standard normal distribution, then this random variable *squared* is governed by the χ^2-*distribution with one degree of freedom.* We write

$$x \sim N(0, 1) \implies x^2 =: y \sim \chi^2(1) . \qquad (A.79)$$

In fact, $\chi^2(1)$ (or its non-central counterpart, see below) is sufficient for almost all our needs if we are working with moment generating functions since via Eqs. A.30 and A.31, the MGFs of sums of independent random variables can be written as products of MGFs of $\chi^2(1)$.

The MGF also proves to be a helpful tool in calculating the density function of the χ^2-distribution. We first generalize the statement of the problem somewhat, keeping in mind that our goal is to arrive at the MGF of $\chi^2(1)$.

Suppose a random variable x is distributed according to some known distribution whose distribution density we denote by pdf_x. Then, for any arbitrary function f, the random variable $y = f(x)$ is distributed according to another distribution function pdf_y. This distribution is in general unknown but from the construction it is immediately obvious that the probability for the function f to assume a value y is just the probability of the original stochastic variable to assume the value $x = f^{-1}(y)$ (for simplicity, we assume here that $f(\cdot)$ is invertible):

$$\text{pdf}_y(y) = \text{pdf}_x(x = f^{-1}(y)) . \qquad (A.80)$$

Let's now look at the MGF. The MGF of the unknown distribution pdf_y is the expectation of a function g of y (specifically, $g(y) = e^{sy}$). According to Eq. A.4, the expectation of a function g of y is given by

$$E[g(y)]_{\text{pdf}_y} = \int g(y)\, \text{pdf}_y(y) dy .$$

Here, the expectation is computed with respect to the (unknown) distribution pdf_y (this fact is emphasized by the subscript in the expectation). Since $y = f(x)$, the function

$$g(y) = g(f(x)) = h(x) \tag{A.81}$$

is also a function of x and therefore the expectation of h is according to Eq. A.4

$$\text{E}\,[h(x)]_{\text{pdf}_x} = \int h(x)\,\text{pdf}_x(x)dx \ .$$

Here, the expectation is computed by means of the *known* distribution pdf_x.

Computing the expectation of $h(x)$ means integrating over *all* possible values of x where each x is weighted with $\text{pdf}_x(x)$. Since by definition $y = f(x)$ this operation can also be viewed as integrating over *all possible* values of y where each random number $y = f(x)$ is weighted with the weight $\text{pdf}_x(x = f^{-1}(y))$:

$$\text{E}\,[h(x)]_{\text{pdf}_x} = \int h(x)\,\text{pdf}_x(x)dx$$

$$= \int h(x)\,\text{pdf}_x(f^{-1}(y))dy \ .$$

But this weight is just the density $\text{pdf}_y(y)$, see Eq. A.80. We can therefore write

$$\text{E}\,[h(x)]_{\text{pdf}_x} = \int h(x)\,\text{pdf}_y(y)dy$$

$$= \int g(y)\,\text{pdf}_y(y)dy \ ,$$

where we have used Eq. A.81 in the last step. The expression we have now arrived at is simply the expectation of the function g with respect to the distribution pdf_y. In summary we thus have the quite general result

$$\text{E}\,[g(y)]_{\text{pdf}_y} = \text{E}\,[g(f(x))]_{\text{pdf}_x} \quad \text{for} \quad y = f(x) \ . \tag{A.82}$$

Choosing the function $g(y) = e^{sy}$, the left hand side becomes the MGF of the unknown distribution pdf_y

$$G_{\text{pdf}_y}(s) = \text{E}[e^{sy}]_{\text{pdf}_y} = \int e^{sy}\, \text{pdf}_y(y)dy$$

$$= \int e^{sf(x)}\, \text{pdf}_x(x)dx = \text{E}[e^{sf(x)}]_{\text{pdf}_x} \; .$$

Thus, for every function f of a random variable x, the moment generating function of the distribution of $f(x)$ can be expressed through the expectation with respect to the distribution of x:

$$G_{\text{pdf}_{f(x)}}(s) = \text{E}[e^{sf(x)}]_{\text{pdf}_x} \; . \tag{A.83}$$

For the χ^2-distribution with one degree of freedom, we have $f(x) = x^2$ and obtain[7] the moment generating function of $\chi^2(1)$ as

$$G_{\chi^2(1)}(s) = \text{E}[e^{sx^2}]_{\text{N}(0,1)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx^2} e^{-x^2/2} dx \; . \tag{A.84}$$

This integral can be solved explicitly using the substitution

$$u = x\sqrt{1 - 2s} \Rightarrow dx = \frac{du}{\sqrt{1 - 2s}} \tag{A.85}$$

$$u(x = -\infty) = -\infty \quad \text{for} \quad s < 1/2$$

$$u(x = +\infty) = +\infty \quad \text{for} \quad s < 1/2 \; .$$

Note that the integration limits remain the same only for $s < 1/2$.

$$\text{E}[e^{sx^2}]_{\text{N}(0,1)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2s)x^2/2} dx$$

$$= \frac{1}{\sqrt{1 - 2s}} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} du}_{1} \; .$$

---

[7]With the notation: $\text{pdf}_{f(x)} = \text{pdf}_{x^2} = \chi^2(1)$, $\text{pdf}_x = \text{N}(0, 1)$.

The remaining integral is the probability that a standard normal random variable will assume *any* arbitrary value and, according to Eq. A.2, equals one. Thus, the MGF is simply

$$G_{\chi^2(1)}(s) = \frac{1}{\sqrt{1 - 2s}} \, . \tag{A.86}$$

Since the random variables in the sum appearing in definition A.78 are all *independent* we can use Eq. A.30 to immediately obtain the MGF for a χ^2-distribution with n degrees of freedom

$$G_{\chi^2(n)}(s) = \frac{1}{(1 - 2s)^{n/2}} \quad \text{for} \quad n = 1, 2, \dots \tag{A.87}$$

The *moments* can be derived directly from Eq. A.27 by differentiating the function in Eq. A.87 with respect to s or simply from Eq. A.76 with $\lambda = 1/2$ and $t = n/2$:

$$
\begin{aligned}
E[x^k]_{\chi^2(n)} &= \frac{t(t+1)\cdots(t+k-1)}{\lambda^k} \\
&= \frac{\frac{n}{2}(\frac{n}{2}+1)\cdots(\frac{n}{2}+k-1)}{\left(\frac{1}{2}\right)^k} \\
&= 2^k \frac{n}{2}(\frac{n}{2}+1)\cdots(\frac{n}{2}+k-1) \\
&= n(n+2)\cdots(n+2(k-1)) \, .
\end{aligned}
$$

Therefore

$$E[x^k]_{\chi^2(n)} = \prod_{i=0}^{k-1}(n + 2i) \, . \tag{A.88}$$

For example, the expectation and variance are given by

$$E[x]_{\chi^2(n)} = n \, , \quad \text{var}[x]_{\chi^2(n)} = 2n \, .$$

From the definition in Eq. A.32, the characteristic function of the χ^2-distribution can be obtained by replacing s with is in Eq. A.87 or from Eq. A.74 with $\lambda = 1/2$ and $t = n/2$.

$$\Phi_{\chi^2(n)}(s) = \frac{1}{(1-2is)^{n/2}} \quad \text{for} \quad n = 1, 2, \ldots \tag{A.89}$$

The Density of the χ^2-Distribution with One Degree of Freedom

The MGF Eq. A.84 has been derived based on the general Transformation A.82. It is also possible to calculate the MGF without using Transformation A.82. This has the advantage that an explicit expression for the density of the χ^2-Distribution emerges.

So let $y := x^2$ with $x \sim \mathrm{N}(0, 1)$ as in Eq. A.79. Denote the desired density of the χ^2-Distribution with p and the Standard normal density with q.

$$p(y) := \mathrm{pdf}_{\chi^2(1)}(y)$$

$$q(x) := \mathrm{pdf}_{\mathrm{N}(0,1)}(x) = \frac{1}{\sqrt{2\pi}}e^{-x/2} \ .$$

Then, the cumulative probability is

$$
\begin{aligned}
\int_{-\infty}^{a} p(y)dy &= P(y < a) \\
&= P(x^2 < a) \\
&= P(-\sqrt{a} < x < \sqrt{a}) \\
&= 2P(0 < x < \sqrt{a}) \\
&= 2\int_{0}^{\sqrt{a}} q(x)dx \ .
\end{aligned}
$$

We now differentiate with respect to the upper integration limit a. The left side becomes

$$\frac{\partial}{\partial a}\int_{-\infty}^{a} p(y)dy = p(a)$$

And with the chain rule the right side reads

$$\frac{\partial}{\partial a}2\int_{0}^{\sqrt{a}} q(x)dx = \frac{1}{2\sqrt{a}}2q(\sqrt{a}) = \frac{q(\sqrt{a})}{\sqrt{a}} \ .$$

Comparing both sides yields the relation between the normal and the χ^2-distribution

$$\text{pdf}_{\chi^2(1)}(a) = \frac{1}{\sqrt{a}} \text{pdf}_{\text{N}(0,1)}(\sqrt{a}) \ . \tag{A.90}$$

Finally, inserting the standard normal density we arrive at the explicit form of the χ^2-distribution

$$\text{pdf}_{\chi^2(1)}(a) = \frac{1}{\sqrt{2\pi a}} e^{-a/2} = \frac{1}{\sqrt{2\pi a \exp(a)}} \ . \tag{A.91}$$

With Eq. A.90, the MGF Gl. A.84 can now be derived directly without referring to the general Transformation A.82:

$$
\begin{aligned}
\text{E}[e^{sy}]_{\chi^2(1)} &= \int_{-\infty}^{\infty} e^{sy} \underbrace{p(y)}_{=0 \text{ für } y<0} dy \\
&= \int_{0}^{\infty} e^{sy} p(y) dy \\
&= \int_{0}^{\infty} e^{sx^2} \frac{\text{pdf}_{\text{N}(0,1)}(\sqrt{y})}{\sqrt{y}} \underbrace{dy}_{=2xdx} \\
&= \int_{0}^{\infty} e^{sx^2} \frac{\text{pdf}_{\text{N}(0,1)}(x)}{x} 2x dx \\
&= 2 \int_{0}^{\infty} e^{sx^2} \text{pdf}_{\text{N}(0,1)}(x) dx \\
&= \int_{-\infty}^{\infty} e^{sx^2} \text{pdf}_{\text{N}(0,1)}(x) dx \\
&= \text{E}[e^{sx^2}]_{\text{N}(0,1)} \ .
\end{aligned}
$$

Here we used Eq. A.90 for $p(y)$ in the 3rd row and made the substitution $x = \sqrt{y}$ in the 4th row.

An explicit expression for the density function of the χ^2-distribution can also be obtained by taking the following approach: if we succeed in transforming Eq. A.84 into the form $\int e^{sx} p(x) dx$, this integral can be interpreted as the expectation of e^{sx} taken with respect to the probability density $p(x)$.

This expectation is then by definition the MGF of $p(x)$. On the other hand, Eq. A.84 is the MGF of $\chi^2(1)$-distribution. Since two distribution are exactly equal if they have the same MGF[8] we could then conclude that p is the density function of $\chi^2(1)$. In this way, knowledge of p would give us an explicit expression for the density function of $\chi^2(1)$. We wish to take this approach now.

First, we observe that in the integrand in Eq. A.84, x appears only in the form x^2, i.e., the integral is symmetric about zero; thus we can write

$$G_{\chi^2(1)}(s) = \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{sx^2} e^{-x^2/2} dx \ .$$

We then perform the following substitution in the integral

$$y := x^2 \Longrightarrow x = +\sqrt{y} \quad \text{since} \quad x \geq 0$$
$$y(x = 0) = 0 \ , \quad y(x = \infty) = \infty$$
$$\frac{dy}{dx} = 2x \Longrightarrow dx = \frac{1}{2\sqrt{y}} dy$$

and obtain

$$G_{\chi^2(1)}(s) = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{sy} e^{-y/2} \frac{1}{\sqrt{y}} dy$$
$$= \int_0^\infty e^{sy} p(y) dy = \mathrm{E}[e^{sy}]_p \ .$$

We have thus attained our goal. The MGF of $\chi^2(1)$ is expressed as the expectation of e^{sy} with respect to a density function p. This density is the explicit expression for the density function of $\chi^2(1)$

$$\mathrm{pdf}_{\chi^2(1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-x/2} x^{-1/2} = \frac{1}{\sqrt{2\pi x e^x}} \quad \text{with} \quad x \in [0, \infty[\ ,$$
$$\text{(A.92)}$$

---

[8]The distribution is completely determined by the MGF. This implies that if two MGFs are equal then they are the MGF of the *same* distribution.

in complete agreement to Eq. A.91. The range of an χ^2-distributed random variable is thus restricted to the positive real numbers which makes sense since it is the *square* of a standard normally distributed random variable. Comparing this density to the density of the gamma distribution in Eq. A.69 shows that $\chi^2(1)$ is equal to the gamma distribution for the parameters $\lambda = 1/2$ and $t = 1/2$. This can also be seen immediately by comparing their moment generating function in Eqs. A.86 and A.73; they are equal for $\lambda = 1/2$ and $t = 1/2$ and thus the associated density functions must be the same for these parameter values as well. Likewise, a comparison of the moment generating functions in Eqs. A.87 and A.73 shows that the χ^2-distribution with n degrees of freedom equals the gamma distribution with parameters $\lambda = 1/2$ and $t = n/2$:

$$\mathrm{pdf}_{\chi^2(n)}(x) = \frac{1}{\Gamma(n/2)}\left(\frac{1}{2}\right)^{n/2} x^{\frac{n}{2}-1} e^{-x/2} \quad \text{with} \quad x \in [0, \infty[\, , \quad n = 1, 2, \dots \, .$$

(A.93)

Due to the recursion relation in Eq. A.71 with the initial values A.72, the gamma functions appearing here can be given explicitly as:

$$\Gamma(n/2) = \begin{cases} (n/2 - 1)! & \text{for even values of } n \\ (n/2 - 1)(n/2 - 2)(n/2 - 3)\cdots(1/2)\sqrt{\pi} & \text{for odd values of } n \end{cases} .$$

The Non-central χ^2-Distribution

The χ^2-distribution described above is the distribution of a sum of n squared independent *standard* normal random numbers x_i, $(i = 1, \dots, n)$. A slight but often needed generalization of this is the situation in which the random numbers x_i have expectations $\mu_i \neq 0$. The distribution of a sum of n squared random numbers of this type is called the *non-central* χ^2-distribution with n degrees of freedom and with *non-central parameter* θ, where θ denotes the sum of the squared expectations μ_i:

$$x_i \sim \mathrm{N}(\mu_i, 1)\,, \quad i = 1, \dots, n\,, \quad x_i \text{ iid}$$

$$\Longrightarrow$$

$$\sum_{i=1}^{n} x_i^2 =: y \sim \chi^2(n, \theta) \quad \text{with} \quad \theta = \sum_{i=1}^{n} \mu_i^2 \, .$$

(A.94)

The square of a single random number $x \sim N(0, \mu)$ thus has the non-central χ^2-distribution with one degree of freedom:

$$x \sim N(\mu, 1) \Longrightarrow x^2 =: y \sim \chi^2(1, \mu^2) .$$

To determine the moment generating function of the non-central χ^2-distribution we again start from the general Eq. A.83:

$$G_{\chi^2(1,\mu^2)}(s) = E[e^{sx^2}]_{N(\mu,1)}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx^2} e^{-(x-\mu)^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ sx^2 - \frac{1}{2}(x-\mu)^2 \right\} dx .$$

The decisive step to calculate this integral is to complete the square in the argument of the exp-function:

$$sx^2 - \frac{1}{2}(x-\mu)^2 = -\frac{1}{2}\left[(1-2s)x^2 - 2\mu x + \mu^2 \right]$$

$$= -\frac{1}{2}\left[\left(\sqrt{1-2s}\, x - \frac{1}{\sqrt{1-2s}}\mu \right)^2 - \frac{1}{1-2s}\mu^2 + \mu^2 \right]$$

$$= -\frac{1}{2}\left(\sqrt{1-2s}\, x - \frac{1}{\sqrt{1-2s}}\mu \right)^2 + \frac{s\mu^2}{1-2s} .$$

Thus, the expectation becomes

$$E[e^{sx^2}]_{N(\mu,1)}$$

$$= \exp\left\{ \frac{s\mu^2}{1-2s} \right\} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2}\left(\sqrt{1-2s}\, x - \frac{1}{\sqrt{1-2s}}\mu \right)^2 \right\} dx$$

$$= \frac{1}{\sqrt{1-2s}} \exp\left\{ \frac{s\mu^2}{1-2s} \right\} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2}\left(u - \frac{\mu}{\sqrt{1-2s}} \right)^2 \right\} du}_{1} ,$$

where in the last step we have used the substitution A.85 (again with the condition $s < 1/2$). The remaining integral is the probability that a normally distributed random number with expectation $\mu/\sqrt{1-2s}$ will assume *any* arbitrary value and, according to Eq. A.2, equals one. Thus, the MGF of the

non-central χ^2-distribution with one degree of freedom is

$$G_{\chi^2(1,\mu^2)}(s) = \frac{1}{\sqrt{1-2s}} \exp\left\{\frac{s\mu^2}{1-2s}\right\} \,. \tag{A.95}$$

Since the random variables in the sum appearing in the definition A.94 are all *independent* we can use Eq. A.30 to immediately obtain the MGF for a non-central χ^2-distribution with n degrees of freedom:

$$G_{\chi^2(n,\theta)}(s) = \frac{1}{(1-2s)^{n/2}} \exp\left\{\frac{s\theta}{1-2s}\right\} \quad \text{with} \quad \theta = \sum_{j=1}^{n} \mu_j^2 \,. \tag{A.96}$$

The characteristic function, Eq. A.32, of the non-central χ^2-distribution follows again by replacing s by is in Eq. A.96.

$$\Phi_{\chi^2(n,\theta)}(s) = \frac{1}{(1-2is)^{n/2}} \exp\left\{\frac{is\theta}{1-2is}\right\} \quad \text{with} \quad \theta = \sum_{j=1}^{n} \mu_j^2 \,. \tag{A.97}$$

A.5   Transformations Between Distributions

It is possible to transform random variables with a certain distribution into random variables which have *another* distribution. This is particularly useful, for instance, when simulating random walks with the Monte Carlo method. Most *random number generators* generate uniformly distributed random numbers. These can be transformed into random numbers distributed according to a more suitable (for example, normal) distribution function. Two examples of transformations into normally distributed random variables and one transformation into random variables governed by any desired distribution will be introduced below.

A.5.1 Summations

By directly applying the central limit theorem, a sufficient number of independent, identically distributed, random variables are added to obtain an

(approximately) normally distributed random variable:

$$z_i \quad \text{iid, uniformly distributed between 0 and } 1 \Rightarrow \sqrt{\frac{12}{n}} \sum_{i=1}^{n}(z_i - \frac{1}{2}) =: x \sim N(0, 1) \,.$$

(A.98)

According to Eq. A.37, uniformly distributed random variables z_i on $[0, 1]$ have an expectation of $1/2$ and a variance of $1/12$. The random numbers $(Z_i - 1/2)$ are uniformly distributed between $-1/2$ and $+1/2$, thus having zero expectation. However, the variance of these variables remains $1/12$. According to Eq. A.18 the variance of the sum of these variables is

$$\text{var}\left[\sum_{i=1}^{n}(z_i - \frac{1}{2})\right] = \sum_{i=1}^{n} \text{var}[(z_i - \frac{1}{2})] = \sum_{i=1}^{n} \frac{1}{12} = \frac{n}{12} \,.$$

Thus, this sum has zero expectation and a variance of $n/12$, or equivalently a standard deviation of $\sqrt{n/12}$. Dividing the sum by the factor $\sqrt{12/n}$ compensates for this standard deviation. The resulting random variable x in Eq. A.98 has a variance of 1 and is thus in consequence of the central limit theorem (approximately) a *standard* normally distributed random variable.

n uniformly distributed *iid* random numbers are required in order to generate a single, approximately normally distributed random variable. For most applications, it is sufficient to take $n = 12$. Figures A.1 and A.2 show the effect of the transformation for $n = 12$.

Note that the transformed random variables can only take on values between $-\sqrt{3n}$ and $+\sqrt{3n}$. Events lying more than $\sqrt{3n}$ standard deviations from the expectation will never occur when using this transformation. The probability of such an event is so small, however, that it does not play a role in most practical applications in finance. For example, for $n = 12$, the probability is approximately 1 to half a billion:

$$1 - \frac{1}{\sqrt{2\pi}} \int_{-6}^{6} e^{-x^2/2}dx \approx 1,973210^{-9}$$
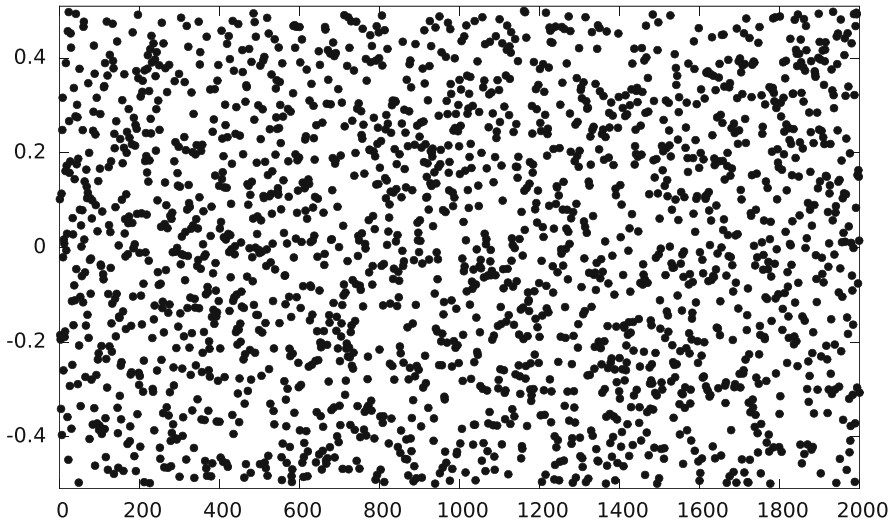
Fig. A.1   2000 *iid* random numbers uniformly distributed between $-1/2$ and $+1/2$
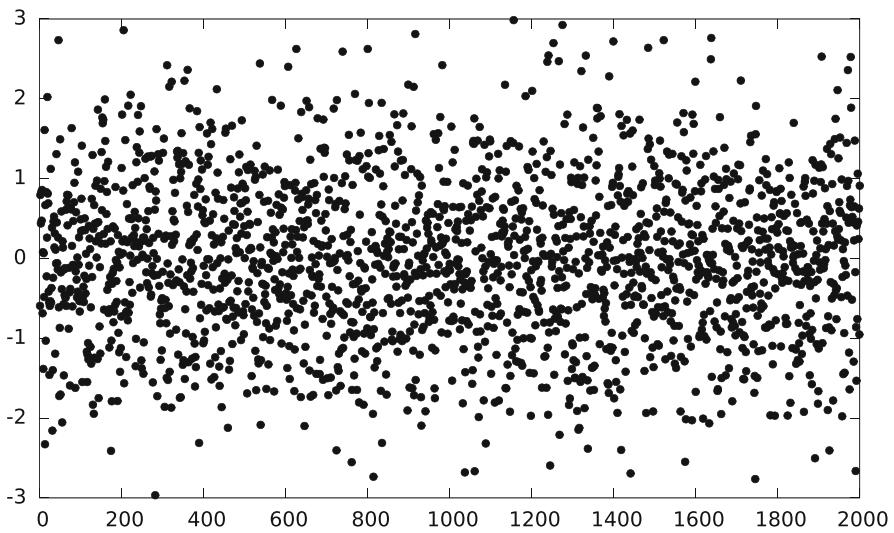
Fig. A.2   2000 approximately normally distributed random numbers, which have been generation according to the transformation in Eq. A.98 with $n = 12$, based on $24,000$ random numbers equally distributed between $-1/2$ and $1/2$ as in Fig. A.1

A.5.2 Box-Muller Transformations

The transformation in Eq. A.98 requires a significant amount of effort, since n random numbers must be generated in order to obtain one single normally distributed random number. The transformation found by *Box* and *Muller* [18] is much more effective. From two independent uniformly distributed random numbers, two independent, normally distributed random variables are generated as follows:

$$x_1 = \sqrt{-2\ln(z_1)}\cos(2\pi\,z_2)\,, \quad x_2 = \sqrt{-2\ln(z_1)}\,\sin(2\pi\,z_2)$$
$$\text{(A.99)}$$

with uniformly distributed z_1, z_2 between 0 and 1 $\Rightarrow$

$$x_1 \sim N(0,1)\,, \quad x_2 \sim N(0,1)\,, \quad \mathrm{cov}\,[x_1, x_2] = 0\,.$$

A.5.3 Inversion of Cumulative Distribution Functions

A very simple method for generating random numbers obeying any desired distribution from a set of random numbers uniformly distributed on the interval between 0 and 1 is to evaluate the inverse cumulative probability function of the desired distribution with the uniformly distributed random numbers as its arguments. Or more precisely: let z be a *uniformly* distributed random variables taking on values in the interval between 0 and 1. Let $f(z)$ be an arbitrary distribution density function with $f(z) > 0 \,\forall z$. Then

$$x := F^{-1}(z) \quad \text{with} \quad F(z) = \int_{-\infty}^{z} f(u)\,du \qquad \text{(A.100)}$$

is distributed according to the distribution associated with the density f. Intuitively, since the uniformly distributed random number z lies between 0 and 1, it can be interpreted as a "probability". The inverse of the cumulative distribution function is then nothing other than the *percentile* of this "probability" z with respect to the desired distribution. The random, uniformly distributed numbers are thus interpreted as probabilities. The percentiles associated with these "probabilities" with respect to the *desired* distribution are then random variables distributed according to the *desired* distribution.

This surprisingly simple method for generating random numbers distributed according to an arbitrary distribution is so important in practice that we will provide the reader with a proof. We begin by proving the existence

of an inverse of the cumulative distribution by showing that the cumulative distribution is strictly increasing if the associated density function is strictly positive:

$$F(x + dx) \equiv \int_{-\infty}^{x+dx} f(u)\, du = \int_{-\infty}^{x} f(u)\, du + \int_{x}^{x+dx} f(u)\, du$$

$$= F(x) + f(x)dx > F(x) \quad \text{since} \quad f(x) > 0 \,.$$

Generally $f(x)dx$ is the probability that a random number governed by a distribution density f will take on a value between x and $x + dx$. This is just the definition of the distribution density, i.e.,

$$f(x)dx = F(x + dx) - F(x)$$

$$= z + dz - z = dz \,,$$

where in the second step, we have made use of the fact that Eq. A.100 implies $F(x) = z$ and, because F is monotone, $F(x + dx) = z + dz$ holds as well. In summary, we may write

$$z = F(x) \iff dz = f(x)dx \,. \tag{A.101}$$

A random variable, uniformly distributed on the interval $[a, b]$, has a constant density function $p(z) = 1/(b - a)$ for all z in the interval $[a, b]$ as was shown in Eq. A.36. In the interval $[0, 1]$ the density of the uniform distribution is thus $p(z) = 1$. The probability that such a uniformly distributed random variable will lie between z and $z + dz$ is therefore simply $p(z)dz = dz$. It follows that Eq. A.101 can be interpreted as follows: if the random variable x is governed by a distribution with density function $f(x)$ then $z = F(x)$ is *uniformly* distributed with density function 1, i.e., uniformly distributed on an interval of length 1. Since equality holds in each step of the above derivation, the conclusion holds in both directions: if z is uniformly distributed on an interval of length 1 (and thus has a distribution density of 1) then $x = F^{-1}(z)$ is distributed according to the density $f(x)$. qed.

   One application of Eq. A.100 is the generation of standard normally distributed random numbers:

$$z \text{ uniformly distributed between 0 and 1} \Rightarrow N^{-1}(z) =: x \sim N(0, 1) \,. \tag{A.102}$$

The cumulative distribution function of the normal distribution has the disadvantage that it and its inverse function cannot be computed analytically, although they can be computed with applications commonly offered in many widely available software packages. For demonstration purposes, however, it is instructive to apply this method to a distribution which *can* be treated analytically. Consider therefore what is known as the *Cauchy distribution.* The Cauchy distribution has a parameter $\lambda > 0$ and the density function

$$f_\lambda(x) = \frac{\lambda}{\pi\left(\lambda^2 + x^2\right)} \ . \tag{A.103}$$

The cumulative distribution function of a Cauchy distributed random variable can be computed analytically with little difficulty. The necessary integral can be found in any collection of mathematical formulas (see for instance [27]):

$$F_\lambda(x) \equiv \int_{-\infty}^x f_\lambda(u)\,du = \frac{\lambda}{\pi} \int_{-\infty}^x \frac{1}{\left(\lambda^2 + u^2\right)}\,du$$

$$= \frac{\lambda}{\pi}\left[\frac{1}{\lambda}\arctan\left(\frac{u}{\lambda}\right)\right]_{u=-\infty}^{u=x} = \frac{1}{\pi}\arctan\left(\frac{x}{\lambda}\right) + \frac{1}{2}\ ,$$

where several well-known properties of the *arctangent function* have been employed. We thus have an analytic expression of the cumulative Cauchy distribution. The inverse function can now be obtained by solving the equation $z = F_\lambda(x)$ for x:

$$F_\lambda(x) = z$$

$$\frac{1}{\pi}\arctan(\frac{x}{\lambda}) + \frac{1}{2} = z$$

$$\arctan(\frac{x}{\lambda}) = \pi\left(z - \frac{1}{2}\right)$$

$$\frac{x}{\lambda} = \tan\left[\pi\left(z - \frac{1}{2}\right)\right]$$

$$x = \lambda \tan\left[\pi z - \frac{\pi}{2}\right]$$

$$= -\lambda \cot(\pi Z)$$

Thus, the inverse of the cumulative distribution function of the Cauchy distribution is

$$F_\lambda^{-1}(z) = x = -\lambda \cot(\pi z)$$

This implies that if z is a uniformly distributed random variable on $[0, 1]$, then $x = -\lambda \cot(\pi z)$ is a Cauchy distributed random variable taking on values between $-\infty$ and ∞ and having a distribution density function as given in Eq. A.103.

References

1. M. Abramowitz, I. Stegun, *Handbook of Mathematical Functions* (Dover Publications, New York, 1972)
2. C. Alexander (ed.), *The Handbook of Risk Management and Analysis* (Wiley, Chichester, 1996)
3. L.B.G. Andersen, R. Brotherton-Ratcliffe, The equity option volatility smile: an implicit finite-difference approach. J. Comput. Finance **1**(2), 5–37 (1998)
4. L.B.G. Andersen, V.V. Piterbarg, *Interest Rate Modeling* (Atlantic Financial Press, New York, London, 2010)
5. N. Anderson, F. Breedon, M. Deacon, et al., *Estimating and Interpreting the Yield Curve* (Wiley, Chichester, 1996)
6. D.F. Babbel, C.B. Merrill, *Valuation of Interest-Sensitive Instruments* (Society of Actuaries, Schaumburg, IL, 1996)
7. L. Bachelier, Théorie de la spéculation. Annales scientifiques de l'École Normale Supérieure **3**(17), 21–86 (1900)
8. Bank for International Settlements, International convergence of capital measurement and capital standards, part 2. http://www.bis.org/publ/bcbs128b.pdf, June 2006
9. J. Baschnagel, W. Paul, *Stochastic Processes – From Physics to Finance* (Springer, Heidelberg, 1996)
10. Basler Ausschuš für Bankenaufsicht, *Richtlinien für das Risikomanagement im Derivativgeschäft.* 7 1994
11. M. Baxter, A. Rennie, *Financial Calculus* (Cambridge University Press, Cambridge, 1997)
12. C.L. Beau, D. Lucas, *Technical Trader's Guide to Computer Analyses of the Futures Market* (Hoppenstedt, 1992)
13. U. Becker, *Lexikon Terminhandel* (Gabler Verlag, Wiesbaden, 1994)

14. P. Best, *Implementing Value at Risk* (Wiley, Chichester, 1998)
15. K. Binder, *Monte Carlo Simulations in Statistical Physics* (Wiley, New York, 1986)
16. J. Bittner, R. Goddard, *Successful Bank Asset-Liability Management* (Wiley, Chichester, 1992)
17. F. Black, P. Karasinski, Bond and option pricing when short rates are lognormal. Financ. Anal. J., 52–59 (1991)
18. G.E.P. Box, M.E. Muller, A note on the generation of random normal deviates. Ann. Math. Stat. **29**, 610–611 (1958)
19. A. Brace, D. Gatarek, M. Musiela, The market model of interest rate dynamics. Math. Financ. **7**(2), 127–147 (1997)
20. E.O. Brigham, *The Fast Fourier Transform* (Prentice-Hall, Englewood Cliffs, NJ, 1974)
21. D. Brigo, F. Mercurio, *Interest Rate Models – Theory and Practice* (Springer Finance, Berlin, Heidelberg, 2007)
22. D. Brigo, M. Morini, A.P. Vicini, *Counterparty Credit Risk, Collateral and Funding* (Wiley Finance, New Jersey, 2013)
23. M. Broadie, P. Glasserman (eds.), *Hedging with Trees* (Risk Publications, London, 1998)
24. O. Brockhaus, M. Farkas, A. Ferraris, D. Long, M. Overhaus, *Equity Derivatives and Market Risk Models* (Risk Books, London, 2000)
25. O. Brockhaus, A. Ferraris, C. Gallus, D. Long, R. Martin, M. Overhaus, *Modelling and Hedging Equity Derivatives* (Risk Books, London, 1999)
26. P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting* (Springer, Heidelberg, 1998)
27. S. Bronstein, *Taschenbuch der Mathematik* (Verlag Harri Deutsch, Frankfurt, 1979)
28. Bundesaufsichtsamt für das Kreditwesen, *Verlautbarung über Mindestanforderungen an das Betreiben von Handelsgeschäften der Kreditinstitute.* Berlin, 10 1995. Geschäftsnummer I4-42-3/86
29. Bundesaufsichtsamt für das Kreditwesen, *Neufassung des Grundsatzes I, Erläuterungen zum Vorentwurf.* Berlin, 5 1996. Geschäftsnummer 17-A223-2/93
30. J.T.M. Byron, *Elements of Simulation* (Chapman & Hall, Boca Raton, FL, 1984)
31. J.P. Calamos, *Convertible Securities* (McGraw-Hill, New York, 1998)
32. J.Y. Campbell, A.W. Lo, A.G. Mackinlay, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, NJ, 1997)
33. G. Cesari, J. Aquilina, N. Charpillon, Z. Filipovic, G. Lee, I. Manda, *Modelling, Pricing, and Hedging Counterparty Credit Exposure: a Technical Guide* (Springer Finance, Berlin, Heidelberg, 2010)
34. A.C. Chiang, *Fundamental Methods of Mathematical Economics*, 4th edn. (McGraw-Hill, New York, 2005)
35. N.A. Chriss, *Black Scholes and Beyond* (McGraw-Hill, New York, 1997)
36. I.J. Clark, *Foreign Exchange Option Pricing* (Wiley Finance, New Jersey, 2011)
37. R. Clarke, Estimating and using volatility: Part 1. Deriv. Q., 40–44 (1994)

38. L. Clewlow, C. Strickland, *Implementing Derivatives Models* (Wiley, Chichester, 1998)

39. F. Collinson, M. Giddings, M. Sullivan, *Financial Products: A Survival Guide* (Euromoney Books, London, 1996)

40. K.B. Connolly, *Pricing Convertible Bonds* (Wiley, Chichester, 1998)

41. E.A. Cornish, R.A. Fisher, Moments and cumulants in the specification of distributions. Rev. Int. Stat. Inst. **5**, 307–320 (1937)

42. J.C. Cox, J.E. Ingersoll, S.A. Ross, A theory of the term structure of interest rates. Econometrica **53**, 385–407 (1985)

43. J.C. Cox, M. Rubinstein, *Options Markets* (Prentice-Hall, Englewood Cliffs, NJ, 1985)

44. L. Culp, M.H. Miller (eds.), *Corporate Hedging in Theory and Practice* (Risk Publications, London, 1999)

45. S. Das, *Swap and Derivative Financing*, revised edition (McGraw-Hill, New York, 1994)

46. S. Das, *Credit Derivatives* (Wiley, New York, 1998)

47. R.E. Dattatreya, F.J. Fabozzi, *Active Total Return Management of Fixed Income Portfolios* (Probus Publishing, Chicago, IL, 1989)

48. W.B. Davenport, *An Introduction to the Theory of Random Signals and Noise* (McGraw-Hill Books, New York, 1958)

49. H.-P. Deutsch, *Quantitative Portfoliosteuerung* (Schäffer-Poeschel Verlag, Stuttgart, 2005)

50. H.-P. Deutsch, M. Beinker, Downloads for all example spreadsheets. https://doi.org/10.1007/978-3-030-22899-6_1

51. A.K. Dixit, R.S. Pindick, *Investment Under Uncertainty* (Princeton University Press, Princeton, NJ, 1994)

52. M. Doi, S.F. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, Oxford, 1986)

53. M. Dothan, *Prices in Financial Markets* (Oxford University Press, Oxford, 1990)

54. K. Dowd, *Beyond Value-at-Risk* (Wiley, Chichester, 1990)

55. D. Duffie, *Security Markets – Stochastic Models* (Academic Press, San Diego, 1988)

56. D. Duffie, *Dynamic Asset Pricing Theory*, 2nd edn. (Princeton University Press, Princeton, NJ, 1996)

57. C. Dunis, B. Zhon, *Nonlinear Modelling of High Frequency Financial Time Series* (Wiley, Chichester, 1996)

58. B. Dupire, Pricing with a smile. Risk **7**(18–20), (1994)

59. B. Dupire, *Monte Carlo – Methodologies and Applications* (Risk Books, London, 1998)

60. R. Eller (ed.), *Handbuch Derivativer Instrumente* (Schäffer-Poeschel Verlag, Stuttgart, 1996)

61. R. Eller (ed.), *Handbuch des Risikomanagements* (Schäffer-Poeschel Verlag, Stuttgart, 1996)

62. R. Eller, W. Gruber, M. Reif (eds.), *Handbuch Bankenaufsicht und interne Risikosteuerungsmodelle* (Schäffer-Poeschel Verlag, Stuttgart, 1998)

63. E. Elton, M. Gruber, *Modern Portfolio Theory and Investment Analysis*, 5th edn. (Wiley, Chichester, 1995)

64. P. Embrechts, C. Klüppelberg, T. Mikosch, *Modelling Extremal Events for Insurance and Finance* (Springer, Heidelberg, 1999)

65. F.J. Fabozzi, *Valuation of Fixed Income Securities and Derivatives*, 3rd edn. (Fabozzi Associates, New Hope, PA, 1998)

66. F.J. Fabozzi, *The Handbook of Fixed Income Securities*, 5th edn. (McGraw-Hill, New York, 1999)

67. Feller, *Probability Theory and Its Applications* (Wiley, New York, 1950)

68. K. Ferris, M. Jones, *The Reuters Guide to Official Interest Rates*, 2nd edn. (Probus Publishing, Chicago, IL, 1995)

69. G.S. Fishman, *Monte Carlo Concepts, Algorithms and Applications* (Springer, Heidelberg, 1996)

70. B. Flury, *A First Course in Multivariate Statistics* (Springer, Heidelberg, 1996)

71. J.C. Francis, J.A. Frost, J. Gregg Whittaker, *The Handbook of Credit Derivatives* (McGraw-Hill, New York, 1998)

72. C.W. Gardiner, *Handbook of Stochastic Methods*, 2nd edn. (Springer, Heidelberg, 1985)

73. M.J. Gardner, D.L. Mills, *Managing Financial Institutions*, 3rd edn. (Dryden, Fort Worth, 1994)

74. N. Gershenfeld, *The Nature of Mathematical Modelling* (Cambridge University Press, Cambridge, 1994)

75. R. Gibson (ed.), *Model Risk – Concepts, Calibration and Pricing* (Risk Publications, London, 1999)

76. M. Gilli, M. Grŏse, E. Schumann, Calibrating the nelson-siegel-svensson model. Comisef working papers series, COMISEF, March 2010

77. B. Graham, D.L. Dodd, *Security Analysis*, 6th edn. (McGraw-Hill Professional, New York, 2008)

78. W.H. Greene, *Econometric Analysis*, 3rd edn. (Prentice-Hall, Englewood Cliffs, NJ, 1997)

79. J. Gregory, *Counterparty Credit Risk and Credit Value Adjustment*, 2nd edn. (Wiley Finance, New Jersey, 2012)

80. Grill, Gramlich, Eller, *Gabler Bank Lexikon*, 11th edn. (Gabler Verlag, Wiesbaden, 1995)

81. G.R. Grimmett, D.R. Stirzaker, *Probability and Random Processes*, 2nd edn. (Oxford University Press, Oxford, 1998)

82. R. Grinold, R. Kahn, *Active Portfolio Management* (Probus Publishing, Chicago, IL, 1995)

83. Group of Thirty, *Derivatives, Practices and Principles* (1993)

84. W. Gruber, L. Overbeck, Working paper. Deutsche Bundesbank, 1996

85. O. Gwilym, C. Sutcliffe (eds.), *High-Frequency Financial Market Data* (Risk Publications, London, 1999)

86. J.D. Hamilton, *Time Series Analysis* (Princeton University Press, Princeton, NJ, 1994)

87. D. Hammer, *Dynamic Asset Allocation* (Wiley, Chichester, 1994)

88. L. Hanenberg, Zur Verlautbarung über Mindestanforderungen an das Betreiben von Handelsgeschäften der Kreditinstitute des Bundesaufsichtsamtes für das Kreditwesen. Die Wirtschaftsprüfung **49**, 637–648 (1996)

89. M. Harrison, S. Pliska, Martingales and stochastic integrals in the theory of continuous trading. Stoch. Process. Appl. **11**, 215–260 (1981)

90. G. Haug, *The Complete Guide to Option Pricing Formulas* (McGraw-Hill, New York, 1997)

91. R.A. Haugen, *Modern Investment Theory*, 5th edn. (Prentice-Hall, Englewood Cliffs, NJ, 2000)

92. D. Heath, R. Jarrow, A. Morton, Bond pricing and the term structure of interest rates: A new methodology for contingent claim valuation. Econometrica **60**(1), 77–105 (1992)

93. T. Heidorn, *Vom Zins zur Option – Finanzmathematik in der Bankpraxis* (Gabler Verlag, Wiesbaden, 1994)

94. D.E. Hess, *Die Dynamik der Zinsstruktur* (Gabler Edition Wissenschaft, Wiesbaden, 1995)

95. A. Hicks, *Foreign Exchange Options*, 2nd edn. (Woodhead Publishing, Sawston, 1995)

96. R.C. Higgins, *Analysis for Financial Management*, 5th edn. (McGraw-Hill, Irwin, 1998)

97. S.Y. Ho, S.-B. Lee, Term structure movements and pricing interest rate contingent claims. J. Financ. **41**, 1011–1029 (1986)

98. T. Ho, *Fixed Income Solutions* (Irwin Professional Publishing, Chicago, 1996)

99. http://www.cdsmodel.com

100. L. Hughston (ed.), *Vasicek and Beyond: Approaches to Building and Applying IR-Models* (Risk Publications/Euro Brokers, London, 1996)

101. L. Hughston (ed.), *Options* (Risk Publications, London, 1999)

102. L. Hughston (ed.), *The New Interest Rate Models: Recent Developments in the Theory and Application of Yield Curve Dynamics* (Risk Books, London, 2000)

103. J.C. Hull, *Options, Futures and Other Derivatives*, 8th edn. (Prentice-Hall, Englewood Cliffs, NJ, 2011)

104. J.C. Hull, A. White, Pricing interest rate derivative securities. Rev. Financ. Stud. **3**(4), 573–592 (1990)

105. J.C. Hull, A. White, *Hull-White on Derivatives* (Risk Publications, London, 1996)

106. P.J. Hunt, J.E. Kennedy, *Financial Derivatives in Theory and Practice* (Wiley, Chichester, 2000)

107. E. Ingersoll Jr., E. Jonathan, *Theory of Financial Decision Making* (Rowman & Littlefield, Bollman Place, 1987)
108. J. James, N. Webber, *Interest Rate Modelling* (Wiley, Chichester, 2000)
109. F. Jamshidian, Options and futures evaluation with deterministic volatility. Math. Financ. **3**(2), 149–159 (1993)
110. R. Jarrow (ed.), *Over the Rainbow* (Risk Publications, London, 1995)
111. R. Jarrow, *Modelling Fixed Income Securities and Interest Rate Options* (McGraw-Hill, New York, 1996)
112. R. Jarrow (ed.), *Volatility – New Estimation Techniques for Pricing Derivatives* (Risk Publications, London, 1998)
113. R. Jarrow, S. Turnball, *Derivative Securities* (Southwestern Publishers, Cincinnati, 1996)
114. N.L. Johnson, S. Kotz, N. Balakrishnan, *Continuous Univariate Distributions* (Wiley, Chichester, 1994)
115. R.A. Johnson, D.W. Wichern, *Applied Multivariate Statistical Analysis*, 4th edn. (Prentice-Hall, Englewood Cliffs, NJ, 1999)
116. I.T. Jollife, *Principal Component Analysis* (Springer, Heidelberg, 1986)
117. D.W. Jordan, P. Smith, *Mathematical Techniques*, 2nd edn. (Oxford University Press, Oxford, 1998)
118. P. Jorion, *Value At Risk: The New Benchmark for Controlling Derivatives Risk* (McGraw-Hill, New York, 1997)
119. M.H. Kalos, P.A. Whitlock, *Monte Carlo Methods* (Wiley, New York, 1986)
120. I. Karatzas, S.E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd edn. (Springer, Heidelberg, 1999)
121. H.M. Karlin, *A First Course in Stochastic Processes* (Academic Press, New York, 1975)
122. R.A. Klein, J. Lederman, *Derivatives Risk and Responsibility* (Irwin Professional Publishing, Chicago, 1996)
123. P.E. Kloeden, E. Platen, H. Schurz, *Numerical Solutions of SDE Through Computer Experiments* (Springer, Heidelberg, 1994)
124. M. Kofler, *VBA-Programmierung mit Excel 97* (Addison-Wesley, Boston, 1997)
125. R.A. Korajzykk (ed.), *Asset Pricing and Portfolio Performance* (Risk Publications, London, 1999)
126. C.J.C. Kruger, Constrained cubic spline interpolation, August 2002. http://www.korf.co.uk/spline.pdf
127. W.J. Krzanowski, *Principles of Multivariate Analysis*, reprinted edition (Oxford Science Publications, Oxford, 1998)
128. Y.K. Kwok, *Mathematical Models of Financial Derivatives* (Springer, Heidelberg, 1998)
129. D. Lamberton, B. Lapeyere, *Introduction to Stochastic Calculus Applied to Finance* (Chapman & Hall, Boca Raton, FL, 1996)
130. M. Leibowitz, S. Kogelman, T. Klaffky, *A Shortfall Approach to Duration Management* (Salomon Brothers, New York, 1990)

131. A.L. Lewis, *Option Valuation Under Stochastic Volatility* (Financepress, New Jersey, 2000)

132. E. Lewis, *An Introduction to Credit Scoring*, reprint edition (Fair, Isaac & Co., San Jose, 1994)

133. O. Loistl, *Kapitalmarkttheorie* (Oldenbourg-Verlag, München, 1994)

134. J. Longerstaey, *Introduction to RiskMetrics* (Morgan Guaranty Trust Company, New York, 1998)

135. F.A. Longstaff, E.S. Schwartz, Valuing American options by simulation: A simple least-square approach. Rev. Financ. Stud. **14**(1), 113–147 (2001)

136. G. Macey, Pricing with standard CSA defined by currency buckets. Technical report, SSRN eLibrary, 2011

137. R.N. Mantegna, H.E. Stanley, *An Introduction to Econophysics* (Cambridge University Press, Cambridge, 1998)

138. H.M. Markovitz, *Mean Variance* (Blackwell, Hoboken, 1987)

139. H.M. Markovitz, *Portfolio Selection*, 2nd edn. (Blackwell, Hoboken, 1998)

140. C. Matten, *Managing Bank Capital* (Wiley, Chichester, 1996)

141. R.C. Merton, *Continuous Time Finance*, revised edition (Blackwell, Malden, MA, 1998)

142. T.C. Mills, *The Econometric Modelling of Financial Time Series* (Cambridge University Press, Cambridge, 1997)

143. P. Miron, P. Swannel, *Pricing and Hedging of Swaps* (Euromoney Books, London, 1991)

144. J. Montier, *Behavioural Finance* (Wiley Finance, New Jersey, 2002)

145. M. Musiela, M. Rutkowski, *Martingale Methods in Financial Modelling* (Springer, Heidelberg, 1997)

146. S.N. Neftci, *An Introduction to the Mathematics of Financial Derivatives*, 2nd edn. (Academic Press, San Diego, 2000)

147. I. Nelken, *The Handbook of Exotic Options* (Probus Publications, Chicago, IL, 1995)

148. C.R. Nelson, A.F. Siegel, Parsimonious modeling of yield curves. J. Bus. **60**(4), 473–89 (1987). University of Chicago Press

149. H. Niederreiter, *Random Number Generation and Quasi Monte Carlo Methods* (SIAM, Philadelphia, 1992)

150. L.T. Nielsen, *Pricing and Hedging of Derivative Securities* (Oxford University Press, Oxford, 1999)

151. B. Oksendal, *Stochastic Differential Equations*, 5th edn. (Springer, Heidelberg, 1998)

152. J.P. O'Shaughnessy, *What Works on Wall Street*, 4th edn. (McGraw-Hill Professional, New York, 2011)

153. H.H. Panjer (ed.), *Financial Economics with Applications to Investments, Insurance and Pensions* (The Actuarial Foundation, Schaumburg, IL, 1998)

154. R.S. Pindyck, D.L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th edn. (McGraw-Hill, New York, 1997)

155. V. Piterbarg, Cooking with collateral. Risk, 58–63 (2012)

156. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in C – The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1994)

157. P.S. Prudenzi, *VBA mit Excel 97* (Addison-Wesley, Boston, 1997)

158. G.S. Questa, *Fixed Income Analysis for the Global Financial Market* (Wiley, Chichester, 1999)

159. R. Rebonato, *Interest-Rate Option Models* (Wiley, Chichester, 1996)

160. R. Rebonato, *Volatility & Correlation* (Wiley, Chichester, 1999)

161. R. Rebonato, K. McKay, R. White, *The SABR/LIBOR Market Model* (Wiley, Chichester, 2009)

162. J.A. Rice, *Mathematical Statistics and Data Analysis*, 2nd edn. (Duxbury Press, Massachusetts, 1999)

163. B.D. Ripley, *Stochastic Simulation* (Wiley, New York, 1987)

164. RiskMetrics, *Technical Document*, 4th edn. (JP Morgan Global Research, New York, 1996)

165. B. Rönz, E. Förster, *Regressions- und Korrelationsanalyse, Grundlagen, Methoden, Beispiele* (Gabler Verlag, Wiesbaden, 1992)

166. C. Rouvinez, Going Greek with var. RISK Mag. **10**(2), 11 (1994)

167. M. Rubinstein, Return to oz. RISK Mag. **10**(11), (1994)

168. M. Rubinstein, *Rubinstein on Derivatives* (Risk Publications, London, 1999)

169. R.Y. Rubinstein, *Simulation and the Monte Carlo Method* (Wiley, Chichester, 1981)

170. H. Schierenbeck, *Bank- Und Versicherungslexikon* (Oldenbourg Verlag, München, 1990)

171. H. Schierenbeck, *Ertragsorientiertes Bankmanagement*, 5th edn. (Gabler Verlag, Wiesbaden, 1997)

172. P.J. Schönbucher, *Credit Derivatives Pricing Models* (Wiley Finance, New Jersey, 2003)

173. W.F. Sharpe, G.J. Alexander, J.V. Bailey, *Investments*, 5th edn. (Prentice-Hall, Englewood Cliffs, NJ, 1995)

174. C.D. Smith, *Option Strategies* (Wiley, New York, 1996)

175. C.W. Smith, C.W. Smithson, *The Handbook of Financial Engineering, New Financial Product Innovations, Applications, and Analyses* (Harper Business Books, New York, 1990)

176. G.D. Smith, *Numerical Solution of Partial Differential Equations*, 3rd edn. (Oxford University Press, Oxford, 1998)

177. L.E.O. Svensson, Estimating and interpreting forward interest rates. Sweden 1992–1994. Papers, *Stockholm - International Economic Studies*, 1994

178. N. Taleb, *Dynamic Hedging* (Wiley, Chichester, 1998)

179. J.M. Tavakoli, *Credit Derivatives* (Wiley, Chichester, 1998)

180. D. Tavella, C. Randall, *Pricing Financial Instruments: The Finite Difference Method* (Wiley, Chichester, 2000)

181. J.W. Thomas, *Numerical Partial Differential Equations* (Springer, Heidelberg, 1995)

182. J.I. Tiner, J.M. Conneely, *Accounting for Treasury Products*, 2nd edn. (Woodhead-Faulkner, Sawston, 1989)

183. R. Tomkins, *Options Explained* (MacMillan Business, London, 1994)

184. H. Tong, *Non-linear Time Series*, reprinted edition (Oxford Science Publications, Oxford, 1999)

185. D.G. Uyemura, D.R. van Deventer, *Financial Risk Management in Banking* (McGraw-Hill, New York, 1993)

186. O.A. Vasicek, An equilibrium characterization of the term structure. J. Financ. **5**, 177–188 (1977)

187. J. Walmsley, *The New Financial Instruments* (Wiley, Chichester, 1988)

188. J. Welcker, J.W. Kloy, K. Schindler, *Professionelles Optionsgeschäft* (MI Verlag, Zürich, 1992)

189. R. White, The pricing and risk management of credit default swaps, with a focus on the isda model. OpenGamma Quantitative Research (2013)

190. D. Williams, *Probability with Martingales* (Cambridge Mathematical Textbooks, Cambridge, 1998)

191. P.J. Wilmott, *Derivatives – The Theory and Practice of Financial Engineering* (Wiley, Chichester, 1998)

192. P.J. Wilmott, J. Dewynne, S. Howison, *Option Pricing – Mathematical Models and Computation* (Oxford Financial Press, Oxford, 1997)

193. P.J. Wilmott, J. Dewynne, S. Howison, *The Mathematics of Financial Derivatives* (Cambridge University Press, Cambridge, 1998)

194. T.C. Wilson, Calculating risk capital, in *Risk Management and Analysis*, ed. by C. Alexander (Wiley, Chichester, 1997)

195. U. Wystup, *FX Options and Structured Products* (Wiley Finance, New Jersey, 2006)

196. P. Zangari, *A VaR Methodology for Portfolios that Include Options* (RiskMetrics Monitor, 1st quarter 1996)

197. P.G. Zhang, *Exotic Options: A Guide to Second Generation Options* (World Scientific, Singapore, 1997)

198. H. Zimmermann, S. Jaeger, Z. Staub, *Asset- und Liability-Management*, 2nd edn. (Verlag Neue Zürcher Zeitung, Zürich, 1996)

Index

30/360, 9, 68
30E/360, 9

Accounting currency, 719
Accrued interest, 85, 369
Act/360, 68
Act/365, 68, 416
Act/Act, 9, 68
Active, 646
Active cash, 647
Active risk
    marginal, 647
Actual/360, 9
Actual/365f, 8
Add-on, 582
Address
    reference, 482
ADP, 318
Aging, 576
Akaike information criterion, 817
Algorithm
    genetic, 688
Alpha, 642, 648
American exercise, 223

Analysis
    fundamental, 618
Annualization, 68
    of volatilities, 511
Annualized, 68
Annuity, 361, 385
    as numeraire, 385
Annuity loan, 401
Approximation, 90
AR(1), 761
AR(p), 757
Arbitrage, 21, 97
Arbitrage free
    market, 257
    short rate model, 310
Arctangent, 869
Arrow-Debreu price, 318
    instantaneous, 319
Asian option, 461
    American, 467
    average price, 443
    average strike, 443
Asian tail, 444
Asset, 588
    management, 588