

Springer Texts in Business and Economics

Gerhard Larcher

The Art of Quantitative Finance Vol.2

Volatilities, Stochastic Analysis and
Valuation Tools



Springer

Springer Texts in Business and Economics

Springer Texts in Business and Economics (STBE) delivers high-quality instructional content for undergraduates and graduates in all areas of Business/Management Science and Economics. The series is comprised of self-contained books with a broad and comprehensive coverage that are suitable for class as well as for individual self-study. All texts are authored by established experts in their fields and offer a solid methodological background, often accompanied by problems and exercises.

Gerhard Larcher

The Art of Quantitative Finance Vol.2

Volatilities, Stochastic Analysis and
Valuation Tools



Springer

Gerhard Larcher
Institute for Financial Mathematics and
Applied Number Theory
Johannes Kepler University of Linz
Linz, Austria

ISSN 2192-4333 ISSN 2192-4341 (electronic)
Springer Texts in Business and Economics
ISBN 978-3-031-23869-7 ISBN 978-3-031-23870-3 (eBook)
<https://doi.org/10.1007/978-3-031-23870-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface to Volume II

Volatility of markets is one of the most important concepts in quantitative finance. We will start this Volume II with an intensive investigation of various approaches to volatility. Next, we will extend our knowledge on derivative pricing and apply the new techniques for the valuation of more complex types of derivatives. Further, we will give an intuitive introduction to the powerful tools of stochastic analysis and we will apply these tools, for example, to the valuation of interest rate derivatives.

About This Book

This book was translated by Nina Sattler-Hovdar, based on the highly successful German book “Quantitative Finance. Strategien, Investitions, Analysen” by Springer Gabler (2020).

Contents

1	Volatilities	1
1.1	Volatility I: Historical Volatility.....	1
1.2	Volatility II: ARCH Models	9
1.3	Volatility III: How to Use and Forecast Volatility	14
1.4	Volatility IV: Magnitude of the Historical Volatility of the S&P500	21
1.5	Volatility V: Volatility in Derivatives Pricing.....	22
1.6	Derivatives Pricing with Time- (and Price-) Dependent Volatility: the Dupire Model	25
1.7	Implied Volatility	30
1.8	Implied Volatilities of Call Options and Put Options with Same Expiration and Strike	37
1.9	Volatility Skews, Volatility Smiles, and Volatility Surfaces	38
1.10	Inferences From Implied Volatilities About the Market-Anticipated Distribution of the Underlying Asset's Price	42
1.11	Volatility Indices	50
1.12	Basic Properties of the VIX	54
1.13	Relation and Correlations Between VIX and SPX	58
1.14	Influence of Price- and/or Time-Dependent Volatility on Delta, Gamma, and Theta	66
1.15	Combined Trading of SPX and VIX for Hedging Purposes	70
1.16	Relation and Correlations of VIX with Historical and Realized Volatility	78
1.17	The CBOE S&P500 Put Write Index	84
1.18	The VIX Calculation Methodology	86
1.19	The Volatility Weekend Effect	93
1.20	Derivatives on the VIX: VIX Futures	96
1.21	VIX Options	106
1.22	Payoff and Profit Functions of a Trading Strategy for Combinations of SPX and VIX Options	109
	References	117

2 Extensions of the Black-Scholes Theory to Other Types of Options (Futures Options, Currency Options, American Options, Path-Dependent Options, Multi-asset Options).....	119
2.1 Introduction and Discussions So Far	119
2.2 Currency Options	121
2.3 Futures Options	126
2.4 Valuation of American Options and of Bermudan Options Through Backwardation (the Algorithm).....	131
2.5 Valuation Examples for American Options in the Binomial and in the Wiener Model	138
2.6 Hedging American Options	142
2.7 Path-Dependent (Exotic) Derivatives, Definition and Examples	144
2.8 Valuation of Path-Dependent Options, the Black-Scholes Formula for Path-Dependent Options.....	150
2.9 Numerical Valuation Example of a Path-Dependent Option in a Three-Step Binomial Model (European and American)	155
2.10 The Complexity of Pricing Path-Dependent Options in an N-Step Binomial Model in General and, For Example, for Lookback Options	157
2.11 Valuation of an American Lookback Option in a Four-Step Binomial Model (Numerical Example).....	163
2.12 Explicit Formulas for European Path-Dependent Options, For Example, Barrier Options	167
2.13 Explicit Formulas for European Path-Dependent Options, For Example, Geometric Asian Options	170
2.14 Brief Comment on Hedging Path-Dependent Derivatives	178
2.15 Valuation of Derivatives Using Monte Carlo Methods, Basic Principle	179
2.16 Valuation of European Path-Dependent Derivatives with Monte Carlo Methods	183
2.17 Monte Carlo Valuation of Asian Options	184
2.18 Monte Carlo Valuation of Barrier Options	185
2.19 Barrier Options in Turbo and Bonus Certificates	189
2.20 Estimating Greeks (Especially Delta and Gamma) of Derivatives with Monte Carlo	192
2.21 Estimating Delta and Delta Hedging for Path-Dependent Derivatives (e.g. Geometric Asian Option)	200
2.22 Some Fundamental Remarks on Monte Carlo Methods and on the Convergence of Monte Carlo Methods.....	207
2.23 Some Remarks on Random Numbers.....	212
2.24 A Remark on Quasi-Monte Carlo Methods	217
2.25 An Example of Low-Discrepancy QMC Point Sets: The Hammersley Point Sets	226
2.26 Variance Reduction Methods for Monte Carlo	229

2.27	Using Monte Carlo with Control Variates to Value an Arithmetic Asian Option	233
2.28	Multi-asset Options	236
2.29	Modelling Correlated Financial Products in the Wiener Model, Cholesky Decomposition	238
2.30	Valuation of Multi-asset Options.....	242
2.31	Example of Pricing a Multi-asset Option with MC and with QMC	245
	References	250
3	Fundamentals: Stochastic Analysis and Applications, Interest Rate Dynamics, and Basic Principles of Pricing Interest Rate Derivatives	251
3.1	Modelling of Interest Rate Dynamics.....	252
3.2	Differential Representation of Stochastic Processes: Heuristic Introduction	254
3.3	Simulation of Ito Processes and Basic Models	266
3.4	Excursus: The Ito Formula and Differential Notation of the GBM	275
3.5	Interest Rate Modelling Using Mean-Reverting Ornstein-Uhlenbeck	279
3.6	Examples of Interest Rate Derivatives and a Principal Methodology for Pricing such Derivatives	285
3.7	Basic Concepts of Frictionless Interest Rate Markets: Zero-Coupon Bonds and Interest Rates.....	287
3.8	Fixed and Floating Rate Coupon Bonds	292
3.9	Interest Rate Swaps	295
3.10	Valuation of Bond Prices and Interest Rate Derivatives in a Short-Rate Approach	296
3.11	The Mean-Reverting Vasicek Model and the Hull-White Model for the Short Rate	303
3.12	Affine Model Structures of Bond Prices	304
3.13	Bond Prices in the Vasicek Model and Calibration in the Vasicek Model	306
3.14	Bond Prices in the Hull-White Model and Calibration in the Hull-White Model	308
3.15	Valuation and Put-Call Parity of Call and Put Options on Bond Prices	311
3.16	Valuation of Caplets and Floorlets (as Well as Interest Rate Caps and Interest Rate Floors)	312
3.17	The Black-Scholes Differential Equation.....	316
3.18	The Stochastic Ito Integral: Heuristic Explanation and Basic Properties.....	319
3.19	Conditional Expectations and Martingales	330
3.20	The Feynman-Kac Formula	333

3.21	The Black-Scholes Formula	337
3.22	The Black-Scholes Model as a Complete Market and Hedging of Derivatives	338
3.23	The Multidimensional Black-Scholes Model and Its Completeness	341
3.24	Incomplete Markets (e.g. the Trinomial Model)	342
3.25	Incomplete Markets (e.g. Non-tradable Underlying Asset).....	349
	References	353



Keywords

Historical volatility · ARCH models · Dupire model · Implied volatility · Volatility indices · VIX · Volatility derivatives

1.1 Volatility I: Historical Volatility

For some time now, we have been pointing out the need to discuss in detail the concept of volatility and the volatility parameter in pricing formulas for derivatives. We are going to open this discussion now. As so often before, however, we do so with a caveat: A thorough and in-depth analysis of the concept of volatility requires intensive study and advanced mathematical techniques. We will have to confine ourselves once again to the strictly necessary concepts and properties. Our main goal in this is to help you gain sufficient understanding of the concept of volatility and of the main techniques that you need to be able to run basic applications in an exact manner, using well-founded methods (all while knowing, of course, that more advanced methods and models might potentially provide even deeper insights into the respective application).

Volatility—which, in its broadest sense, is defined as a measure of how strongly a financial asset's price fluctuates—is a key metric used in risk management and portfolio management as well as for pricing derivatives, for classification purposes within investment fund valuation systems, and for classification and definition of risk classes in investment fund management.

In principle, put simply and inexactly, volatility of a financial asset's price expresses the fluctuation of the (continuous or discrete) returns of that asset's price (which, for simplicity, we will refer to as the “stock price” in the following).

Yet this simplified “definition” of volatility is severely flawed, as we will see in the following:

First of all, we will encounter widely different types of volatilities, some of which will be far removed from the concept of volatility as conveyed by the above definition. And secondly, even if we accept that concept, we will still be dealing with different types of volatilities depending on the time periods they relate to.

Let us start with the first basic concept, the **historical volatility** of a price, and revisit a few basics from Volume I Chapter 4.1.

Once again we will start with a time range $[0, T]$, which we divide into N equal parts of length dt . (All time data are given in years.) The values A_0, A_1, \dots, A_N denote the prices of a stock at times $0, dt, 2 \cdot dt, \dots, N \cdot dt$.

First, we calculate the returns a_i of the price movements for the time periods $i \cdot dt$ to $(i + 1) \cdot dt$ and for $i = 0, 1, \dots, N - 1$.

All of the following could be carried out for both discrete and continuous returns. We will focus on continuous returns here, thus: $a_i = \log\left(\frac{A_{i+1}}{A_i}\right)$

We calculate the mean over the returns a_i in the observed time range

$$\mu'_A = \frac{1}{N} \sum_{i=0}^{N-1} a_i.$$

μ'_A is the stock's (historical) trend in the time range $[0, T]$ per dt .

We use the unwieldier notation μ'_A instead of μ_A for the trend per dt . The simpler notation μ_A is reserved for the normalized variable $\frac{\mu'_A}{dt}$, i.e. the annualized trend of A over $[0, T]$.

The standard deviation σ'_A of the returns measures the square root of the mean square deviation of the returns from the trend, i.e.

$$\sigma'_A = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (a_i - \mu'_A)^2}$$

We refer to the standard deviation σ'_A as the “**historical volatility of the stock A per time unit dt in the time range $[0, T]$** ”.

Again, we are using the somewhat unwieldier notation σ'_A instead of σ_A to denote the volatility per dt . The simpler notation σ_A is reserved for the normalized variable $\frac{\sigma'_A}{\sqrt{dt}}$, i.e. the annualized volatility of A on $[0, T]$.

Volatility is often not expressed in absolute terms, but rather as a percentage, i.e. multiplied by 100. We often hear statements along the following lines to describe volatility (which may hold more or less true but is definitely very illustrative):

“A financial asset with a trend of $x\%$ p.a. and a volatility of $y\%$ p.a. suggests an average annual return of $x\%$ at an average deviation of $y\%$, so that an annual return between $x - y\%$ and $x + y\%$ can usually be expected”.

Assuming **normally distributed** returns and volatility of σ'_A per dt , you can expect, with probability of

68.27% in a time step of **length dt** , to get a return in the range of

$$[\mu'_A - \sigma'_A, \mu'_A + \sigma'_A]$$

95.45% in a time step of length dt , to get a return in the range of

$$[\mu'_A - 2\sigma'_A, \mu'_A + 2\sigma'_A]$$

and of

99.73% in a time step of length dt , to get a return in the range of

$$[\mu'_A - 3\sigma'_A, \mu'_A + 3\sigma'_A].$$

Assuming **independent and normally distributed** returns and volatility of σ'_A per dt , you can expect, with probability of

68.27% in a time step of **length $T = N \cdot dt$** , to get a return in the range of

$$\left[\begin{array}{l} \mu'_A \cdot \frac{T}{dt} - \sigma'_A \cdot \sqrt{\frac{T}{dt}}, \mu'_A \cdot \frac{T}{dt} + \sigma'_A \sqrt{\frac{T}{dt}} \\ \mu_A \cdot T - \sigma_A \cdot \sqrt{T}, \mu_A \cdot T + \sigma_A \cdot \sqrt{T} \end{array} \right] =$$

95.45% in a time step of length $T = N \cdot dt$, to get a return in the range of

$$\left[\begin{array}{l} \mu'_A \cdot \frac{T}{dt} - 2\sigma'_A \cdot \sqrt{\frac{T}{dt}}, \mu'_A \cdot \frac{T}{dt} + 2\sigma'_A \sqrt{\frac{T}{dt}} \\ \mu_A \cdot T - 2\sigma_A \cdot \sqrt{T}, \mu_A \cdot T + 2\sigma_A \cdot \sqrt{T} \end{array} \right] =$$

and of 99.73% in a time step of length $T = N \cdot dt$, to get a return in the range of

$$\left[\begin{array}{l} \mu'_A \cdot \frac{T}{dt} - 3\sigma'_A \cdot \sqrt{\frac{T}{dt}}, \mu'_A \cdot \frac{T}{dt} + 3\sigma'_A \sqrt{\frac{T}{dt}} \\ \mu_A \cdot T - 3\sigma_A \cdot \sqrt{T}, \mu_A \cdot T + 3\sigma_A \cdot \sqrt{T} \end{array} \right] =$$

This simply follows from the fact that the return a on the interval $[0, T]$ is given by $a = a_0 + a_1 + \dots + a_{N-1}$, that is, as the sum of N independent normally distributed random variables with mean μ'_A and volatility σ'_A , and is therefore normally distributed with mean $N \cdot \mu'_A = \mu'_A \cdot \frac{T}{dt}$ and volatility $\sqrt{N} \cdot \sigma'_A = \sigma'_A \cdot \sqrt{\frac{T}{dt}}$.

However, as we know, stock prices generally do not really follow a normal distribution and do not generally move in complete independence. It is therefore not necessarily true that the per annum volatility for any dt results from $\sigma_A = \sigma'_A \cdot \sqrt{\frac{1}{dt}}$.

Or, put differently, for different time ranges dt , the “normalization” $\sigma'_A \cdot \sqrt{\frac{1}{dt}}$ does not always give the same value σ_A , i.e. not always a well-defined annualized volatility.

To illustrate this with an example, we calculated and annualized the continuous returns, trends, and volatilities on a daily, weekly, monthly, and annual basis from the daily, weekly, monthly, and annual closing prices of the S&P500 from 1957 to 2017, of the DAX from 1988 to 2017, of the EuroStoxx50 from 1992 to 2017, and of the IBM stock from 1982 to 2017 (as from October in each case).

Table 1.1 Annualized trends and volatilities of S&P500, DAX, EuroStoxx50, and IBM

	Trend p.a.	Vol p.a. on daily basis	Vol p.a. on weekly basis	Vol p.a. on monthly basis	Vol p.a. on annual basis
S&P500	6.49%	15.74%	15.43%	15.23%	14.21%
DAX	8.51%	22.38%	22.88%	21.35%	23.20%
EuroStoxx50	5.04%	21.76%	20.81%	20.03%	21.31%
IBM	6.26%	26.89%	26.02%	25.70%	25.33%

This gave us the following Table 1.1 with the annualized trends and annualized volatilities: The results for the trend per annum are independent of the chosen time period. The results for the annualized volatilities show slightly different values, depending on which period lengths were chosen. Volatilities tend to fall the longer the chosen period is.

The statement made further above along the lines that:

“Assuming **independent and normally distributed** returns and (unchanged) trend μ_A and volatility σ_A per annum, one can expect, with a probability of 68.27% in a time step of **length $T = N \cdot dt$** , to get a return in the range of $[\mu_A \cdot T - \sigma_A \cdot \sqrt{T}, \mu_A \cdot T + \sigma_A \cdot \sqrt{T}]$ ”, implies that, for any time 0 and stock price S_0 , we can expect to see, with the same probability, a stock price in the range of $[S_0 \cdot e^{\mu_A \cdot T - \sigma_A \cdot \sqrt{T}}, S_0 \cdot e^{\mu_A \cdot T + \sigma_A \cdot \sqrt{T}}]$ at time T . Let us illustrate this using the example of the S&P500:

For this purpose, we use the trend $\mu_A = 6.49\%$ that we calculated above as of October 2017 and an average value of the volatility calculated for that time of $\sigma_A = 15.23\%$.

The value of the S&P500 as at the reporting date in October 2017 was 2597 points. Figure 1.1 shows how the S&P500 moved until October 2017, i.e. until the value of $S_0 = 2597$ points. To the right of this, you see the area $[S_0 \cdot e^{\mu_A \cdot T - u\sigma_A \cdot \sqrt{T}}, S_0 \cdot e^{\mu_A \cdot T + u\sigma_A \cdot \sqrt{T}}]$ for $u = 1$ (red), $u = 2$ (green) and $u = 3$ (turquoise) for the following 10 years from October 2017 to October 2027.

If we select an arbitrary point in time T in the future, then, under the above conditions, the probability for the S&P500 value to be between the two red curves at that time would be 68.27%, the probability for it to be between the two green curves would be 95.45%, and the probability for it to be between the two turquoise curves would be 99.73%.

Viewed over a longer period of time, there are generally phases where a stock price (index price, exchange rate ...) experiences stronger volatility and phases where it moves more smoothly. In the examples above, we calculated the average standard deviation of the returns over a very long period of time. As a result, we obtain virtually no information about the “current fluctuation intensity” of a price.

Yet what exactly does “the current fluctuation intensity” really mean? Something like the “current energy” in the way the price moves? How should we measure and describe such “current energy”? Obviously, for any such measurement, we would always have to draw on historical data. The only question then is: How

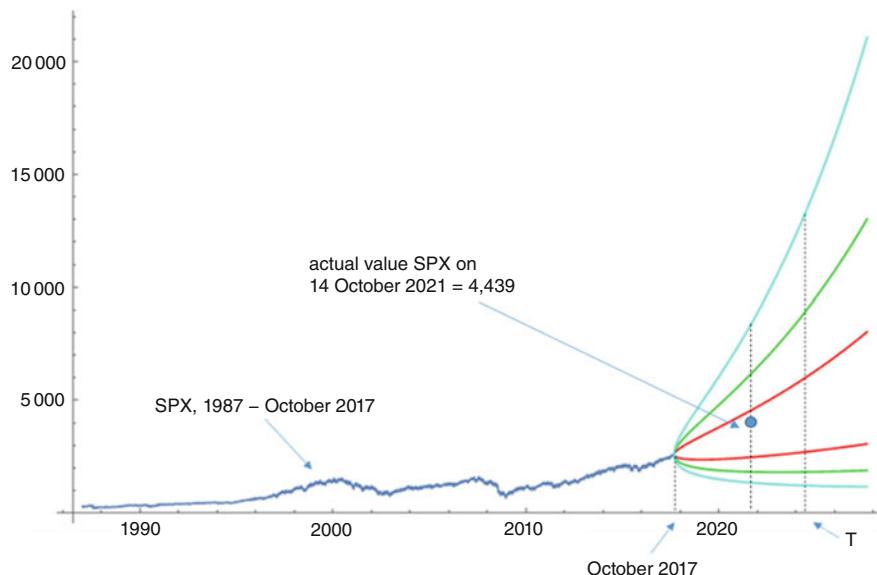


Fig. 1.1 Probability ranges for future trends of the S&P500 seen from October 2017 and actual value in October 2021

far back should the historical data go that we want to include (to what extent?) and at what intervals should such data be measured? We are going to reveal at this point already that in addition to this “historical” approach, there is also an alternative, conceptually completely different method, that is, the “implicit” method for determining volatility, known as “implied volatility”.

For the time being, however, we are going to proceed with the concept of “historical volatility”. The usual reason for measuring or estimating historical volatility is the need to estimate or forecast the intensity of a financial product’s price variations over a certain period $[0, T]$ for purposes of risk management, or valuation of a derivative, or classification of a financial product into a risk class at time 0. The idea behind this is that, if a financial product’s price has exhibited a certain (relatively stable) “fluctuation energy” over a certain period in the past, then—unless unforeseen material events occur—this “energy” can be expected to remain more or less that way for a while. The validity of the “idea” that certain volatilities tend to persist at a similar level over longer periods of time is a hypothesis that we will examine more closely later.

To estimate the volatility of a stock price A at time 0 for a period of time $[0, T]$, one could then proceed, in a first attempt and rather naively, as follows:

- Choose a (mostly short) time interval dt .
- Select a time range $[-U, 0]$ in the past where $U = N \cdot dt$.
- Look at the stock prices of A at the times

$$-N \cdot dt, -(N - 1) \cdot dt, -(N - 2) \cdot dt, \dots, -2 \cdot dt, dt, 0.$$

Denote these prices by $B_0, B_1, B_2, \dots, B_{N-2}, B_{N-1}, B_N$.

- From these prices, determine the associated continuous returns b_0, b_1, \dots, b_{N-1} .
- For this sequence of returns, calculate the standard deviation and denote it by $hv'(0, dt, N)$.
- Annualize this, using $hv(0, dt, N) := \frac{hv'(0, dt, N)}{\sqrt{dt}}$.
- $hv(0, dt, N)$ will then be the (annualized) (N, dt) -historical volatility of A at time 0.

If we run this process at an arbitrary time t (instead of time 0), we always calculate the value $h(t, dt, N)$, i.e. the (N, dt) -historical volatility of A at time t .

For example, if you read about the historical 20-day volatility of a stock price at a point in time t , then this means $h\left(t, \frac{1}{255}, 20\right)$ (assuming 255 trading days in a year), where this value is calculated from the daily prices of the 20 trading days immediately preceding the point in time t .

In Chart 1.2, the annualized historical volatilities for the S&P500 were calculated and plotted daily based on daily price quotations. The calculation period varies between 10, 50, and 200 days.

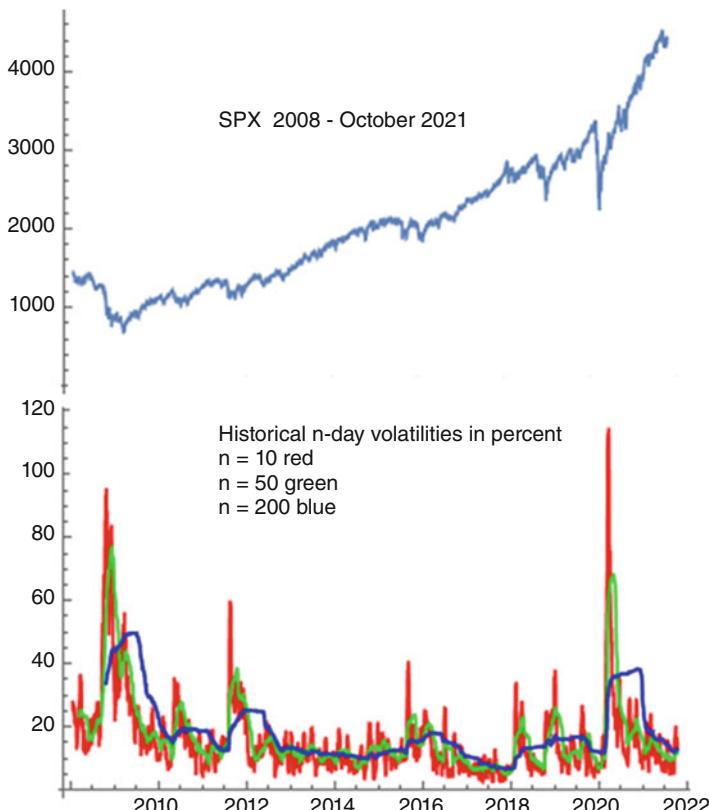


Fig. 1.2 S&P500 and annualized historical volatilities of S&P500 based on daily prices

It is obvious and naturally logical that the longer the period with (n) prices is from which the historical volatility is calculated, the smoother the graph $h\left(t, \frac{1}{255}, n\right)$ becomes.

It is equally obvious and also quite logical that the historical volatilities attain higher values wherever the S&P500 exhibits a stronger tendency to fluctuate (particularly in the second half of 2008, in the last quarters of 2012 and 2016, in 2018, and above all during the first covid wave in early 2020) and that the historical volatilities only react with a certain delay to fluctuation changes (increase or decrease in fluctuation intensity), with that delay getting longer, the longer the calculation period is.

The maximum and minimum historical volatilities are of course all the more pronounced the shorter the calculation period is (the smaller n is). In the case of a longer calculation period, more data is averaged and so the resulting values even out more.

Figure 1.3 plots the historical volatilities of the S&P500 from when it was first calculated in 1957.

The average value over all of the volatilities reported over the entire period is 13.93. The overwhelming majority of the historical volatility values were in the range from around 5% to around 30%. Outliers in ranges above 30% occurred above all in the high-volatility periods after 17 October 1987, the period of the Internet bubble up until it burst (1999–2003), during and in the aftermath of the financial crisis 2008–2011, and during the first covid wave in early 2020.

A significant disadvantage of this (naive) historical method for calculating volatilities is that extreme singular events can cause volatility (especially when using longer calculation periods, i.e. for larger n) to remain at a high level for longer

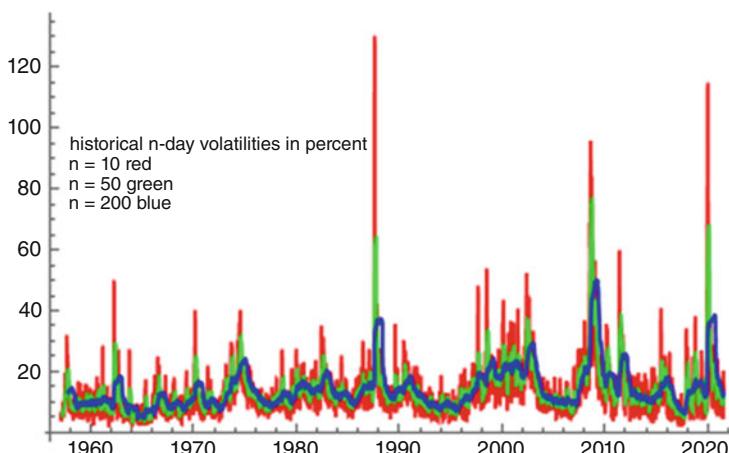


Fig. 1.3 Annualized historical volatilities ($n = 10, 50, 200$) based on daily data, 1957–October 2021

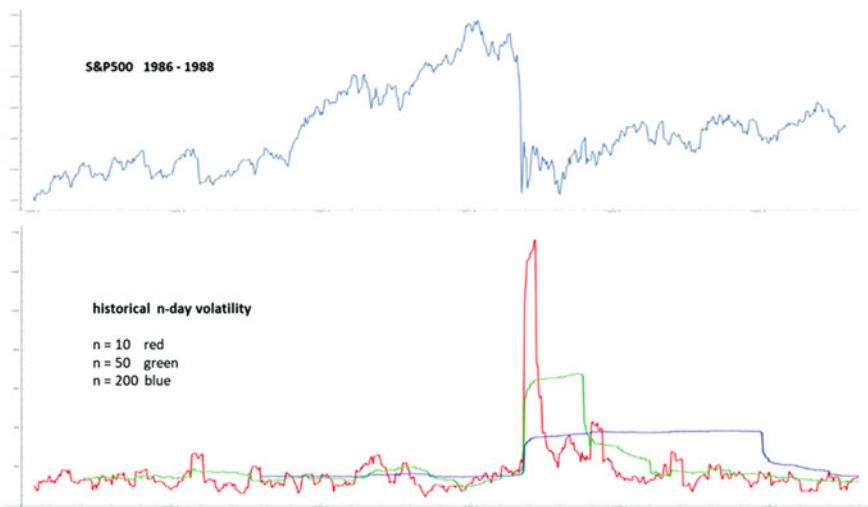


Fig. 1.4 S&P500 and annualized historical volatilities of the S&P500 based on daily rates 1986–1988

periods of time even if, on the face of it, the fluctuation tendency of the financial asset's price has already subsided considerably.

This can be clearly seen, for example, in the performance of the S&P500 after 17 October 1987. Figure 1.4 shows the S&P500 and the three different historical volatilities once more for the period from 1986 to 1988. The S&P500 exhibits an extreme plunge in the 2 or 3 days around 17 October 1987. Then, it calms down very quickly and immediately settles—as per the graph—at a fluctuation level similar to the one before 17 October 1987.

While the historical volatility for $n = 10$ (red graph) reacts very quickly and returns to earlier levels when the energy of the S&P500 “cools off”, the extreme event of 17 October has a long-lasting impact on the historical volatility for $n = 200$ (blue graph), which remains at a very high level for a long period of time, for then to drop sharply (after about 200 days, i.e. with substantial delay).

Does this mean that a shorter calculation period yields more useful historical volatility values? At first glance, it would indeed seem so. However, using shorter calculation periods means that all too quickly, the past would get completely ignored. And it would still have a significant flaw: One extreme event would enter into the calculation of the historical volatility with a sudden, relatively massive impact, for then to disappear again completely after n days.

1.2 Volatility II: ARCH Models

It is quite obvious therefore that we need to consider a different concept of historical volatility—one where volatility calculations include price movements over a longer past period but where price values dating further back play a much smaller role than more recent ones. In this concept, the influence of earlier price movements diminishes over time. This approach is used in the ARCH and GARCH models to estimate historical volatility. Before introducing these models just briefly (!), a side note seems called for.

Side Note So far, we have confined ourselves to calculating and estimating historical volatilities on the basis of daily data and to presenting them on the basis of daily calculations, and we will continue to do so in the following. All calculations could of course also be carried out on the basis of data drawn every minute, week, month In such cases, however, the following would need to be taken into account:

If we were to calculate historical volatilities on a daily basis from, for example, weekly data, we would regularly encounter the phenomenon that the calculated volatility may take very different values on successive days. This is due to the fact that, say, on a Tuesday, different price values (values of the last n Tuesdays) are used for calculating the historical n -week volatility than on the day before (values of the last n Mondays), and that causes instability. The volatility graph would then typically look something like the following Fig. 1.5:

To obtain a consistent representation when using, for example, weekly data, volatility should be calculated and presented only once a week for a fixed weekly date. Then, you would typically get a smoother curve, as shown in Fig. 1.6 (green curve compared to the above red curve).

As noted above, we will continue to use daily data in the following and calculate historical volatility on a daily basis.

So the idea now is to look at a type of historical volatility, where earlier returns are given a lower weighting in calculating volatility than more recent returns. This is done using the so-called ARCH and GARCH models for volatilities.

When using an **ARCH model** to estimate the volatility of a stock price A at time 0 for a time range $[0, T]$, the procedure is as follows:

- Choose a (mostly short) time interval dt .
- Select a time range $[-U, 0]$ in the past where $U = N \cdot dt$.
- Select non-negative weights $\alpha_0 \geq \alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_{N-1}$ and β so that $\alpha_0 + \alpha_1 + \alpha_2 + \dots + \alpha_{N-1} + \beta = 1$.

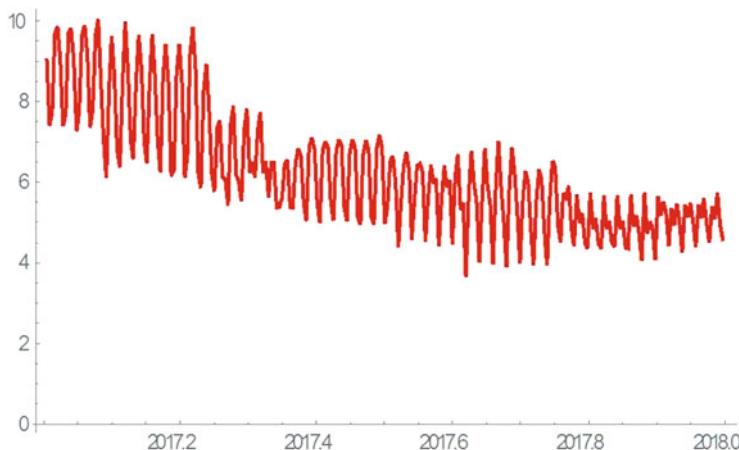


Fig. 1.5 Historical volatility of the S&P500 in 2017, daily calculation based on 20 weekly price values



Fig. 1.6 Historical volatility of the S&P500 in 2017, daily calculation (red) and weekly calculation (green), each based on 20 weekly price values

- The β parameter only comes into play if you have a reliable estimate sig_L^2 of a relatively stable average value, taken over many years, for the square of the stock's volatility per annum. Otherwise, you set $\beta = 0$. We write $V := dt \cdot sig_L^2$.
- Look at the prices of A at the times $-N \cdot dt, -(N-1) \cdot dt, -(N-2) \cdot dt, \dots, -2 \cdot dt, dt, 0$.
We denote these prices by $B_0, B_1, B_2, \dots, B_{N-2}, B_{N-1}, B_N$.
- From these prices, we determine the associated continuous returns b_0, b_1, \dots, b_{N-1} .
- The mean of these returns is denoted by μ_0 .

- We determine $u_i^2 := (b_i - \mu_0)^2$ for $i = 0, 1, \dots, N - 1$.
- Then, we calculate the volatility estimate at time 0 using $\sigma'(0, dt, N)^2 := \beta \cdot V + \sum_{i=0}^{N-1} \alpha_i \cdot u_i^2$.
- We annualize again using $\sigma(0, dt, N) := \frac{\sigma'(0, dt, N)}{\sqrt{dt}}$. We then denote $\sigma(0, dt, N)$ as the (annualized) (N, dt) -historical volatility of A at time 0 based on the ARCH model with parameters $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_{N-1}, \beta$, and V .

If we perform this procedure for an arbitrary point in time t (instead of time 0), we calculate the value $\sigma(t, dt, N)$. Let us first examine, purely informally, the expectation $E(\sigma^2(0, dt, N))$ of this estimator $\sigma^2(0, dt, N)$:

$$E(\sigma^2(0, dt, N)) = E\left(\frac{\sigma'^2(0, dt, N)}{dt}\right) = \beta \cdot \frac{V}{dt} + \sum_{i=0}^{N-1} \alpha_i \cdot \frac{E(u_i^2)}{dt}.$$

If the estimate sig_L^2 for the long-term average value of the square of the standard deviation per annum is truly reliable, then $E(u_i^2) \approx dt \cdot sig_L^2$ should hold. Hence:

$$\beta \cdot \frac{V}{dt} + \sum_{i=0}^{N-1} \alpha_i \cdot \frac{E(u_i^2)}{dt} \approx \left(\beta + \sum_{i=0}^{N-1} \alpha_i\right) \cdot sig_L^2 = sig_L^2.$$

Thus, the expected value of our squared estimate is indeed equal to the estimated long-term mean of the squared volatility.

A common choice for the α_i parameters is as follows: $\alpha_i = \lambda \cdot \alpha_{i-1}$ for $i = 1, 2, \dots, N - 1$ with a parameter λ between 0 and 1. The value for α_0 then results from the condition

$$1 = \alpha_0 + \alpha_1 + \alpha_2 + \dots + \alpha_{N-1} + \beta = \alpha_0 \cdot (1 + \lambda + \lambda^2 + \dots + \lambda^{N-1}) + \beta = \alpha_0 \cdot \frac{1-\lambda^N}{1-\lambda} + \beta \text{ i.e. } \alpha_0 = (1 - \beta) \cdot \frac{1-\lambda}{1-\lambda^N}.$$

The parameter β , which can be freely selected, controls how much weight is assigned to the estimated long-term mean sig_L^2 in the model.

The parameter λ , which can be freely selected, controls how much weight is assigned to current returns in the volatility estimate versus returns in the past. A λ value close to 1 weights earlier returns more heavily than a λ close to 0.

Let us look at an example using the S&P500. First, we choose:

$$N = 100 \\ dt = \frac{1}{255}$$

$\lambda = 0.94$ (this λ value is suggested by J.P. Morgan, e.g. see ([1], Chapter 23.2))

The only other choice we need to make now is that of the long-term mean sig_L^2 of the square deviation of the returns from the mean as well as the β weighting of this mean in the model. For this, we are proceeding very superficially at this point.

The appropriate methods for calibrating the parameters in an ARCH model will not be discussed here, as the example is mainly intended for illustration purposes.

For the parameter β , we will select a couple of different values in the following graphs and observe the influence of these different choices.

If we estimate the long-term mean for the squared standard deviation of returns using all the S&P500 data from 1957 until October 2021, we get an annualized value of 0.0257 (which corresponds to an average volatility of 16.02%). However, if we test this value for stability and calculate this long-term mean for different 10-year blocks, we get the following picture, for example:

Mean of the annualized squared standard deviation in the time range:

1957–1966	0.0110 (corr. to standard deviation of 10.52%)
1967–1976	0.0175 (corr. to standard deviation of 13.22%)
1977–1986	0.0184 (corr. to standard deviation of 13.57%)
1987–1996	0.0251 (corr. to standard deviation of 15.86%)
1997–2006	0.0331 (corr. to standard deviation of 18.20%)
2007–2016	0.0436 (corr. to standard deviation of 20.88%)
2017–October 2021	0.0379 (corr. to standard deviation of 19.46%)

Thus, there is a certain degree of fluctuation when estimating long-term squared standard deviation over various longer time blocks. Still, we can certainly choose the overall mean as the value for sigL^2 , i.e. set $\text{sigL}^2 = 0.0257$. However, the β value should not give too much weight to this estimate. In the following graphs, we see some examples using these parameters for different time periods of the S&P500 and in comparison with the naive approaches for determining historical volatility.

Figure 1.7 (above) shows the S&P500 again in the period around 17 October 1987. In the lower graph, the blue curve represents the 20-day historical volatility, the green curve the 100-day historical volatility, and the red curve the volatility estimate modelled with ARCH and the above parameters where $\beta = 0.1$. The great advantage of the ARCH approach is clear to see, as it plots the very plausible scenario of an immediately beginning yet continuous abatement of the “explosion of volatility” caused by the extreme event on 17 October 1987.

Figure 1.8 shows the S&P500 and the same volatilities with the same parameters as above for the period from 2008 to the end of October 2018. Here again, the advantages of the ARCH approach are clearly visible: Calmer behavior than the short-term historical volatility yet responding faster and subsiding more uniformly than the long-term historical volatility.

In Fig. 1.9, we illustrate the influence of the parameter β , which was selected (with all other parameters remaining the same) as 0.2 (blue curve), 0.4 (red curve), 0.6 (green curve), and 0.8 (orange curve). The closer β is chosen to 1, the smoother the ARCH curve becomes, approaching the long-term mean.

Finally, Fig. 1.10 shows the effect of changing the λ parameter. In this chart, we chose the values 0.98 (red curve), 0.88 (blue curve), and 0.78 (turquoise curve) for λ , with all other parameters remaining the same ($\beta = 0.2$).

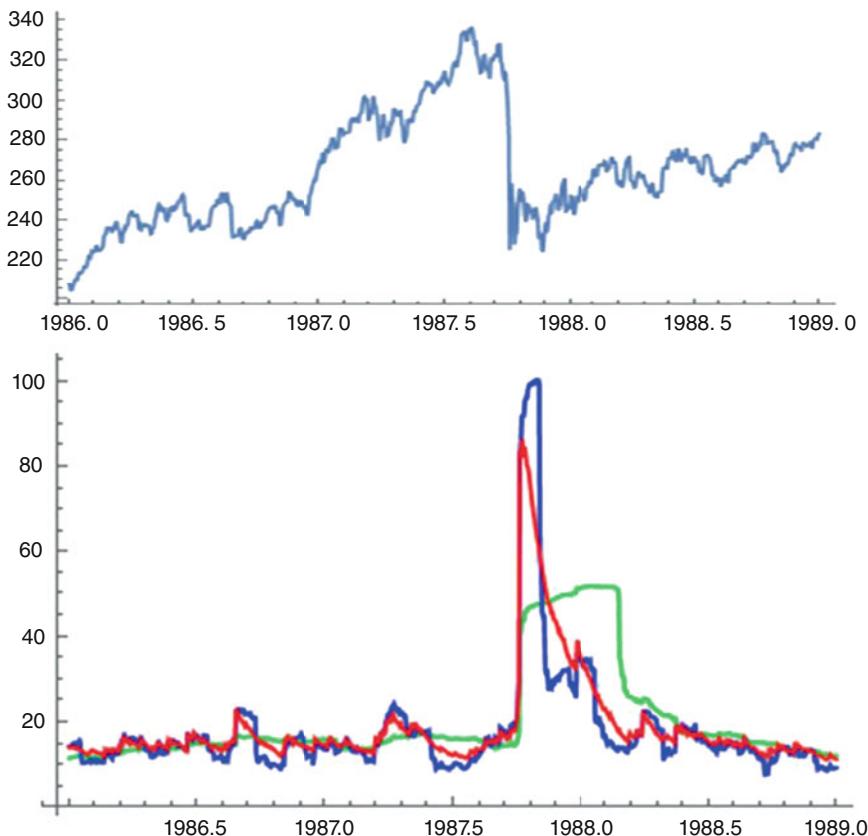


Fig. 1.7 S&P500 from 1986 to 1988 (above) and modelling of volatilities, 20-day historical (blue), 100-day historical (green), 100-day ARCH (red)

The closer λ is to 1, the more the ARCH modelling resembles the original historical volatility (apart from the β value and thus from the long-term average).

(If we choose $\beta = 0$ and $\lambda = 1$, we get exactly the original historical volatility.)

An alternative to the ARCH model for modelling volatilities are the GARCH(p, q)—models. Most widely used is the GARCH(1, 1)—model, which bears a strong resemblance to the ARCH model discussed above with the above weights α_i . We are not going to look at these models in more detail at this point.

Our software naturally offers all the various options for determining historical or ARCH volatilities.

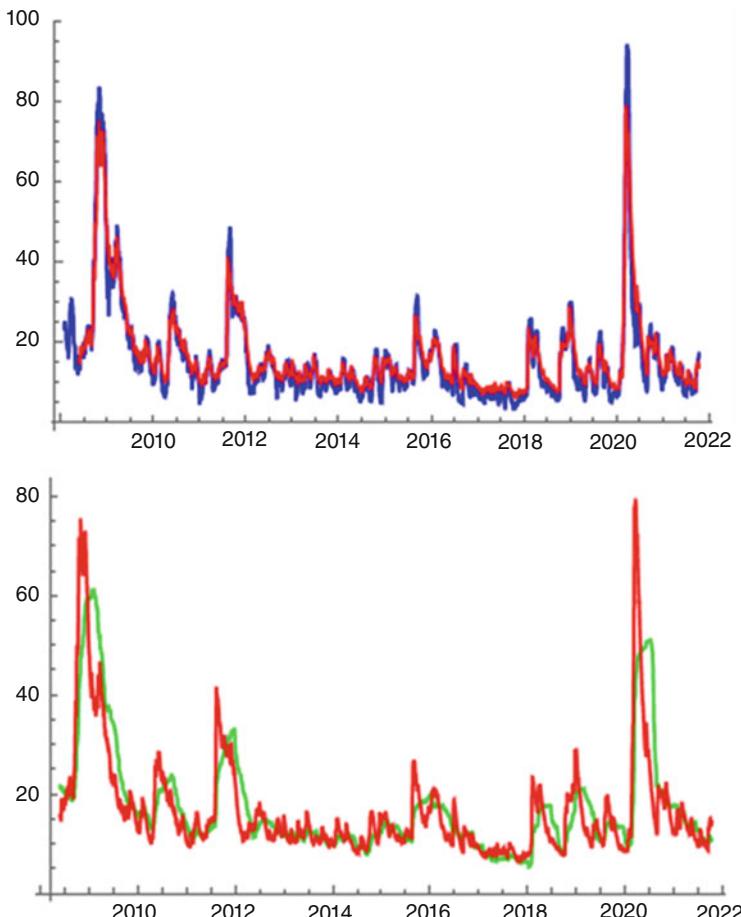


Fig. 1.8 Modelling of volatilities, 20-day historical (blue), 100-day historical (green), 100-day ARCH (red)

1.3 Volatility III: How to Use and Forecast Volatility

We return once again to our reasons for looking into volatility, its measurement, its modelling, and its forecasting, and we ask ourselves to what extent our above discussions have already helped us gain substantial insights.

Our starting point was that we wanted to force the intuitively plausible concept of “volatility” of a financial asset’s price into a quantitative format, i.e. be able to quantify it by means of a plausible measure.

The first difficulties we have identified so far, and the questions that will continue to arise, are:

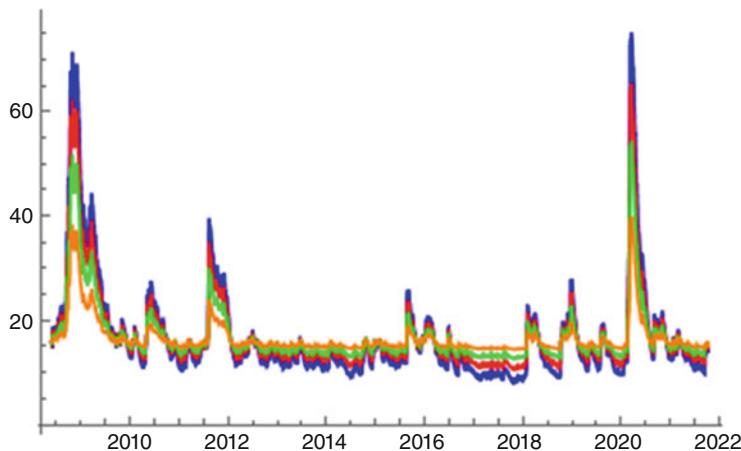


Fig. 1.9 ARCH volatility modelling for the S&P500 from 2008 to October 2021 for various β values

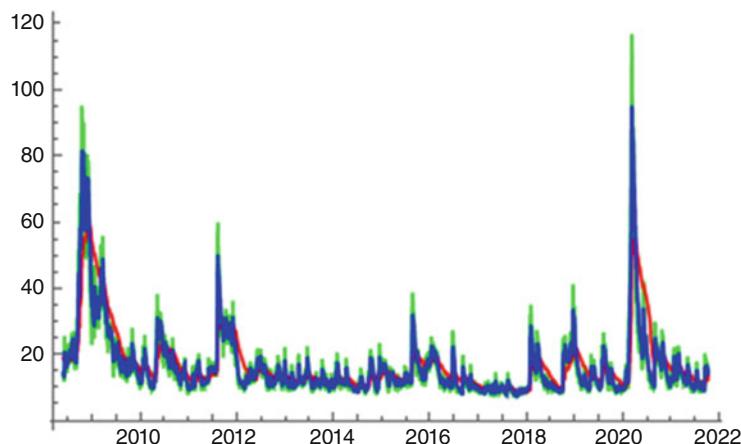


Fig. 1.10 ARCH volatility modelling for the S&P500 from 2008 to October 2021 for various λ values.

- There are various possible concepts for quantifying volatility.
- We need to understand that in order to determine volatility (based on existing price data), we will always have to analyse data over a certain period in the past.
- The actual results from such measurements depend on the data that we select (tick data, daily data, closing prices, highs, lows, ...) and the time period in which these data are observed and analysed.
- All concepts presented so far provide information about volatility in a certain (albeit perhaps short) time interval in the past, but not about the **current** inherent **volatility** of a financial product's price, i.e. its inherent volatility regardless of

past data. The question then begs: What “energy” is inherent in the stock’s price at this very moment?

- Is it possible to accurately formulate the idea of such a current inherent volatility, and if so, what would it be needed for?
- What are the benefits of knowing a financial product’s historical (or current) volatilities?

Let us revisit one of the questions that we asked ourselves at the outset of our reflections on volatility: Why do we need volatility estimates in financial engineering? Let us look once more at the main areas where volatility is used:

Risk Classification of (Structured) Financial Products For this purpose, a posteriori measurement of historical volatilities (of different types), calculated from historical price data of the product that is to be classified, is often sufficient. This can be done using the techniques described above. Here is an example: The Austrian Investment Fund Act requires a risk/return classification into seven different risk classes using a synthetic risk and return indicator. This indicator is based on the past volatility of the relevant fund series, more specifically, on the weekly returns of the last 5 years, taking into account any distributed income and dividends.

Risk Management Tools One of the key metrics used in the field of risk management is the value at risk (VAR) of a financial product or financial portfolio. We will discuss the concept of VAR in more detail in later chapters on risk management. For now, let us simply say: The VAR concept provides information on the probability with which the price of a financial product will exceed or fall below a certain value within a certain period of time $[0, T]$ in the future. It is obvious that, to calculate such a probability, we need (at least) an estimate of the financial product’s (the portfolio’s) volatility (in whatever sense) for the future time range $[0, T]$. One way to obtain such an estimate consists in analysing that financial product’s (or portfolio’s) historical volatilities and drawing (substantiated) conclusions therefrom as to future volatility. The question is, of course, how reliable such forecasts based on historical volatilities are.

Pricing of Derivatives In the Black-Scholes approach for pricing a derivative D on an underlying asset A , the critical key parameter is the volatility of A over the derivative’s life $[0, T]$. In discussing that approach, we proceeded from the assumption that in the future time range $[0, T]$, the underlying asset follows a Wiener model, in which the returns are normally distributed with a (constant) standard deviation σ . But don’t we have to ask ourselves what kind of standard deviation is meant here? The (annualized) standard deviation of returns from time 0 to time T ? Or the (annualized) standard deviation of returns in short time intervals of length dt in the interval $[0, T]$? The answer is simple: If we assume normally distributed and independent returns (with constant σ over $[0, T]$), then all approaches (when applying normalization) lead to the same annualized value. It is therefore sufficient to give an estimate for the standard deviation of returns

for time intervals of any arbitrary length dt in $[0, T]$. The question then is, again, whether (and if so, which) historical volatilities are suited for predicting such future volatilities.

To summarize: There are many settings where estimates or forecasts of standard deviations σ of financial product's returns in a time range $[0, T]$ are needed at a point in time 0.

For these returns, we continue to assume (for now) that they are independent of one another and are normally distributed over time.

Before we go any further, let us first ask ourselves the following question: If we arrived at an estimate for σ at time 0 (by analysing historical volatilities, for example), how can we verify at time T how good this forecast actually was?

What was the actual **realized volatility** in the time range $[0, T]$?

Now, again under the usual assumption (of normally distributed and independent returns), we can again divide $[0, T]$ into many small subintervals of length dt , determine the returns in each of these intervals and their standard deviation, and annualize this value. The result essentially reflects the volatility realized in $[0, T]$.

With this, we can now reframe the question as to the predictive ability of the above concepts in more precise terms:

Which estimate at time 0 (historical volatility, ARCH volatility, based on which historical data, completely different approach?) provides, on average, the best forecast for the actual realized volatility in the time range $[0, T]$? In adding “a completely different approach” above, we were already dropping a hint, as you may well have noted, that we will soon discuss a completely different—and highly essential—alternative approach (i.e. the concept of “implied volatility”).

At this point, however, we are just going to run a few simple (representative and possibly stimulating) experiments on the predictive ability of historical volatilities and ARCH volatilities for future realized volatilities. As a representative numerical example, let us calculate the realized volatilities of the S&P500 from 1957 to October 2018 for periods of 21 trading days each (≈ 1 trading month), calculated on the basis of daily returns in each case. In a first step, we are going to estimate these realized volatilities based on the realized volatilities (the historical volatility) of the previous interval. Thus, denoting the successive realized volatilities by w_1, w_2, \dots, w_M , we estimate w_{i+1} based on w_i in each case and ask ourselves about the quality of this estimate. Is there a strong positive correlation between w_i and w_{i+1} or are these values generally quite independent of each other?

To answer this question, we will first graphically represent the pairs (w_i, w_{i+1}) two-dimensionally in Fig. 1.11 and then analyse the correlations between the sequence $(w_1, w_2, \dots, w_{M-1})$ of the forecast values and the sequence (w_2, w_3, \dots, w_M) of the predicted values. That is, we determine the first-order autocorrelation of the sequence (w_1, w_2, \dots, w_M) . As most of the values w_i are in the range between approx. 3% and 35%, the graphic representation is limited to this range. For visual comparison, we also plotted the same number of pairs (u_i, u_{i+1}) in Fig. 1.11 with independent randomly chosen coordinates ranging from 3 to 35. These random points fill the area of the square $[3, 35] \times [3, 35]$ fairly evenly. The pairs (w_i, w_{i+1}) on the other hand are often close to the straight line $f(x) = x$,

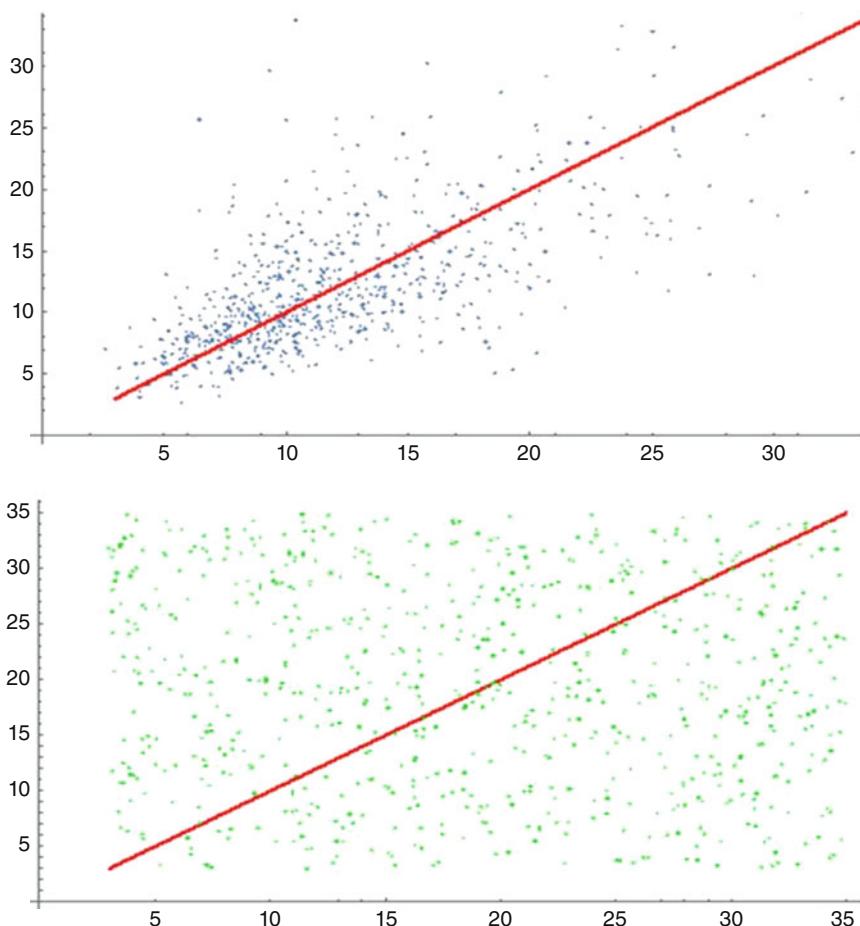


Fig. 1.11 Pairs (forecast, realized volatility) through and for historical 1-month volatilities based on daily data, S&P500, 1957–October 2018 (above) compared with random pairs (below)

which means that the two values w_i and w_{i+1} are often very similar in size. We can therefore say that an approximate value of w_{i+1} can indeed be inferred with a certain probability from w_i . The correlation between the sequence $(w_1, w_2, \dots, w_{M-1})$ and the sequence (w_2, w_3, \dots, w_M) is 0.658, so clearly positive.

It is interesting to note that the forecasts show a clearly positive correlation even with subsequent realized volatilities. More precisely: The correlation (autocorrelations of orders 2, 3, 4, and 5)

between

$(w_1, w_2, \dots, w_{M-2})$ and the sequence (w_3, w_4, \dots, w_M) is 0.537,

between

$(w_1, w_2, \dots, w_{M-3})$ and the sequence (w_4, w_5, \dots, w_M) is 0.449,

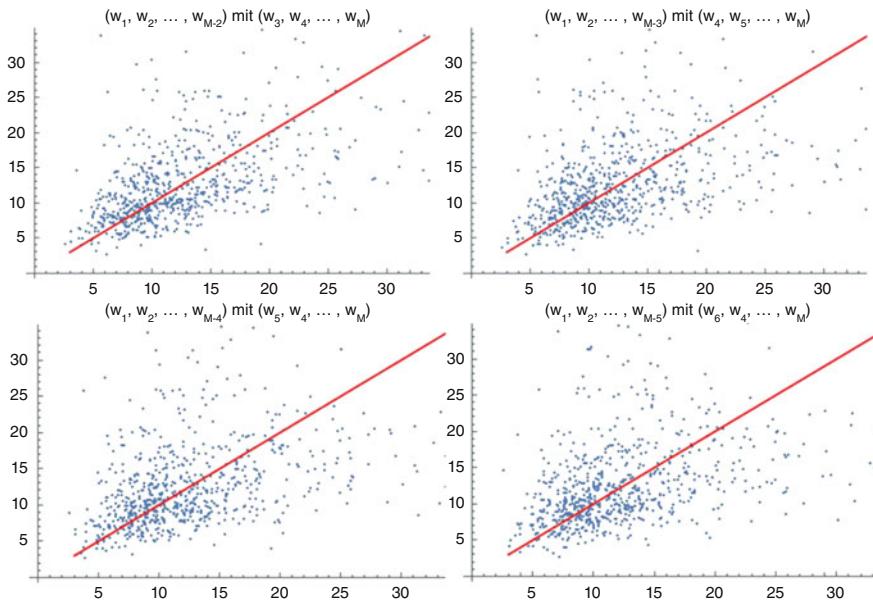


Fig. 1.12 Autocorrelations of orders 2, 3, 4, and 5 of the 20-day volatility based on daily closing prices

between

$(w_1, w_2, \dots, w_{M-4})$ and the sequence (w_5, w_4, \dots, w_M) is 0.362,
between

$(w_1, w_2, \dots, w_{M-5})$ and the sequence (w_6, w_5, \dots, w_M) is 0.358.

The corresponding pair cloud images are shown in Fig. 1.12.

We have so far purposely estimated the future historical 21-day volatility using the preceding 21-day volatility (obtaining a correlation of 0.658). Of course, this begs the question whether the historical volatility over a different time range would not provide an even better estimate of the subsequent 21-day volatility. In the following, we have therefore also used forecast values based on historical volatilities over other time periods and measured the correlations to the subsequent 21-day volatilities. This gives us the following Table 1.2:

Thus, a forecast period between 20 and 26 days seems to give the best forecasts here (i.e. forecast period of a length that roughly corresponds to the length of the period to be forecast).

We now want to create analogous (but shorter) Tables 1.3 and 1.4 for a period to be forecast of half a year (i.e. 126 days) and for a period to be forecast of 1 week (i.e. 5 days).

Based on this very small sample for the S&P500, we can see that the best forecast values for the $\frac{1}{2}$ -year volatility are those based on a shorter period of about 1 month

Table 1.2 Correlation to 21-day volatility for different time ranges of forecast values

Time range for forecast values	5	6	7	8	9	10	11	12	13
Correlation to 21-day volatility	0.633	0.577	0.582	0.602	0.615	0.616	0.628	0.631	0.633
Time range for forecast values	14	15	16	17	18	19	20	21	22
Correlation to 21-day volatility	0.636	0.639	0.651	0.650	0.652	0.655	0.659	0.658	0.661
Time range for forecast values	23	24	25	26	27	28	29	30	31
Correlation to 21-day volatility	0.662	0.662	0.661	0.660	0.647	0.647	0.649	0.650	0.651
Time range for forecast values	40	50	60	70	80	90	100	110	120
Correlation to 21-day volatility	0.653	0.640	0.637	0.614	0.608	0.598	0.595	0.590	0.584

Table 1.3 Correlation to 1/2-year volatility for different time ranges of forecast values

Time range for forecast values	$\frac{1}{2}$ month	1 month	$\frac{1}{4}$ year	$\frac{1}{2}$ year
Correlation to $\frac{1}{2}$ -year volatility	0.455	0.460	0.443	0.445

Table 1.4 Correlation to 1-week volatility for different time ranges of forecast values

Time range for forecast values	3 days	1 week	2 weeks	3 weeks
Correlation to 1-week volatility	0.523	0.600	0.641	0.637

and for the 1-week volatility based on a longer period (relative to the length to be estimated) of 2 weeks.

Just briefly—and only for the 1-month volatility—we want to examine whether better forecast values for future realized volatility can be obtained using, say, a suitable ARCH model.

For this purpose, we estimated the 1-month historical volatilities in each case using an ARCH model based on n previous daily closing prices and various parameters β and λ . A quick—not exhaustive—search for good parameter values to estimate 1-month volatilities over the entire life of the S&P500 yielded the following values:

$$n = 55$$

$$\beta = 0$$

$$\lambda = 0.95$$

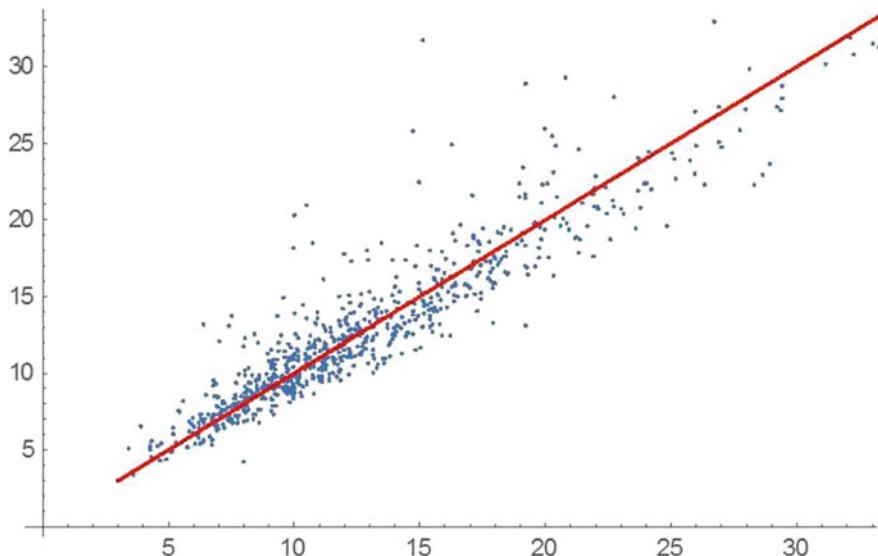


Fig. 1.13 Pairs (forecast, realized volatility) using ARCH values and for historical 1-month volatilities based on daily data, S&P500, 1957–October 2018

The corresponding point cloud of estimated value and realized volatility has the form shown in Fig. 1.13. The correlation between forecast values and realized volatilities is 0.931771!

The tests and observations made here for forecasting future historical (realized) volatilities have of course only been carried out very superficially. Much more extensive and in-depth tests are required to obtain truly relevant results. For example, the stability of the quality of the estimates would have to be tested, the search for optimal parameters would have to be substantially expanded, a wide range of other financial products would have to be investigated (not only the S&P500), and, above all, other measures of the quality of these forecasts would have to be examined in addition to correlation.

The basis for conducting such advanced experiments is given on our website.

1.4 Volatility IV: Magnitude of the Historical Volatility of the S&P500

To conclude our discussions of historical volatility and its predictability, we want to present some observations regarding the S&P500's historical volatility over time, simply for orientation and to provide a sense of the magnitude of historical volatilities. To this end, we once again visualize the 20-day historical volatility of the S&P500 from 1957 to October 2021 in Fig. 1.14.

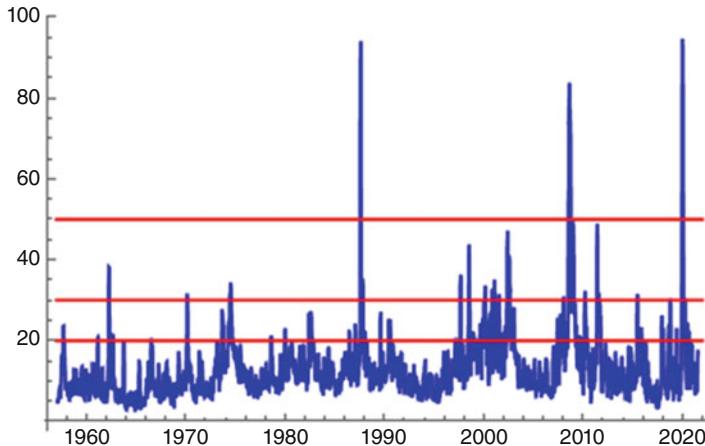


Fig. 1.14 20-day historical volatility of the S&P500 from 1957 to October 2021

Three extraordinary events over the course of the S&P500's life of 65 years to date resulted in historical (20-day) volatilities well above 50%. One was the extreme price collapse on 17 October 1987, the second one was the financial crisis in late 2008, and the third one was the event of the first covid wave in early 2020. In addition, the 30% volatility mark was exceeded at about seven points in time or time periods. They were all rather isolated events in 1963, 1970, 1975, the flash crash in 2010, late 2011 and late 2015, and the high-volatility period (Internet bubble, September 2001, Enron scandal) from 1998 to 2003. The 20% mark was also briefly breached on about the same number of occasions (without leading to volatilities above 30%, however). Most of the time, however, volatility was in the range below 20% (approx. 5%–20%). Such low-volatility phases often lasted for several years. The two longest periods of low historical volatility in the history of the S&P500 to date were from 1963 to 1970 and 1991 to 1997 (see Fig. 1.15). The two longest periods of almost consistently high volatility were from 1999 to 2003 and 2008 to 2012.

1.5 Volatility V: Volatility in Derivatives Pricing

Volatility plays a central role in derivatives pricing. As can be seen, for example, from the Black-Scholes formula for pricing derivatives through a Wiener model, the volatility of the underlying asset (in this setting) is the only sensitive parameter in this formula, and it is therefore a parameter that must be carefully estimated in order to arrive at reliable results.

But what kind of volatility is it exactly that goes into the Black-Scholes formula, that is, into pricing derivatives using the Wiener model in a time range $[0, T]$? Answer: It is precisely the volatility that is used in the Wiener model for the

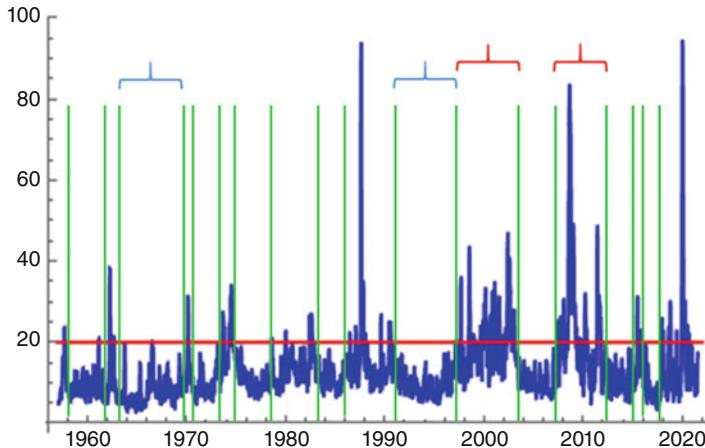


Fig. 1.15 Low-volatility and high-volatility phases in the S&P500

underlying asset, i.e. the (annualized) standard deviation of the (continuous) return of the underlying in the time range $[0, T]$. At time 0, the most I can do is try to forecast this volatility using certain metrics (historical volatility, ARCH forecasts) and/or based on fundamental considerations.

But is it even possible to measure this volatility retrospectively at time T , and if so, how? After all (even in retrospect), we only have the actual realized return on $[0, T]$ to go on. How can this *one* realized return help us infer the standard deviation of the random variable that generated that return?

If we proceed from the assumption of constant volatility over short sub-periods of length dt throughout the derivative's life (which, by the way, was also our assumption in proving the Black-Scholes formula through the binomial model), then (as already explained above) we can indeed estimate the realized volatility on $[0, T]$ retrospectively, by estimating the volatility for the sub-periods of length dt and normalizing to $[0, T]$. However, especially in the case of a longer (remaining) life to expiration at T , we cannot assume unreservedly that volatility will remain constant. So, if we were to retrospectively estimate the volatility realized in $[0, T]$ and see that, for example, there was a high-volatility phase in $[0, \frac{T}{2}]$ and a low-volatility phase in $[\frac{T}{2}, T]$, can we then still proceed in the same way for estimating total volatility? What's more is the following: Is it even possible, in such cases, to prove the Black-Scholes formula? After all, our proof was explicitly based on the assumption of constant volatility of the underlying over the derivative's life!

Luckily, the answer to these questions is as follows: Yes, it is possible. Although it is by no means self-evident. The answer is owed to Bruno Dupire and is based on the following considerations and Dupire's result below:

Let us again assume that at time 0, we want to price a derivative D expiring at time T on an underlying asset with price $S(t)$ and payoff function Φ . We also assume that $S(t)$ follows a Wiener model.

However, this time we don't assume that the underlying asset's (per annum) volatility σ remains constant but that it changes over time (and possibly also as a function of the underlying asset's price). As an example, let us again briefly take the a posteriori position at time T : Looking backward from this point, it might (!) be possible to retrospectively determine for σ —through detailed analysis—a (per annum) performance of $\sigma(t)$ for $t \in [0, T]$. However, it is most assuredly not possible to definitively determine a $\sigma(t, S(t))$, i.e. the dependence of variable t and variable $S(t)$, since the only empirical values that we can draw on are the actually occurred price $S(t)$ and the prevailing volatilities for this actual price, and so we cannot determine which values $\sigma(t, S(t))$ would have had for any other potential values of $S(t)$.

We noted above that it *might* (!) be possible, and we did so because at this point, we are not yet entirely sure how a reliable estimation of the function $\sigma(t)$ could be made—even a posteriori.

So, let us now do the following, as suggested further above: Assuming we are at time T , we

- Subdivide the interval $[0, T]$ into N equal parts of length dt (with “large N ” and “small dt ”).
- Determine the (per annum) returns $r(i \cdot dt)$ for each time interval of the form $[i \cdot dt, (i + 1) \cdot dt]$, $i = 0, 1, \dots, N - 1$.
- Calculate the (per annum) variance of the returns, i.e. $\bar{\sigma}^2 := \frac{1}{N} \sum_{i=0}^{N-1} \sigma'^2(i \cdot dt) = \frac{1}{T} \sum_{i=0}^{N-1} \sigma'^2(i \cdot dt) \cdot dt$, where $\sigma'^2(i \cdot dt)$ is the squared deviation of $r(i \cdot dt)$ from the mean.

This gives us an estimate $\bar{\sigma}$. But of what, exactly?

Now, if the function $\sigma(t)$ does not behave too irregularly, and if we choose a sufficiently small dt , then $\bar{\sigma}^2$ should be a relatively reliable estimate for $\frac{1}{T} \sum_{i=0}^{N-1} \sigma^2(i \cdot dt) \cdot dt$. If we let dt go to 0 and assuming again that σ is not too irregular, this estimate will converge to $\frac{1}{T} \int_0^T \sigma^2(t) dt$.

Note: The estimate $\bar{\sigma}$ is nothing other than the (retrospectively) determined historical (per annum) volatility of $S(t)$ on $[0, T]$ calculated for N time periods of length dt .

Why is that estimate of interest? After all, it only provides an average value of $\sigma(t)^2$ over the entire period $[0, T]$, but it does not provide any detailed information about $\sigma(t)$!

The answer is as follows: This estimate gives us everything we need. Why? Because of the following theorem of Dupire, which is a generalization of the Black-Scholes formula's basic version that we discussed in Volume I Section 4.18.

1.6 Derivatives Pricing with Time- (and Price-) Dependent Volatility: the Dupire Model

Theorem 1.1 (Black-Scholes Formula for Time-Dependent Volatility) Let D be a European-style derivative with expiration date T and payoff function Φ on an underlying asset with price $S(t)$, which in the time range $[0, T]$ moves according to a Wiener model with parameters μ and time-dependent (continuously differentiable) $\sigma(t)$. (It is assumed that no payments or costs are generated by the underlying asset.) The fair price $F(0)$ of D at time 0 is then defined as follows:

$$F(0) = e^{-rT} \cdot E(\Phi(\tilde{S}(T)))$$

where \tilde{S} is defined as

$$\tilde{S}(T) = S(0) \cdot e^{T \cdot \left(r - \frac{1}{2} \cdot \left(\frac{1}{T} \int_0^T \sigma^2(t) dt\right)\right) + w \sqrt{T} \cdot \sqrt{\frac{1}{T} \int_0^T \sigma^2(t) dt}}$$

with a standard normally distributed random variable w . “E” in this equation denotes the expected value, and r is the risk-free interest rate $f_{0,T}$.

Some comments on this:

- This result holds even if the trend μ depends on t and possibly also on $S(t)$ (and is continuously differentiable). However, this is not relevant in our context.
- The condition of continuous differentiability of $\sigma(t)$ is not relevant for practical applications.
- What is essential, however, is the following: To determine what the fair price of the derivative is in this case, or to determine a posteriori what the fair price would have been at time 0, it is not necessary to know $\sigma(t)$ explicitly. It is sufficient to know the average value of $\sigma^2(t)$, i.e. $\frac{1}{T} \int_0^T \sigma^2(t) dt$. This value can also be estimated a posteriori. The reason why we emphasize this will become clear later in two of our case studies. The above-defined value of $\bar{\sigma}^2 := \frac{1}{N} \sum_{i=0}^{N-1} \sigma'^2(i \cdot dt)$ provides a good a posteriori estimate for the required variable $\int_0^T \sigma^2(t) dt$ in the case of small dt . Here, $\bar{\sigma}$ is nothing other than the historical volatility of $S(t)$ on $[0, T]$ for the period length dt .
- This gives us the following price formulas specifically for call and put options:

$$C(t, s) = s \cdot \mathcal{N}(d_1) - e^{-r(T-t)} \cdot K \cdot \mathcal{N}(d_2)$$

$$P(t, s) = e^{-r(T-t)} \cdot K \cdot \mathcal{N}(-d_2) - S(t) \cdot \mathcal{N}(-d_1)$$

where

$$d_1 = \frac{\log\left(\frac{s}{K}\right) + (T-t) \cdot \left(r + \frac{1}{2} \cdot \frac{1}{T-t} \int_t^T \sigma^2(u) du\right)}{\sqrt{T-t} \cdot \sqrt{\frac{1}{T-t} \cdot \int_t^T \sigma^2(u) du}}$$

and

$$d_2 = \frac{\log\left(\frac{s}{K}\right) + (T-t) \cdot \left(r - \frac{1}{2} \cdot \frac{1}{T-t} \int_t^T \sigma^2(u) du\right)}{\sqrt{T-t} \cdot \sqrt{\frac{1}{T-t} \cdot \int_t^T \sigma^2(u) du}}$$

- Remember: The fair price of the derivative reflects the price for perfect hedging of the derivative! As in the case of constant volatility, hedging is done by continuously holding $\Delta(t, S(t))$ units of the underlying asset. Since $\sigma(t)$ in this case depends only on t and not on $S(t)$, the delta is calculated, i.e. differentiated with respect to $s = S(t)$, in the same way as in the case of constant volatility. So, we get exactly the same formulas for delta, only that in this case, we replace the constant value σ by $\sigma(t)$.
- In stating Dupire's theorem, the condition to be satisfied was expressed rather laconically as *... an underlying asset with price $S(t)$ which in the time range $[0, T]$ moves according to a Wiener model with parameters μ and time-dependent (continuously differentiable) $\sigma(t)$* But what exactly does that mean, in practical terms? In the case of constant σ , i.e. in the Wiener model in our original version, this condition meant that the relation

$$S(t + \tau) = S(t) \cdot e^{\mu\tau + \sigma\sqrt{\tau}w}$$

with a standard normally distributed random variable w has to hold for any period $[t, t + \tau]$. For two disjoint periods $[t_1, t_2]$ and $[t_3, t_4]$, the respective random variables w are independent of each other. In particular, the following has to hold:

$$S(T) = S(0) \cdot e^{\mu T + \sigma\sqrt{T}w}$$

and for every very small time interval dt :

$$S(t + dt) = S(t) \cdot e^{\mu \cdot dt + \sigma \sqrt{dt}w}.$$

From now on, however, we will be working on the assumption of time-dependent volatility $\sigma(t)$. Intuitively, for any arbitrary time t and an infinitesimally small dt , this would obviously mean:

$$S(t + dt) = S(t) \cdot e^{\mu \cdot dt + \sigma(t) \cdot \sqrt{dt}w}$$

and that is exactly the condition that was to be satisfied in Dupire's theorem. However, the expression *an infinitesimally small* dt may be somewhat troubling and is indeed not exactly interpretable with the mathematical tools we have discussed so far. (There exists an exact version of this intuitive relation, but for that we need the language of stochastic analysis, which we will look at later.)

We will therefore attempt to come up with a formulation equivalent to $S(t + dt) = S(t) \cdot e^{\mu \cdot dt + \sigma(t) \cdot \sqrt{dt} w}$ for infinitesimally small dt without using the concept of infinitesimally small. For purposes of deriving such an equivalent formulation in an illustrative (heuristic!) way, we will once again picture the time interval $[0, T]$ divided into N equal parts of the (very—infinitesimally—small) length dt . On the basis of $S(t + dt) = S(t) \cdot e^{\mu \cdot dt + \sigma(t) \sqrt{dt} w}$ (where w is standard normally distributed), we then get

$$\begin{aligned} S(T) &= S(T - dt) \cdot e^{\mu \cdot dt + \sigma(T - dt) \cdot \sqrt{dt} w_{N-1}} = \\ &= S(T - 2 \cdot dt) \cdot e^{\mu \cdot dt + \sigma(T - 2 \cdot dt) \cdot \sqrt{dt} w_{N-2}} \cdot e^{\mu \cdot dt + \sigma(T - dt) \cdot \sqrt{dt} w_{N-1}} \\ &= S(T - 2 \cdot dt) \cdot e^{2\mu \cdot dt + \sigma(T - 2 \cdot dt) \cdot \sqrt{dt} w_{N-2} + \sigma(T - dt) \cdot \sqrt{dt} w_{N-1}} = \\ &= \dots = \\ &= S(0) \cdot e^{N\mu \cdot dt + \sigma(0) \cdot \sqrt{dt} w_0 + \dots + \sigma(T - 2 \cdot dt) \cdot \sqrt{dt} w_{N-2} + \sigma(T - dt) \cdot \sqrt{dt} w_{N-1}} = \\ &= S(0) \cdot e^{\mu T + \sigma(0) \cdot \sqrt{dt} w_0 + \dots + \sigma((N-2) \cdot dt) \cdot \sqrt{dt} w_{N-2} + \sigma((N-1) \cdot dt) \cdot \sqrt{dt} w_{N-1}}. \end{aligned}$$

The random variable occurring here in the exponent

$$w := \sigma(0) \cdot \sqrt{dt} w_0 + \dots + \sigma((N-2) \cdot dt) \cdot \sqrt{dt} w_{N-2} + \sigma((N-1) \cdot dt) \cdot \sqrt{dt} w_{N-1}$$

is a sum of N independent normally distributed random variables

$$\sigma(0) \cdot \sqrt{dt} w_0, \dots, \sigma((N-2) \cdot dt) \cdot \sqrt{dt} w_{N-2}, \sigma((N-1) \cdot dt) \cdot \sqrt{dt} w_{N-1}.$$

These random variables all have expected value 0 and the standard deviations

$$\sigma(0) \cdot \sqrt{dt}, \dots, \sigma((N-2) \cdot dt) \cdot \sqrt{dt}, \sigma((N-1) \cdot dt) \cdot \sqrt{dt}.$$

Again, the random variable W is normally distributed with expected value 0, and its standard deviation is

$$\begin{aligned} \sqrt{\sigma^2(0) \cdot dt + \dots + \sigma^2((N-2) \cdot dt) \cdot dt + \sigma^2((N-1) \cdot dt) \cdot dt} &= \\ &= \sqrt{\sum_{i=0}^{N-1} \sigma^2(i \cdot dt) \cdot dt} \end{aligned}$$

Here, we used the following general property of the sum of two normally distributed random variables:

Let X and Y be independent normally distributed random variables with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y . Then the random variable $Z := X + Y$ is also normally distributed, with expected value $\mu_X + \mu_Y$ and with standard deviation $\sqrt{\sigma_X^2 + \sigma_Y^2}$.

The random variable W can therefore be expressed in the form

$W = w \cdot \sqrt{\sum_{i=0}^{N-1} \sigma^2(i \cdot dt) \cdot dt}$, where w is a standard normally distributed random variable.

The value $\sum_{i=0}^{N-1} \sigma^2(i \cdot dt) \cdot dt$ is a Riemann sum that converges to the integral $\int_0^T \sigma^2(t) dt$ for $dt \rightarrow 0$. (Provided that $\sigma^2(t)$ is an integrable function, which we can assume as a given in practice.) So, we find that:

$$S(T) = S(0) \cdot e^{\mu T + w \cdot \sqrt{\int_0^T \sigma^2(t) dt}}$$

with a standard normally distributed random variable w . Of course, this relation holds analogously for any time interval $[t, t + \tau]$ instead of $[0, T]$, hence

$$S(t + \tau) = S(t) \cdot e^{\mu \tau + w \cdot \sqrt{\int_t^{t+\tau} \sigma^2(s) ds}}.$$

Thus, we have deduced the mathematically unique definition for the premise in Dupire's theorem:

The price $S(t)$ of a financial product moves over $[0, T]$ according to a Wiener model with parameters μ and time-dependent (square-integrable) $\sigma(t)$ if the following holds for each subinterval $[t, t + \tau]$ of $[0, T]$:

$$S(t + \tau) = S(t) \cdot e^{\mu \tau + w \cdot \sqrt{\int_t^{t+\tau} \sigma^2(s) ds}} = S(t) \cdot e^{\mu \tau + w \cdot \sqrt{\tau} \sqrt{\frac{1}{\tau} \cdot \int_t^{t+\tau} \sigma^2(s) ds}}$$

with a standard normally distributed random variable w . Here, the random variables occurring for disjoint subintervals $[t_1, t_1 + \tau_1]$ and $[t_2, t_2 + \tau_2]$ are always independent of each other.

As noted above, it is not possible, not even a posteriori, to determine or reliably estimate a time- **and** price-dependent version of a financial product's volatility. Yet this would seem to contradict what we have occasionally practiced so far: In Volume I Section 4.28, we pointed out, and in the analyses in the following sections, we made use of the fact that we can often observe increased volatilities when stock prices are falling and decreasing volatilities when stock prices are going up. For

example, when determining break-even points of various derivative portfolios at different times t over their lives, our analyses would factor in that higher volatilities are generally to be expected when stock prices fall and that this should be taken into account when determining, for example, break-even points. For this purpose, we also proposed various formulations for the dependence of volatility on the price of the underlying asset.

However: In these chapters, we also pointed out that we did not know (at the time) what exactly the “current volatility” of an underlying asset was supposed to mean, in particular the current volatility that we intended to use to price a derivative (to determine the fair price of a derivative). And indeed, the above observation regarding the dependence of volatility on the price of the underlying asset applies above all to the so-called implied volatility—which we will discuss in the next sections—and not to the historical volatility discussed so far nor to the specific volatility occurring in the Wiener model.

But let us nevertheless suppose that we have a reliable estimate of the volatility σ occurring in the Wiener model in the time interval $[0, T]$ as a function of time t and the relevant stock price $S(t)$, that is, an estimate for the function $\sigma(t, S(t))$. This would mean that (here again we are going to write this in heuristically intuitive notation) for every time t in $[0, T]$ and every infinitesimally small dt , the following would hold:

$$S(t + dt) = S(t) \cdot e^{\mu \cdot dt + \sigma(t, S(t)) \cdot \sqrt{dt} w}$$

Would we then still be able to deduce a formula for the fair price of a derivative on that underlying asset?

The answer is yes and is given in Dupire’s more general result below. However, it is not given as an explicit formula for the price of such a derivative but as a solution of a certain partial differential equation (PDE). In most cases, this PDE cannot be solved explicitly, yet it is possible to find arbitrarily accurate approximations. How to approximate solutions of this PDE (and of PDEs in general) is not a topic we will go further into here.

Theorem 1.2 (Black-Scholes Formula for Time- and Price-Dependent Volatility)
Let D be a European-style derivative with expiration T and payoff function Φ on an underlying asset with price $S(t)$, which in the time range $[0, T]$ evolves according to a Wiener model with parameters μ and time- and price-dependent (continuously differentiable) $\sigma(t, S(t))$.

(It is assumed that no payments or costs are generated by the underlying asset.)

Then, for the fair price $F(t, s)$ of D at time t and price s of the underlying, we get the following partial differential equation:

$$\frac{dF(t, s)}{dt} + \frac{\sigma(t, s)^2 \cdot s^2}{2} \cdot \frac{d^2F(t, s)}{ds^2} + r \cdot s \cdot \frac{dF(t, s)}{ds} - r \cdot F(t, s) = 0$$

with the boundary condition $F(T, s) = \Phi(s)$.

1.7 Implied Volatility

We are now, at time 0, going to price an option with expiry T on an underlying asset that follows a Wiener model. What value should we choose for the volatility parameter?

We know that the correct choice would be to use the value $\sqrt{\frac{1}{T} \int_0^T \sigma^2(t) dt}$ in the first of the two Dupire theorems mentioned above, thus a value that could be estimated quite well a posteriori from time T . But how can we reliably estimate this value from the perspective of time 0?

Let us ask ourselves another question first: Why, in practice, would we even want to price a derivative, and what is the benefit of knowing the fair price of a derivative? There are essentially three reasons why we need to price derivatives (or portfolios of derivatives, or derivative trading strategies):

- We identify a derivative on the market that we find interesting, and we want to know whether the price at which the derivative is currently offered (on the exchange or the OTC market) is reasonable or too high or too low. Our purchase decision will depend on the outcome of that analysis.
- We wish to develop and market a new derivative of an underlying asset: What is the fair price at which we could offer it for sale? In addition, our analysis would also give us information on how much a perfect hedging of the derivative would cost and how it would have to be carried out.
- Often, the question is not about finding the explicit fair price of a derivative or whether the current price of a derivative is fair and reasonable but rather whether the relation of the prices of two derivatives on the same underlying is correct, that is, whether one of the derivatives is overpriced compared to the other. (Significant deviations from the correct relation could be exploited in arbitrage attempts—by dynamically selling the overpriced derivative and simultaneously buying the underpriced derivative.)

For most of the relevant underlying assets traded on major exchanges, a wide variety of derivatives are nowadays traded on exchanges or OTC.

In the following, we assume (without explicitly stating so in every case): a fixed underlying asset with price $S(t)$ which follows a Wiener model and a risk-free interest rate r that remains essentially unchanged throughout the derivative's life. The prices $F(0)$ at time 0 of all these derivatives (with payoff function Φ) are given by the Black-Scholes formula. We write this schematically as

$$F(0) = BS(T, \sigma, \Phi).$$

This representation is meant to indicate: The price of every derivative on this underlying is given by the derivative's expiration, by its payoff function Φ and by the choice of the underlying's volatility σ . For simplification, we substituted

the symbol σ here for the actual value $\sqrt{\frac{1}{T} \int_0^T \sigma^2(t) dt}$. We immediately notice, of course, that this value $\sigma = \sqrt{\frac{1}{T} \int_0^T \sigma^2(t) dt}$ is also dependent on T . So instead of σ , we have to set a volatility parameter $\sigma(T) = \sqrt{\frac{1}{T} \int_0^T \sigma^2(t) dt}$ that is dependent on the derivative's expiration T , and this gives us the correct form of the above formula as

$$F(0) = BS(T, \sigma(T), \Phi).$$

Now, if such a derivative D is traded or offered on the market at the price $F(0)$, *the formula $F(0) = BS(T, \sigma(T), \Phi)$ can be used to calculate the value $\sigma(T)$.*

We refer to this **value $\sigma(T)$** as the **implied volatility of the underlying asset until expiration T of the derivative D** .

If all market participants were aware of the underlying asset's correct volatility $\sigma(T)$ for the upcoming time interval $[0, T]$ and if all derivatives were correctly priced on the market, i.e. if they were all traded at their fair price, then the underlying value $\sigma(T)$ would consequently be the same for every derivative D ! Put differently: The underlying asset's implied volatility contained in a derivative D would yield the same value $\sigma(T)$, i.e. the implied volatility of the underlying until expiration T , for all derivatives. The fact is: If we have an underlying asset for which there is a reasonably liquid derivatives market and if we have a fixed expiration T , then the implied volatilities $\sigma_D(T)$ calculated from derivatives D with the same expiration T on this underlying generally yield similar but not one common constant value $\sigma(T)$. Thus, there is generally no such thing as **the implied volatility $\sigma(T)$ of an underlying asset**.

As an example, let us look at Figs. 1.16 and 1.17:

These are screenshots taken from the Interactivebrokers Trader Workstation option price charts on 26 October 2018. Table 1.16 shows options with expiration 16 November 2018. The left column shows the implied volatilities of the respective call options, and the last column shows the implied volatilities of the respective put options. We see that, as the strike price K goes up, the implied volatilities of these options fall, from 25.3% to 18.9%. Thus, the implied volatilities resulting from the respective derivatives do not provide a uniform picture; there is no uniform value for the implied volatility of the S&P500 until expiration on 16 November 2018. We get an approximate range for the implied volatility of the S&P500 roughly in the low 20% range, but not an exact single value.

The situation is analogous for options on the S&P500 expiring on 31 January 2019. Again, we see that, as the strike price goes up, volatilities fall, from 22.5% to 16.1%. Again there is no unique value $\sigma(T)$ for the implied volatility of the S&P500. However, the average approximate value resulting from the implied volatilities of the derivatives is significantly lower here than for the shorter-dated options.

Before we contemplate further how these implied volatilities come about, how meaningful they are in terms of predictive significance, and how we should handle

Fig. 1.16 IB Trader

Workstation, options on 26 October 2018 expiring 16 November 2018, with implied volatilities

implied volatility CALLS	16 November 2018 STRIKES	implied volatility PUTS
25.3%	2600	25.3%
25.1%	2605	25.2%
25%	2610	25%
24.8%	2615	24.8%
24.7%	2620	24.6%
24.4%	2625	24.4%
24.2%	2630	24.3%
24.1%	2635	24.1%
23.9%	2640	23.9%
23.7%	2645	23.7%
23.5%	2650	23.5%
23.3%	2655	23.3%
23.2%	2660	23.1%
23%	2665	22.9%
22.8%	2670	22.8%
22.6%	2675	22.6%
22.4%	2680	22.4%
22.2%	2685	22.2%
22%	2690	22%
21.9%	2695	21.8%
21.7%	2700	21.6%
21.5%	2705	21.5%
21.3%	2710	21.3%
21.1%	2715	21.2%
20.9%	2720	20.9%
20.7%	2725	20.8%
20.5%	2730	20.6%
20.3%	2735	20.4%
20.2%	2740	20.2%
20%	2745	20%
19.8%	2750	19.8%
19.7%	2755	19.7%
19.5%	2760	19.5%
19.3%	2765	19.2%
19.2%	2770	19.2%
19%	2775	18.9%

Fig. 1.17 IB Trader

Workstation, options on 26 October 2018 expiring 31 January 2019, with implied volatilities

implied volatility CALLS	31 January 2019 STRIKES	implied volatility PUTS
22.5%	2500	22.5%
22.3%	2510	22.3%
22%	2520	22.1%
21.9%	2525	22%
21.7%	2530	21.9%
21.6%	2540	21.7%
21.4%	2550	21.4%
21.1%	2560	21.2%
20.9%	2570	21%
20.8%	2575	20.9%
20.7%	2580	20.8%
20.5%	2590	20.6%
20.3%	2600	20.3%
20.1%	2610	20.1%
19.9%	2620	19.9%
19.7%	2625	19.8%
19.6%	2630	19.6%
19.4%	2640	19.4%
19.2%	2650	19.2%
19%	2660	19%
18.8%	2670	18.8%
18.6%	2675	18.7%
18.5%	2680	18.6%
18.3%	2690	18.3%
18.1%	2700	18.1%
17.9%	2710	17.9%
17.7%	2720	17.7%
17.6%	2725	17.6%
17.5%	2730	17.5%
17.3%	2740	17.3%
17.1%	2750	17.1%
16.9%	2760	16.9%
16.7%	2770	16.6%
16.6%	2775	16.6%
16.5%	2780	16.5%
16.3%	2790	16.3%
16.1%	2800	16.1%

their various values, let us briefly take a closer look at the calculations of implied volatilities (as shown in the tables above):

For the above call options on the S&P500, the equation

$F(0) = BS(T, \sigma(T), \Phi)$ is:

$$F(0) = s \cdot \mathcal{N} \left(\frac{\log \left(\frac{s}{K} \right) + T \left(r + \frac{1}{2} \cdot \sigma^2(T) \right)}{\sqrt{T} \cdot \sigma(T)} \right) - e^{-rT} \cdot K \cdot \mathcal{N} \left(\frac{\log \left(\frac{s}{K} \right) + T \left(r - \frac{1}{2} \cdot \sigma^2(T) \right)}{\sqrt{T} \cdot \sigma(T)} \right)$$

The implied volatility cannot be explicitly expressed from this equation, but it can be calculated numerically to an arbitrarily high degree of accuracy for specific values of $F(0)$, s , K , T , and r .

For example, let us consider the call option with expiration 16 November 2018 and strike $K = 2650$. The bid/ask quotes at the time of quotation were 75.70 // 76.90, the last traded price was 78.90, and the implied volatility (according to the table) was 23.5%.

The price s of the S&P500 was 2680.43. The option's remaining time to expiration at T , as calculated from 26 October 2018 (or rather, as from half of that trading day) to 16 November 2018 (start of trading day), was approximately 14.5 out of a total of approximately 252 annual trading days, thus approximately $\frac{14.5}{252} = 0.0575397$ years. The overnight US Dollar LIBOR on 26 October 2018 was 2.17425%.

Since it is not clear from the table which option price (bid, ask, mid, last price?) was used to calculate the implied volatility, i.e. exactly which option price was used as $F(0)$ in the equation for implied volatility, we invert the process and start by calculating which price results from the given implied volatility $\sigma(T)$ of 23.5%.

So, we input all given values (e.g. with the help of the corresponding program on our website, see <https://app.lsqt.org/book/implicit-vola>) into the formula

$$s \cdot \mathcal{N} \left(\frac{\log \left(\frac{s}{K} \right) + T \left(r + \frac{1}{2} \cdot \sigma^2(T) \right)}{\sqrt{T} \cdot \sigma(T)} \right) - e^{-rT} \cdot K \cdot \mathcal{N} \left(\frac{\log \left(\frac{s}{K} \right) + T \left(r - \frac{1}{2} \cdot \sigma^2(T) \right)}{\sqrt{T} \cdot \sigma(T)} \right)$$

and obtain the value $F(0) = 78.27$. This value does not match the call option's bid/ask prices in Table 1.16 (75.70 // 76.90) nor its last price (78.90). There may be several reasons for this: IB may have chosen to use a different interest rate r or a

different day convention for its computations, or maybe the implied volatilities in the table had not yet been updated when the screenshot was taken.

Therefore, we are now going to calculate the implied bid volatility and the implied ask volatility from the bid and ask prices using our own parameter choices and numerically solving the equation, and we get the following volatility values:

for Strike 2650: Bid // 22.46 and Ask // 22.95
 (which deviate slightly from the stated value of 23.5%).

For comparison, we also calculate the implied bid/ask volatilities for the 16 November calls with strikes 2600 and 2775 from the respective bid/ask prices of the options and get:

for Strike 2600: Bid // 23.86 and Ask // 24.41
 for Strike 2800: Bid // 18.51 and Ask // 18.82

This confirms once again that the implied volatility, as calculated from the options, decreases as the strike price increases.

How can these observations be interpreted?

If market participants would all coincide in their estimate of the S&P500's future volatility $\sigma(T)$ until time T , if all call and put options were fairly priced, and if it were true that the S&P500 follows a Wiener model, then the implied volatility $\sigma(T)$ for the S&P500 calculated from each option would always result in the same value. This is obviously not the case!

We conclude either:

- (a) The market participants do not coincide in their estimate of the S&P500's future volatility $\sigma(T)$ until time T .
- (b) Or some of the options are overpriced or underpriced in relation to some other options.
- (c) Or the S&P500 does not follow a Wiener model.

We will discuss these three possible conclusions below. Before we do so, however, we want to point out that the implied volatilities that we calculated from the different options give us at least an approximate reference value for the volatility $\sigma(T)$. The implied volatility $\sigma(T)$ of the S&P500 from 26 October 2018 through 16 November 2018 will be around 22% and the implied volatility $\sigma(T)$ of the S&P500 from 26 October 2018 through 31 January 2019 should be around 18%.

So, if we assume that a large number of professional market participants has, at least on average, a roughly accurate estimate of the future volatility of the S&P500 up to time T (remember: It is precisely these market participants that generate this volatility through their trading activities on the financial markets in the first place!), then we can confidently use this reference value, obtained by analysing the liquidly traded derivatives on the respective underlying, as the starting point for pricing derivatives on this underlying. In order to arrive at this volatility estimate, it

is not necessary to study historical prices of the underlying, as this estimate is based exclusively on a snapshot; it is based exclusively on the current prices of derivatives on this underlying.

Let us now briefly look at the three possible explanations given above for the diverging implied volatilities:

ad (a)

For highly liquid derivatives traded by many different market participants on underlying assets that are themselves very liquid and of great interest to many professional traders and arbitrageurs—such as the S&P500 and exchange-traded options on the S&P500—the dynamic system of the financial market, with its many intertwined actions and actors all reacting to one other, leads to a certain equilibrium. Due to these dynamics, the prices of products, especially of highly liquid exchange-traded derivative products, which, for reasons of no-arbitrage arguments, need to be in certain relations to each other, essentially attain the fair prices applicable to them, and these prices then automatically include the applicable implied volatilities of the underlying. (We have no way of knowing, however, in what form they have been included, whether via the Wiener model and the Black-Scholes formula or via other more adequate models.)

ad (b)

We have basically already refuted this possible justification for any such discrepancies in the above comments ad a). In the case of liquid exchange-traded derivatives on liquid underlying assets, we can assume that the price relations are essentially correct.

ad (c)

We already know that the Wiener model exhibits certain simplifications that are not entirely consistent with reality. Among others, we have fat-tail phenomena and certain dependencies between successive returns that are not reflected by the Wiener model. Certain differences between implied volatilities that have been determined from different derivatives (although probably not all) are undoubtedly due to this shortcoming in the model.

To summarize:

If we assume a liquid underlying asset with a liquid market of exchange-traded derivatives, then we can assume that these derivatives are essentially fairly priced. This is essentially—without going into further detail—one of the fundamental statements of the **efficient market hypotheses** postulated by Eugene Fama around 1970 and has been the subject of intense and controversial debate ever since. If we had an exact stochastic model for the underlying and a derivative pricing theory

that provided unique fair derivative prices based on this model and based on the underlying asset's volatility, then we could determine an underlying asset's implied volatility from any derivative and for any available expiration T . This implied volatility could then be used to price all other derivatives on the underlying with the same time to expiration. Otherwise, arbitrage would be possible. Since we do not have an exact model, but only approximate models, such as the Wiener model, different derivatives often yield different implied volatilities for the underlying. However, these implied volatilities are mostly all in a similar range and therefore serve as a benchmark value for pricing derivatives on the underlying in such an approximation model. We will informally refer to this benchmark as the implied volatility of the underlying for the respective expiration (and in the respective model). When we price a derivative using this volatility benchmark, we price it in line with the market. However, analysts are obviously free to make their own subjective assessment of future volatility and use that "individual" volatility to determine an alternative (subjectively accurate) price for the derivative in question.

The implied volatilities that are supplied by a wide variety of financial data providers have virtually always been computed with the Wiener model and the Black-Scholes formula. These implied volatilities from the Black-Scholes universe will be discussed in more detail below.

1.8 Implied Volatilities of Call Options and Put Options with Same Expiration and Strike

Looking at the tables of implied volatilities in the Tables 1.16 and 1.17, it is interesting to note, among other things, that although the implied volatility of the call and put options changes from strike to strike, the implied volatility of a call option always has the same value as the implied volatility of the put option with the same expiration and the same strike.

This equality between the implied volatilities of calls and puts with the same fixed expiration T and the same fixed strikes K must indeed hold (assuming the no-arbitrage principle) irrespective of what model is chosen for the underlying. The only premises we need in order to prove this equality are the following:

- The no-arbitrage principle applies.
- We have some (!) kind of formulas for the call option's and the put option's fair price, $CF(\sigma)$ and $PF(\sigma)$, with a well-defined (injective) dependence on a parameter σ of the underlying (regardless of how this parameter σ is to be interpreted, that is not relevant).

Let $CF(\sigma)$ and $PF(\sigma)$ be the theoretical fair prices of the two options, where σ is given, and let C and P be the actual current prices of the two options. We continue to use the term "implied volatility" for the parameter σ .

We know that the put-call parity equation must hold in an arbitrage-free market regardless of any model assumptions. What's more: The put-call parity equation has

to hold both for the fair prices CF and PF (otherwise, arbitrage would theoretically be possible, which would contradict the definition of fair prices) and for the actual prices C and P (since otherwise arbitrage would in fact be possible). So, for any σ :

$$CF(\sigma) + K \cdot e^{-rT} = PF(\sigma) + S$$

and

$$C + K \cdot e^{-rT} = P + S.$$

By subtracting the two equations, we get

$$CF(\sigma) - C = PF(\sigma) - P$$

Now we let σ_C be the value for which $CF(\sigma_C) = C$, that is, we let σ_C be the call option's current implied volatility. Then, because of the above equation,

$$0 = CF(\sigma_C) - C = PF(\sigma_C) - P \quad \text{i.e.} \quad PF(\sigma_C) = P.$$

Since the put option's implied volatility σ_P is defined such that $PF(\sigma_P) = P$ and since, according to the premise, an option's implied volatility is uniquely given by the option price, it follows that $\sigma_P = \sigma_C$, irrespective of whatever model is chosen.

This observation supports our argument in the previous section that differences between listed implied volatilities of various derivatives with the same expiration on the same underlying assets are likely due above all to the choice of the not entirely perfect Wiener model.

1.9 Volatility Skews, Volatility Smiles, and Volatility Surfaces

In the examples shown in Tables 1.16 and 1.17, we see that implied volatilities fall as the strikes fall in each case. Thus, if we plot these implied volatilities in relation to the strikes and connect them, we get lines sloping downward from left to right. Downward sloping volatility lines (which theoretically, given a correct model for the underlying, should be horizontal) are typical when the underlying assets are stocks or stock indices. When such downward sloping volatility lines occur, they are referred to as **volatility skews**. Figure 1.18 illustrates the volatility skews of Tables 1.16 and 1.17. For better comparison in such illustrations, the curve of implied volatilities is often shown (as in Fig. 1.18) not as a function of the strike price K but as a function of $\frac{K}{S_0}$ (i.e. of the strike expressed as a percentage of the underlying asset's price).

In Fig. 1.19, we take another look at the volatility curves of the S&P500 but this time for even longer periods to expiration and a wider range of strikes. We see—and this too is typical for the implied volatilities of options on stocks or stock indices—that the longer the option's time to expiration, the less pronounced the skew seems

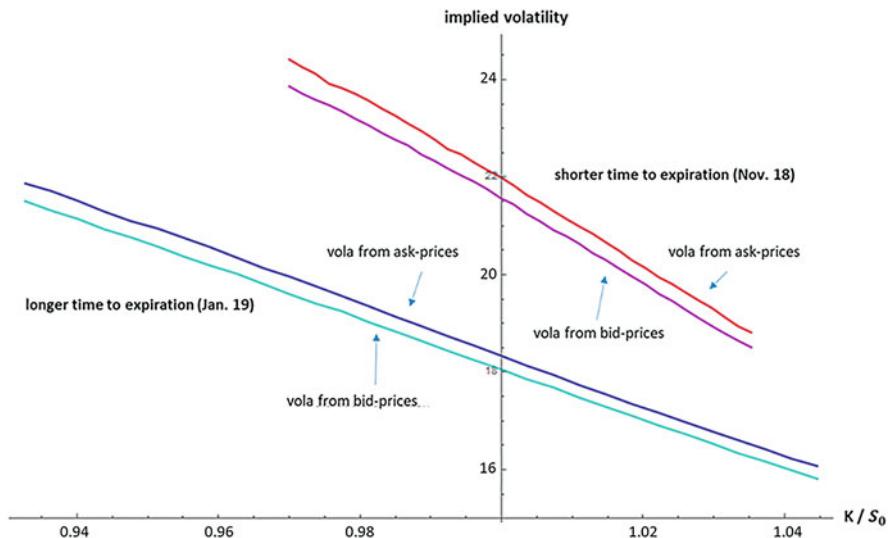


Fig. 1.18 Volatility skews as calculated on 26 October 2018 from the bid/ask prices of S&P500 options shown in Tables 1.16 and 1.17

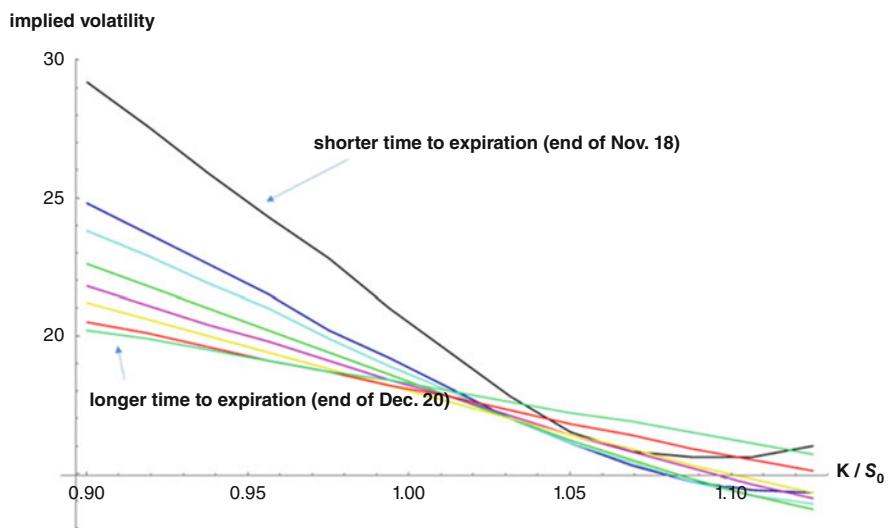


Fig. 1.19 Volatility skews as calculated on 29 October 2018 from the mid-prices of S&P500 options with different expiration dates

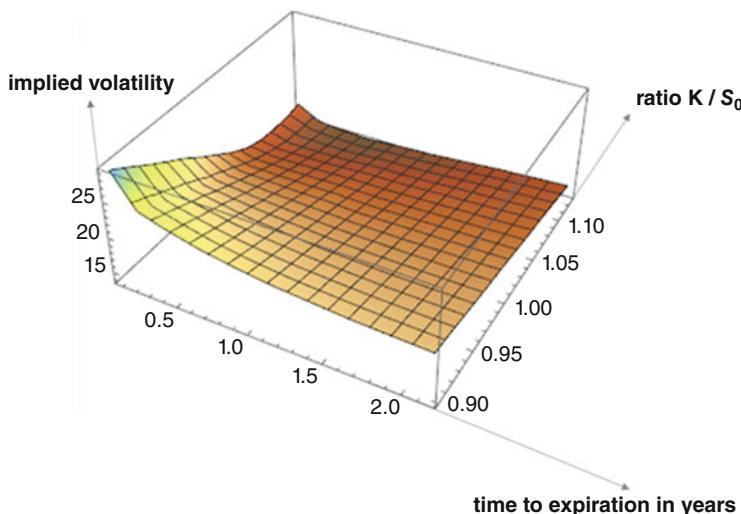


Fig. 1.20 Volatility surface of S&P500 options on 29 October 2018

to be. So, in our example, the implied volatilities decrease with the option's time to expiration. This is especially true in the range where the ratio between the strike price K and the underlying asset's price S_0 is below approximately 1.2. Yet that is not always case! In our example, the volatilities before the date on which these data were selected had been very stable over a long period of time in a very low range of around 12%. It was only in the last 2 weeks before that date that volatility had risen sharply into the temporary range of around 20%. Apparently, the prevailing view in the market at that point—we infer this from the implied volatilities—was that volatilities would stay in a somewhat higher range in the short term and would then, over time, drop back to the usual lower ranges.

The current values of the implied volatilities thus depend both on the time left to expiration and the ratio between the strike price K and the underlying asset's price S_0 . If we plot the implied volatility as a function of these two variables in a three-dimensional coordinate system, we get the underlying asset's current “volatility surface”. The volatility surface of the S&P500 on 29 October 2018 based on the data in Fig. 1.19 is shown in Fig. 1.20.

In the case of currencies, the implied volatility curves tend to have a rather symmetric shape, with implied volatility generally going up at the edges (high and low strike). This is then referred to as a **volatility smile**.

Figure 1.21 lists the quotes for call options on the Euro/US dollar—exchange rate on 29 October 2018. Figure 1.22 shows the corresponding volatility smile.

underlying	type	strike	implied vola	ex-piration	bid	ask
EUR/USD (Euro / US-Dollar...)	Call	1,02	24,48	07.12.2018	10,85	10,89
EUR/USD (Euro / US-Dollar...)	Call	1,03	22,37	07.12.2018	10	10,04
EUR/USD (Euro / US-Dollar...)	Call	1,04	20,46	07.12.2018	9,11	9,15
EUR/USD (Euro / US-Dollar...)	Call	1,05	18,74	07.12.2018	8,15	8,19
EUR/USD (Euro / US-Dollar...)	Call	1,06	17,19	07.12.2018	7,28	7,32
EUR/USD (Euro / US-Dollar...)	Call	1,07	17,19	07.12.2018	6,46	6,5
EUR/USD (Euro / US-Dollar...)	Call	1,08	15,8	07.12.2018	5,62	5,66
EUR/USD (Euro / US-Dollar...)	Call	1,09	13,68	07.12.2018	4,78	4,82
EUR/USD (Euro / US-Dollar...)	Call	1,1	12,87	07.12.2018	3,93	3,97
EUR/USD (Euro / US-Dollar...)	Call	1,11	11,59	07.12.2018	3,09	3,13
EUR/USD (Euro / US-Dollar...)	Call	1,12	10,72	07.12.2018	2,32	2,36
EUR/USD (Euro / US-Dollar...)	Call	1,13	9,97	07.12.2018	1,69	1,73
EUR/USD (Euro / US-Dollar...)	Call	1,14	9,32	07.12.2018	1,2	1,24
EUR/USD (Euro / US-Dollar...)	Call	1,15	8,59	07.12.2018	0,74	0,78
EUR/USD (Euro / US-Dollar...)	Call	1,16	8,33	07.12.2018	0,41	0,45
EUR/USD (Euro / US-Dollar...)	Call	1,17	8,15	07.12.2018	0,21	0,25
EUR/USD (Euro / US-Dollar...)	Call	1,18	8,07	07.12.2018	0,09	0,13
EUR/USD (Euro / US-Dollar...)	Call	1,19	8,22	07.12.2018	0,04	0,08
EUR/USD (Euro / US-Dollar...)	Call	1,2	8,76	07.12.2018	0,01	0,05
EUR/USD (Euro / US-Dollar...)	Call	1,21	9,15	07.12.2018	0,01	0,05
EUR/USD (Euro / US-Dollar...)	Call	1,22	9,93	07.12.2018	0	0,04
EUR/USD (Euro / US-Dollar...)	Call	1,23	11,04	07.12.2018	0	0,04
EUR/USD (Euro / US-Dollar...)	Call	1,24	12,09	07.12.2018	0	0,04
EUR/USD (Euro / US-Dollar...)	Call	1,25	13,14	07.12.2018	0	0,04

Fig. 1.21 Quotes for call options on the Euro/US dollar exchange rate on 29 October 2018 at the then current exchange rate of 1.138

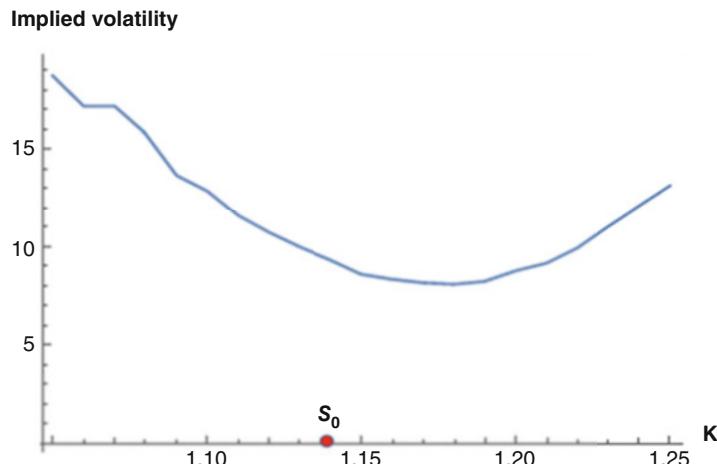


Fig. 1.22 The implied volatilities listed in Table 1.21 form a volatility smile

1.10 Inferences From Implied Volatilities About the Market-Anticipated Distribution of the Underlying Asset's Price

What causes volatility smiles and volatility skews? Why isn't there just one single implied volatility per underlying and time horizon? One of the reasons—as said before—is certainly that the assumption of the Wiener model (with independent normally distributed returns of the underlying) is not correct.

Assuming a Wiener model with a fixed value σ for volatility and with respect to a risk-neutral measure, the option prices that we obtain—for a fixed time to expiration T —are the discounted expected payoffs. But if we look at real option prices, we find—if we maintain the hypothesis of a risk-neutral Wiener model—that these option prices cannot be the discounted expected payoff for a fixed volatility σ . This can only be the case if we assume different values for σ . Yet the underlying can only have *one* current volatility σ ! Thus, if we want option prices to be interpreted as discounted expected payoffs, we have to abandon the assumption of the Wiener model! But what other model should then be used for the underlying price?

So, the question is as follows: If the real option prices are the discounted expected payoff values with respect to a fixed (risk-neutral) distribution for the underlying price, what does this fixed distribution look like?

For Example

Let us look again at the prices of the call options on the S&P500 of 26 October 2018 with expiration 31 January 2019 as shown in Table 1.17. To do this, let us focus on the mid-prices and the strikes between 2500 and 2800 that are divisible by 10. We have compiled the mid-prices that we are going to work with in the following in the Table 1.5 below:

It is important to note here that we cannot achieve an entirely exact representation of the probability distribution we are looking for, since we only have the option prices for a specific range of strikes (increments of ten from 2500 to 2800) (see Fig. 1.23).

Table 1.5 Mid-prices and strike prices of call options on the S&P500 of 26.10.2018 with an expiration 31.01.2019

Strike	2500	2510	2520	2530	2540	2550	2560	2570
Mid-price	236.15	228.15	220.2	212.2	204.45	196.85	189.2	181.7
Strike	2580	2590	2600	2610	2620	2630	2640	2650
Mid-price	174.3	166.9	159.7	152.6	145.7	138.75	132	125.4
Strike	2660	2670	2680	2690	2700	2710	2720	2730
Mid-price	118.9	112.6	106.45	100.35	94.6	88.8	88.3	77.95
Strike	2740	2750	2760	2770	2780	2790	2800	
Mid-price	72.8	67.85	63.05	58.4	54.05	49.8	45.85	

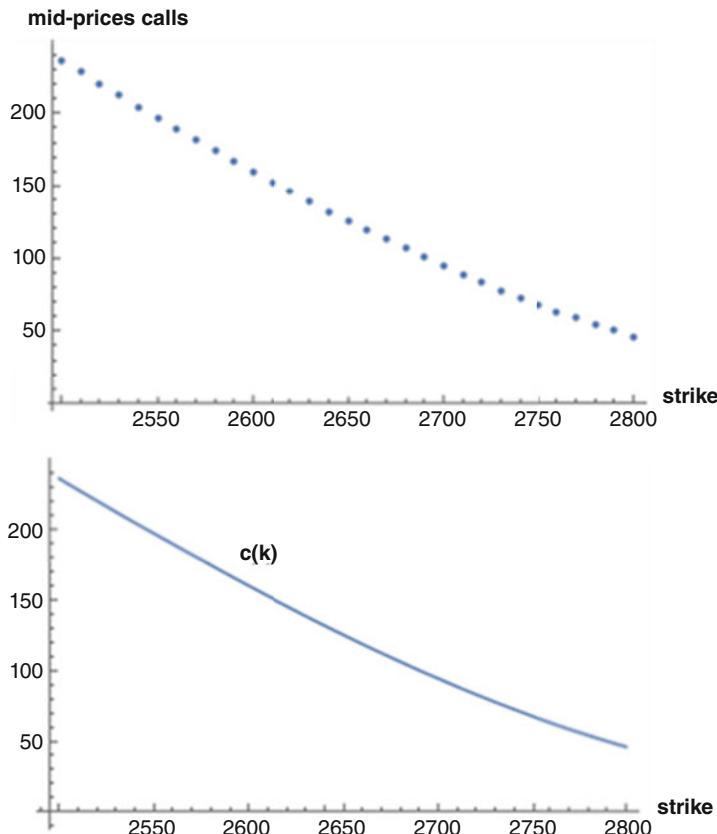


Fig. 1.23 Mid-prices of call options expiring 31 January 2019 as on 26 October 2018 (above) and interpolated (below)

However, it appears that option prices can be interpolated quite well by a smooth function $C(K)$ (see Fig. 1.23, above), which is defined for *all strikes* K (in the relevant range). In order for us to be able to really work with this interpolation function and obtain really useful and meaningful results, we need a representation for C that goes substantially beyond the defined range of 2500 to 2800 and that satisfies certain properties that we will specify below.

In Fig. 1.24, we also plotted the implied volatilities that were calculated from the mid-prices. The blue line indicates the average value of those implied volatilities, which is 18.8%.

Purely intuitively, the shape of the volatility skew suggests the following:

- The volatility line is skewed, so the assumption of a normal distribution for the returns is not correct.

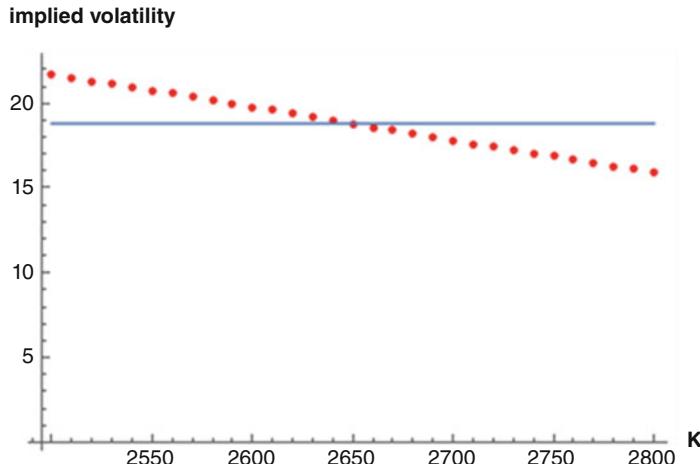


Fig. 1.24 Implied volatility of call options expiring 31 January 2019 as on 26 October 2018

- (*) The assumption of normal distribution of the returns leads to higher volatilities on the “left” than on the “right”. This means: The real option prices for small strikes are too high for an assumed normal distribution and the real option prices for high strikes are too low for an assumed normal distribution.
- (*) in the case of small strikes:
Under the actual distribution of price returns, the expected payoff would have to be smaller than is the case under the assumption of normal distribution.
- The expected payoff of a call option is smaller where the probability of small underlying prices is higher.
- This means: The actual distribution of price returns allows a higher probability of small prices than the normal distribution does.
- (*) in the case of large strikes:
Under the actual distribution of price returns, the expected payoff would have to be larger than is the case under the assumption of normal distribution.
- The expected payoff of a call option is larger where the probability of large underlying prices is higher.
- This means: The actual distribution of price returns allows a higher probability of large prices than the normal distribution does.
- To summarize: The real actual distribution of price returns maps higher probabilities to very small and very large events than the normal distribution does.
- To summarize: **A volatility skew (a curve sloping downward to the right) indicates heavy tails in the distribution of returns.**

Now, can the actual distribution of returns be calculated unambiguously from the volatility line (from its interpolant), and if so, how?

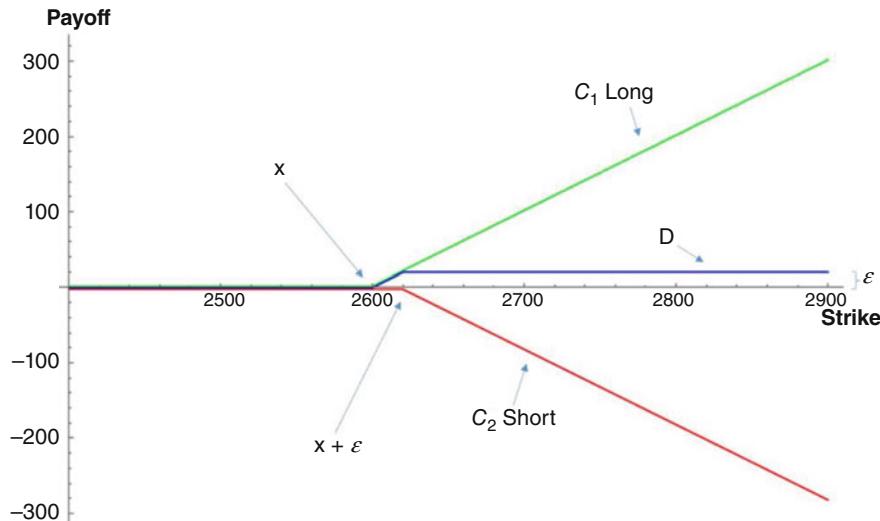


Fig. 1.25 Payoff of portfolio D

To answer this question, we look at two special call options: One, call (C_1) with strike x , and another call (C_2) with strike $x + \varepsilon$ where ε denotes a very small positive number. We then look at the portfolio D consisting of the first long call option and the second short call option. The payoff Φ from D at time T has the shape as shown in blue in Fig. 1.25.

The fair price of the portfolio D is $C(x) - C(x + \varepsilon)$ with the function C from 1.23.

On the other hand (according to our assumption above), the price of D is also given by

$e^{-rT} \cdot E(\Phi(S(T)))$, where E denotes the expected value of the actual distribution of returns of the underlying price. However, as we can see from Fig. 1.25,

$$E(\Phi(S(T))) < \varepsilon \cdot W(S(T) > x) \text{ und } E(\Phi(S(T))) > \varepsilon \cdot W(S(T) > x + \varepsilon).$$

Here, W denotes the probability with respect to the actual distribution of returns. Hence:

$$\begin{aligned} \varepsilon \cdot W(S(T) > x + \varepsilon) \cdot e^{-rT} &< e^{-rT} \cdot E(\Phi(S(T))) = \\ &= C(x) - C(x + \varepsilon) < \varepsilon \cdot W(S(T) > x) \cdot e^{-rT}, \end{aligned}$$

that is,

$$W(S(T) > x + \varepsilon) < \frac{1}{\varepsilon}(C(x) - C(x + \varepsilon)) \cdot e^{rT} < W(S(T) > x),$$

and if we assume that $W(S(T) > x)$ is a continuous function in x and we let ε go to 0, we get

$$W(S(T) > x) = -e^{rT} \cdot C'(x),$$

where C' denotes the derivative of C with respect to the strike price x .

But what exactly is it that we want to calculate? We want to know what the distribution of the annualized continuous returns z of the underlying price relative to the actual distribution of the underlying price looks like. In other words: We want to find the distribution function of the annualized returns, that is, find out what value the probability $W(z < y)$ has for any y . Now, we have (where z' denotes the return for $[0, T]$):

$$\begin{aligned} W(z < y) &= W\left(\frac{z'}{T} < y\right) = W(z' < yT) = W\left(S_0 \cdot e^{z'} < S_0 \cdot e^{yT}\right) = \\ &= W\left(S_T < S_0 \cdot e^{yT}\right) = 1 - W\left(S_T > S_0 \cdot e^{yT}\right) = \\ &= 1 + e^{rT} \cdot C'\left(S_0 \cdot e^{yT}\right). \end{aligned}$$

Thus, we have an explicit representation of the distribution function of the actual price return distribution. The density f of the distribution is now simply obtained by differentiating the distribution function with respect to y . Hence:

$$f(y) = S_0 \cdot T \cdot e^{yT} \cdot e^{rT} \cdot C''(S_0 \cdot e^{yT})$$

One might think that the second derivative C'' of the interpolating function C can be used to explicitly calculate the density function f and have it plotted. However, on testing this, we quickly find out that the result is actually extremely unstable and is rarely of any real use.

In order for us to obtain really useful and meaningful results, C needs to satisfy certain requirements:

- First of all, C must of course be differentiable at least twice on $(0, \infty)$; we already assumed this in our above calculations.
- The function f is supposed to be a density function, which means that f must always be positive, and this is satisfied if and only if C'' is always positive. C therefore has to be continuously skewed to the left.
- $W(z < y)$ (as a function of y) is a distribution function. Hence, $\lim_{y \rightarrow -\infty} W(z < y) = 0$ and $\lim_{y \rightarrow +\infty} W(z < y) = 1$. So we get:

$$\lim_{y \rightarrow -\infty} \left(1 + e^{rT} \cdot C'\left(S_0 \cdot e^{yT}\right)\right) = 0 \text{ und } \lim_{y \rightarrow +\infty} \left(1 + e^{rT} \cdot C'\left(S_0 \cdot e^{yT}\right)\right) = 1,$$

that is:

$$\lim_{x \rightarrow 0^+} C'(x) = -e^{-rT} \text{ and } \lim_{x \rightarrow +\infty} C'(x) = 0.$$

In order for us to determine the implied distribution function explicitly for our example, we need an analytically given approximation function C that satisfies the above requirements. In the following, we deliberately choose an approximant that is not really ideal, with a view to emphasizing that the determination of the distribution function is relatively unstable and that the result of the calculation can react very strongly and often in an undesirable manner to small changes in the approximation function C .

In Fig. 1.24, we illustrated that the average implied volatility of the options we were looking at was about 18.8%. If all options had been priced according to this average volatility, the options' price line would have looked like the green line in Fig. 1.26 (compared to the actual option prices in blue).

Figure 1.27 shows (assuming constant volatility of 18.8%) the associated density function (left) and associated distribution function (right).

For the twice-differentiable approximation function C , we now chose the (magenta-colored) curve shown in Fig. 1.28. It consists of a quadratic function in the range 0 to 2800 that is seamlessly followed at $x = 2800$ by a function of the form $\frac{a}{x^b}$ with suitable parameters a and b . We have deliberately chosen C such that C' is once continuously differentiable in $x = 2800$ but the second derivative in $x = 2800$ does not exist. The other required properties are satisfied by C .

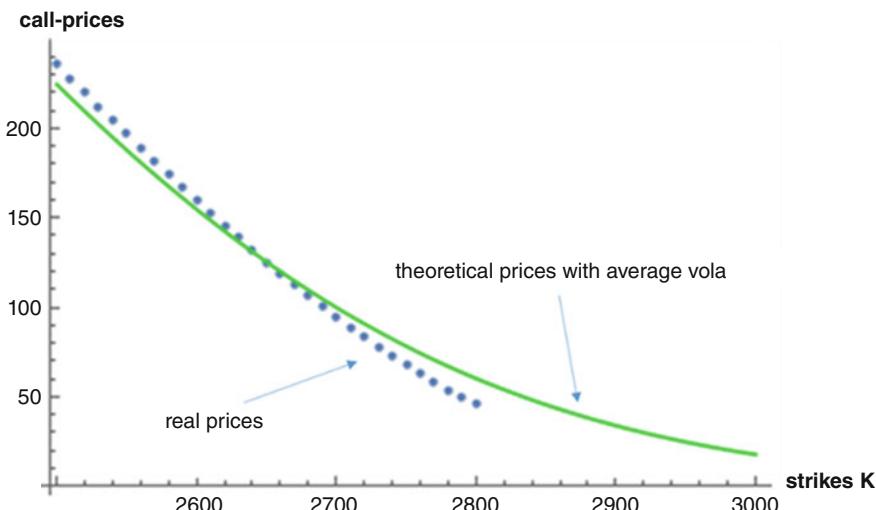


Fig. 1.26 Actual prices of call options (blue) and theoretical prices of call options with constant average volatility

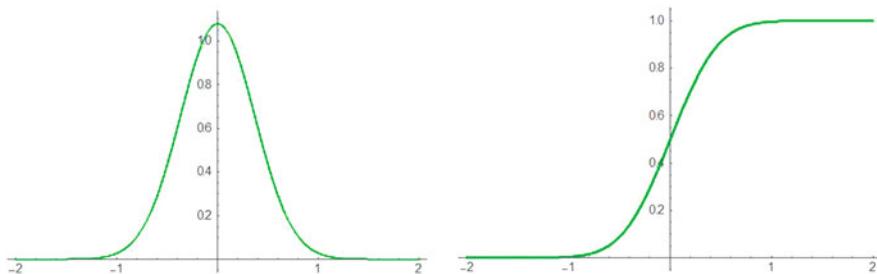


Fig. 1.27 Density function and distribution function assuming constant volatility of 18.8%

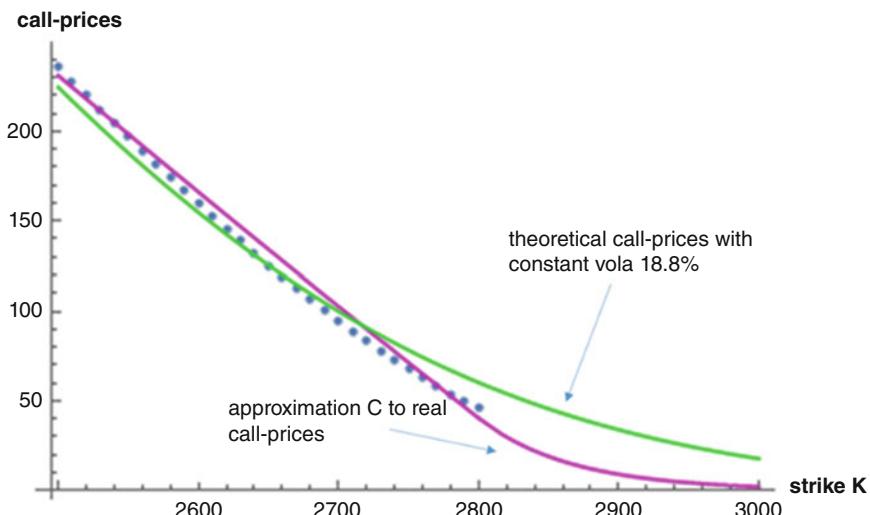


Fig. 1.28 Real call option prices (blue), theoretical call option prices at constant volatility (green), approximation C to real call option prices (magenta)

In Fig. 1.29, the red lines represent the implied density function (above), given by the approximation function C and the distribution function (below) compared with the density and distribution functions of the normal distribution with volatility $\sigma = 18.8\%$ and, of course, with the risk-neutral trend $\mu = r - \frac{\sigma^2}{2}$. We obtain the density function and the distribution function for C simply by using the formulas derived above for $f(y)$ and for probability $W(z < y)$, respectively.

Although the implied density function for C and the implied distribution function for C satisfy all the properties required for density and distribution functions, the density function is not continuous at a point x_0 , and the distribution function is not differentiable at this point x_0 . x_0 is of course precisely the return corresponding to a resulting price of 2800 points at time T . Looking at the density function, you can actually see a “higher middle part” and heavy tails on the left edge.

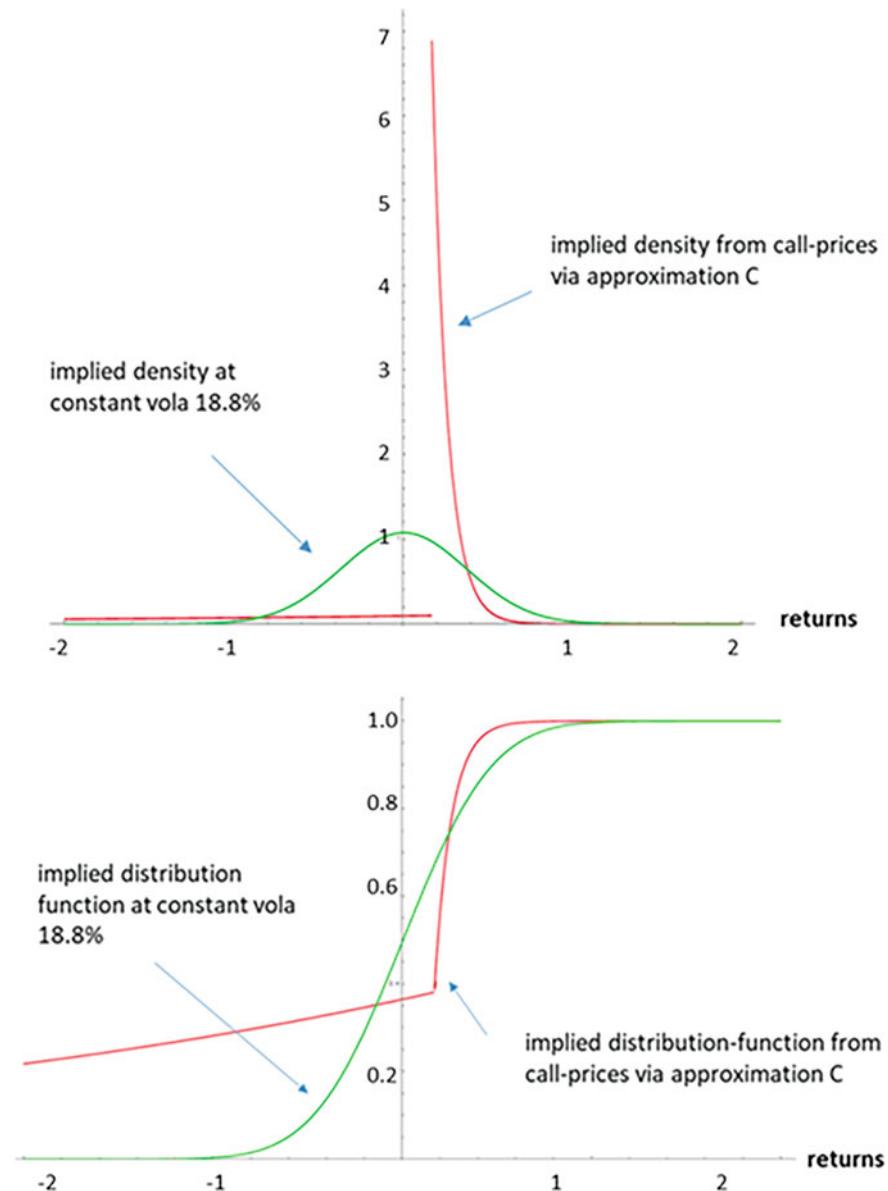


Fig. 1.29 Implied density functions (above) and distribution functions (below) for theoretical call option prices at constant volatility 18.8% (green) and for call option prices given by the approximation function C (red)

1.11 Volatility Indices

As we have seen, there is normally no such thing as *the* implied volatility of a financial product. The implied volatility depends on the time to expiration of the derivatives from which it is calculated and (for options) also on the strike prices. So, if we need to know the implied volatility of an underlying asset in order to obtain a derivative's fair price, we have some leeway as to which value we actually want to use for that implied volatility.

For some major stock indices, there are volatility indices that are calculated and published on a regular basis. These are averages which are calculated in a certain way over the implied volatilities of various options on the index and serve as a reference for the implied volatility of the index. Many market participants observe them very closely and use them as reference for implied volatility.

Some of the best known volatility indices are:

VIX	Volatility index of the S&P500 stock index
VXN	Volatility index of the NASDAQ100 stock index
VXD	Volatility index of the DowJones Industrial Average stock index
VDAX-NEW	Volatility index of the DAX (successor to the VDAX)
VSTOXX	Volatility index for the Euro Stoxx 50 stock index
VSMI	Volatility index for the SMI, the Swiss Market Index

The calculation methods for the various volatility indices differ and depend significantly on the options market they relate to. In the early years, these volatility indices were generally calculated and published in the following way:

A certain period was specified which the volatility indexed by the index should refer to (for example, 30 days). Then, from the range of traded options with the longest time to expiration less than 30 days and the shortest time to expiration greater than 30 days, all implied volatilities were determined using the Black-Scholes formula. The current value of the volatility index was then obtained as a specifically weighted average of these implied individual volatilities. From the turn of the millennium, a new computation method was introduced for most volatility indices which is largely model-independent, that is, it does not rely on the calculation of implied volatility using the Black-Scholes formula. (The method was essentially developed by Goldman-Sachs.) Another advantage of this method is that the volatility index calculated in this way becomes in itself a financial product that can be hedged by the underlying options. We will explain the method (at least to some extent) in Sect. 1.18 using the VIX as an example.

But first, let us provide a brief overview of the development of the most important volatility indices (VIX, VXN, VDAX-NEW) over the years.

The VIX is calculated and published by the CBOE (Chicago Board Options Exchange). From 1993 onwards, it was defined and computed by means of the old method mentioned above, using the implied volatility of individual options on the S&P100. Since 2003, as explained in Sect. 1.18, it has been defined and

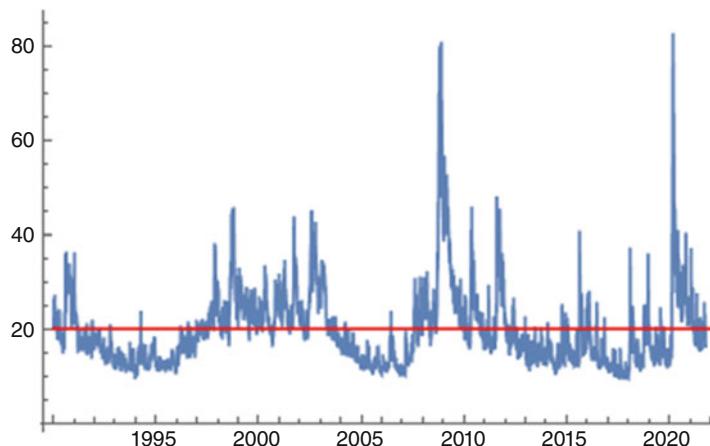


Fig. 1.30 VIX November 1990–October 2021

calculated directly using S&P500 option prices. However, back-calculations of the VIX according to the new method have been made to as far back as 1990 (and for the VXO index, which was based on S&P100 options, there are back-calculations based on the old method going back to 1986). These values are extremely useful in certain back-testing scenarios. In the following analyses, we use the (partly back-calculated) data of the VIX in its new version from 1990. The price movements of the VIX are shown in Fig. 1.30.

We see a picture that is similar in principle to that of the historical volatilities of the S&P500 illustrated in Sect. 1.4. The values are mainly in the range between 10% and 40% with occasional outliers above 40% and two massive peaks in the 80% range at the time of the financial crisis in late 2008 and during the first covid wave in early 2020. The long-term mean since 1990 is 19.48% (red line in Fig. 1.30).

Visually, the VIX oscillates around this mean. Long periods of below-average VIX values can be seen in 1991–1997, 2004–2007, and 2012–2015. Periods of constantly high VIX values are, for example, 1998–2003 and 2008–2012. The VIX values in the range above 80% in 2008 and 2020 are singular events in this picture. In this context, it is therefore of interest to compare the values of the VIX with the values of the VXO from 1990 to 2021 and to further back-calculate the VXO to 1986.

In Fig. 1.31, the two indices, the VXO from 1986 onward (red) and the VIX from 1990 onward (blue), have been superimposed. We see an almost synchronous trend of the two indices from 1990 onward. It can therefore be assumed that the values of the VXO in the period from 1986 to 1990 also provide a good indication of what the values of the VIX would have been in this period. And here, of course, we immediately notice an extreme value of over 150% for the VXO, which was reached on 19 October 1987. This trading day with its massive price drop was already analysed earlier in the book. The volatility value is almost twice as high

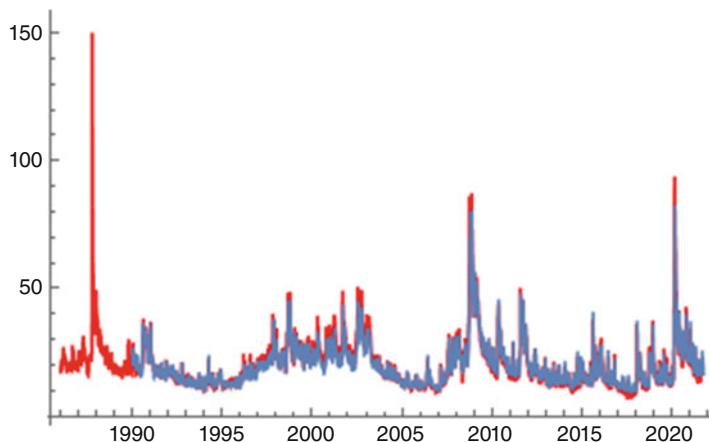


Fig. 1.31 VXO from 1986 (red) and VIX from 1990 (blue)

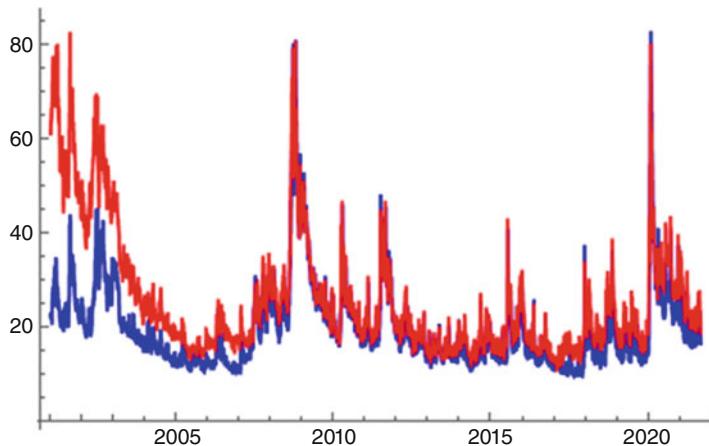


Fig. 1.32 VIX (blue) and VXN (red) from 31 January 2001 to 14 October 2021

as the highest values during the financial crisis in 2008 and in early 2020. It is also interesting to note how quickly the VXO reverts to a period of average to low values after 19 October 1987.

Another peculiarity becomes apparent when comparing the VIX with the VXN, the volatility index of the NASDAQ100: From around 2006 onwards, we notice a certain synchronicity (or at least parallelism) between the two trends in the following Fig. 1.32. In the period from 2001 to 2003, however, the VXN consistently shows significantly higher values than the VIX. This is quite certainly due to the high variability in stock prices during this period of new technology stocks, which have a higher weighting in the NASDAQ than in the S&P500.

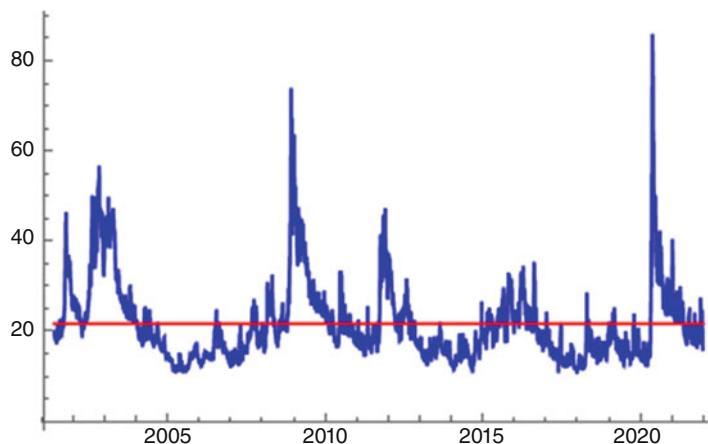


Fig. 1.33 VDAX-NEW, May 2001–October 2021 (partly back-calculated)

Figure 1.33 also shows the VDAX-New from May 2001 to October 2021. The VDAX-New has been calculated and published every minute since 18 April 2005. The values prior to 2005 are back-calculations. The average value of the VDAX-New in the time range shown below is 21.66% (red line in Fig. 1.33). In the same period, the VIX has an average value of 19.48%, which means that the DAX exhibits a slightly higher variability than the S&P500.

In addition to the volatility indices of major stock indices mentioned so far, there is a whole plethora of more specialized indices for various underlying assets. For example, the following page on the CBOE website <http://www.cboe.com/products/vix-index-volatility/volatility-indexes> lists all the volatility indices calculated and published by the CBOE. Among these indices you will find, for example, are the CBOE S&P500 9-Day Volatility Index and the CBOE S&P500 6-Month Volatility Index, which are constructed in exactly the same way as the VIX but are based on shorter-dated and longer-dated SPX options, respectively. You will also find indices on the volatility of prices that depend on interest rate movements, such as the CBOE/CBOT 10-year US Treasury Note Volatility Index or the CBOE Interest Rate Swap Volatility Index (which is calculated from the prices of swaptions on the 10-year US dollar interest swap).

Furthermore, there are volatility indices for the futures on oil and gold prices, on foreign exchange rates to the dollar, and even of individual NYSE stocks. A volatility index that goes one step deeper is the VVIX, the volatility index of the VIX. It is calculated from options on the VIX and, as such, tracks the volatility of the volatility of the S&P500. For some of these volatility indices, there exists a future and/or option market. In a later chapter, we will briefly discuss VIX futures and VIX options and how they are traded.

Financial information systems obviously provide information on implied volatilities for a wide variety of other prices, as well as historical data and tools for technical

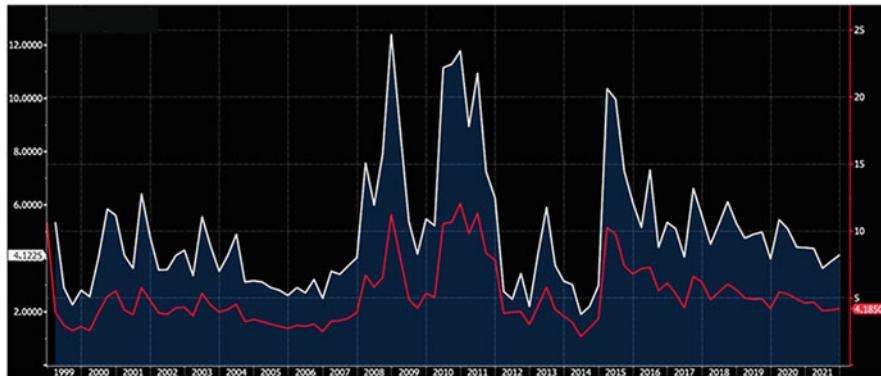


Fig. 1.34 Implied 1-month volatility of the EUR CHF exchange rate from 1999 to 2021 (white) compared with the historical 3-month volatility (red) (source: Bloomberg)

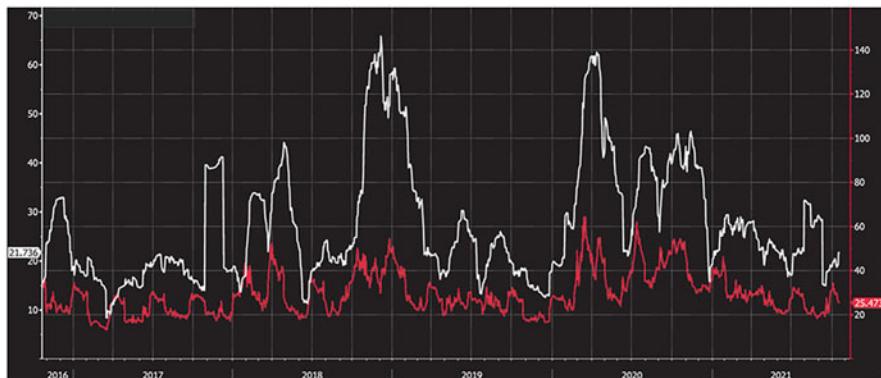


Fig. 1.35 Comparison of 1-month implied volatility (red) and historical 30-day volatility (white) of Amazon stock from 2016 to 2021 (source: Bloomberg)

analyses of all kinds. Figures 1.34 and 1.35, for example, show sample screenshots of Bloomberg pages on volatilities of the EUR CHF exchange rate and the Amazon stock.

1.12 Basic Properties of the VIX

In the following, we will focus exclusively on the VIX, the volatility index of the S&P500. In this section, we take another look at some basic metrics of the VIX. In the following chapters, we will then discuss the relationship between the implied volatility represented by the VIX and the historical volatility and the realized volatility, as well as the relationship between movements in the VIX and the SPX.

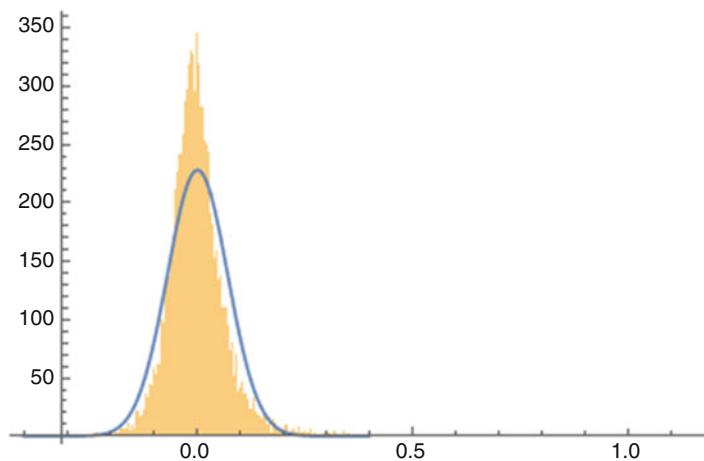


Fig. 1.36 Frequency distribution of the daily returns of the VIX from 1990 to October 2021 compared to the adapted density of the normal distribution with the same mean and same standard deviation (blue)

Compiled in a frequency histogram, the (discrete) daily returns of the VIX show the following picture (Fig. 1.36):

The mean of the daily returns of the VIX is 0.0022 (i.e. 0.22%), the standard deviation of the daily returns is 0.0700 (i.e. 7%). As we can see, the standard deviation is much higher than is generally the case for daily returns on stock prices. Changes in the implied volatility of 6%–7% per day are thus well within the usual range. Furthermore, we can see the (rare, but by no means marginal) occurrence of daily returns in the range of plus/minus 20% (approx. 3 times the standard deviation).

For comparison, we have once again graphed the distribution of daily returns of the S&P500 in a frequency histogram and the corresponding (adapted) density of the normal distribution in Fig. 1.37. The mean of the daily returns of the S&P500 is 0.00034 (i.e. 0.034%); the standard deviation of the daily returns is 0.010 (i.e. 1.0%). Here, daily returns of plus/minus 3% can be found in the range of approximately 3 times the standard deviation.

The empirical returns distribution of the SPX deviates—visibly—from the normal distribution in the usual way. More conspicuous, however, is the deviation of the empirical returns distribution of the VIX from the normal distribution. This significant deviation of the empirical returns distribution of the VIX from the normal distribution of mutually independent values is also evident from the following fact:

As noted further above, the standard deviation of daily returns for the VIX is 0.0700, hence much higher than the 0.01 standard deviation of daily returns for the S&P500. If we hypothetically assume a normal distribution with mutually independent returns for both distributions, we get normalized annual volatility for

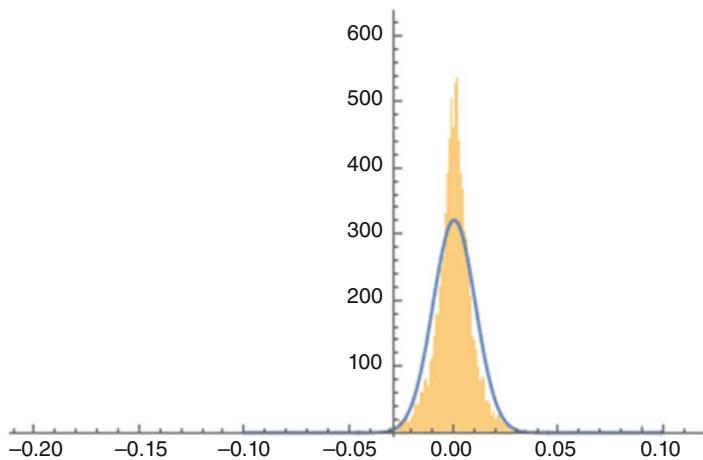


Fig. 1.37 Frequency distribution of the daily returns of the S&P500 from 1990 to October 2021 compared to the adapted density of the normal distribution with the same mean and same standard deviation (blue)

the VIX of $0.07 \times \sqrt{252} = 1.111 (= 111.1\%)$ and normalized annual volatility for the SPX of $0.01 \times \sqrt{252} = 0.1587 (= 15.87\%)$.

If we normalized the VIX and the SPX values to an initial value of 100 and superimposed the two charts (Fig. 1.38), the purely visual impression would not suggest such a large difference in their fluctuations.

However, it is imperative that the chart not be looked at as a whole; instead, one needs to recognize that the high volatility of the VIX occurs primarily in the micro area, that is, on the basis of very high daily fluctuations. This becomes quite clear if we take only a small section of Fig. 1.38 (see Fig. 1.39 with the data from 1990, where the much higher volatility of the VIX is clearly shown).

However, if we direct our attention from daily returns to real (!) annual returns (see Fig. 1.40), it is absolutely impossible to see any volatility in the VIX, which should be six times higher than the volatility of the SPX.

If we calculate the actual volatilities of the annual returns of the VIX and the SPX, we get the following values:

Annual volatility of SPX returns: 16.29% (roughly equivalent to the value of 15.87% as extrapolated from the daily returns)

Annual volatility of VIX returns: 46.53% (differs widely from the value of 111.1% as hypothetically extrapolated from the daily returns)

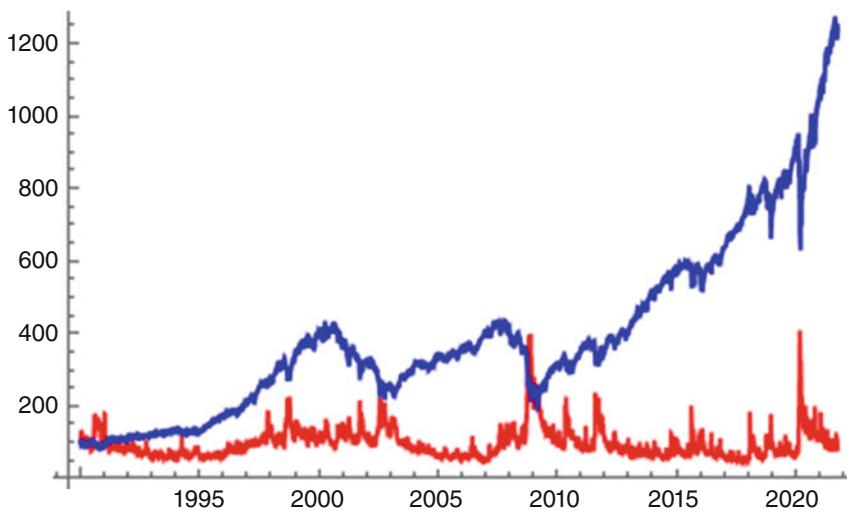


Fig. 1.38 Normalized SPX (blue) and VIX (red) in percent from 1990 to October 2021 based on daily data

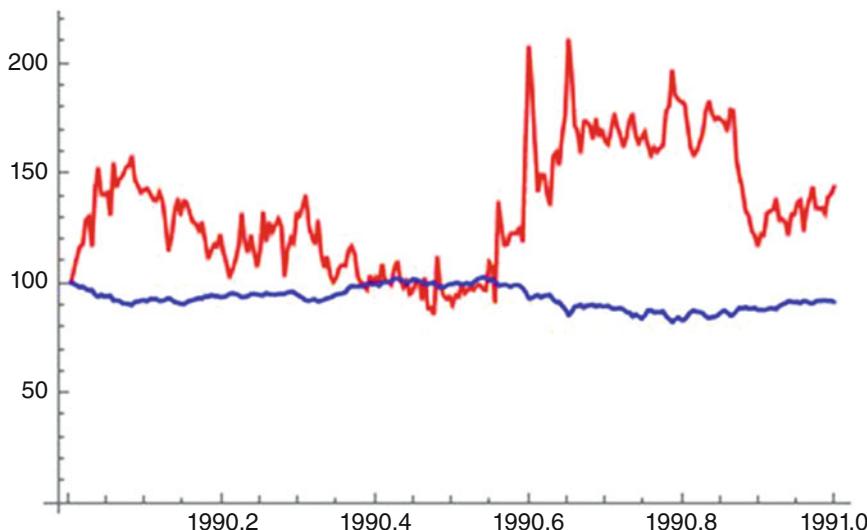


Fig. 1.39 Normalized SPX (blue) and VIX (red) in percent in 1990 based on daily data

Almost three times the volatility of the VIX returns relative to the SPX returns is visually plausible (based on Fig. 1.40). In any case: For VIX returns, the annual volatility can never be approximated from daily returns by usual normalization (as in the case of independent normally distributed variables).

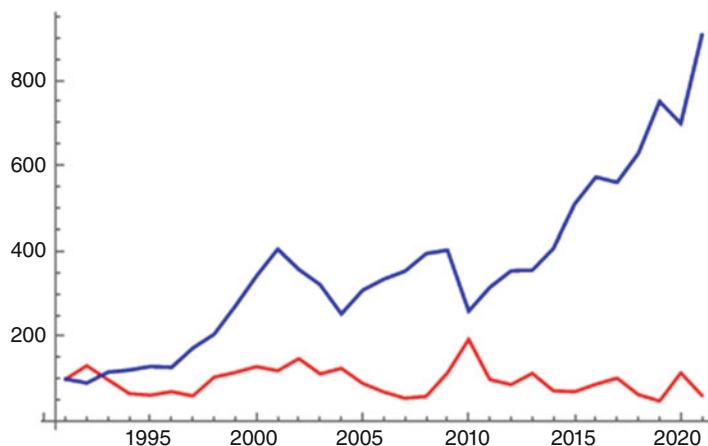


Fig. 1.40 Normalized SPX (blue) and VIX (red) in percent from January 1990 to January 2021 based on annual data

1.13 Relation and Correlations Between VIX and SPX

It is not uncommon to hear opinions along the following lines (and in fact, until recently, the German Wikipedia entry on the “CBOE Volatility Index” even specifically said so):

The VIX has an inverse correlation to the S&P500. When the volatility of the VIX goes up, the S&P500 usually falls. When the volatility of the VIX falls, the S&P500 goes up.

A similar argument was presented to the author of this book as a reproach by an expert in a court hearing in which the author had been summoned to appear as witness. The expert’s argument was as follows:

At the moment when the investment was made in the SPX, the VIX was so high that every market observer should reasonably have concluded that the SPX would now fall. Buying the SPX (knowing that losses were certain to occur) was therefore grossly negligent.

It is correct, as we will see below, that there is a certain correlation between movements of the S&P500 and the VIX. However, it is not true that movements of the VIX have a significant and longer-term impact on subsequent movements of the SPX; what is true is that certain movements of the SPX (especially a strong decline in the SPX price) often have at least a short-term influence on the immediately subsequent movement of the VIX. Sharp price drops in the SPX, for example, make derivatives traders nervous: Hedges (i.e. put options in particular) become more expensive as short positions in put options seem riskier. The implied volatility of put options goes up and, at the same time—for reasons of the no-arbitrage principle—the implied volatility of the call options also goes up. Implied volatility thus reacts to price movements of the underlying asset and not vice versa.

Table 1.6 Correlations of different VIX parameters with different SPX parameters for various periods of time

	1990–Nov. 2018	2000–Nov. 2018	2010–Nov. 2018
Closing price VIX/ Closing price SPX same day	-0.18	-0.50	-0.58
Closing price VIX/ Return SPX same day	-0.12	-0.126	-0.174
Closing price VIX/ Return SPX previous day	-0.106		
Closing price VIX/ Return SPX following day	0.034		
Return VIX/ Return SPX same day	-0.696	-0.717	-0.79
Return VIX/ Return SPX previous day	0.052		
Return VIX/ Return SPX following day	0.023		

Incidentally, the German Wikipedia entry has been changed at our instigation and now reads:

If the S&P500 falls, the VIX usually rises. If the S&P500 rises, the VIX usually falls.

With this in mind, let us look at some correlations summarized in the Table 1.6. Correlations of different VIX parameters with different SPX parameters:

We see that the strongest negative correlations (of around -0.7) is observable between same-day daily returns of the SPX and of the VIX. Significant negative correlations can also be seen (especially from the turn of the millennium, in the range of about -0.5) between same-day daily closing prices of the SPX and of the VIX. Less markedly negative (-0.1 to -0.2), yet still clearly noticeable, are the correlations between the closing price of the VIX and the daily return of the SPX on the same trading day.

No correlations can be seen between values or returns of the VIX and values or returns of the SPX for the respective following day.

Let us look at all trading days from 2 January 1990 to 4 November 2018, on which the **SPX fell by more than 1%**. This was the case on 875 out of 7270 trading days.

On 834 of those 875 trading days, the VIX went up.

Let us look at all trading days from 2 January 1990 to 4 November 2018, on which the **SPX fell by more than 2%**. This was the case on 249 out of 7270 trading days.

On 243 of those 249 trading days, the VIX went up.

Let us look at all trading days from 2 January 1990 to 4 November 2018, on which the **SPX rose by more than 1%**. This was the case on 936 out of 7270 trading days.

On 869 of those 936 trading days, the VIX fell.

Let us look at all trading days from 2 January 1990 to 4 November 2018, on which the **SPX rose by more than 2%**. This was the case on 219 out of 7270 trading days.

On 209 of those 219 trading days, the VIX fell.

So, almost consistently, when the SPX changes sharply, the values of the SPX and the VIX move in opposite directions. Surprisingly, however, there were a few days when, for example, the SPX fell by more than 2% and the VIX did not go up on that day. Let us take a look at the last trading day on which this happened: It was 14 April 2009, towards the end of the most intense phase of the financial crisis

On that day, the SPX had fallen from 858.73 to 841.5 points. Before that, however, the SPX had risen quite strongly, for example, from 787.53 points on 30 March 2009 to 858.73 points on 13 April 2009 (i.e. by more than 9%) as displayed in the Table 1.7. The VIX was still very high, at around 40%. Apparently, the temporary decline of the SPX on 14 April (the SPX continued its recovery again on the following days) was indeed interpreted by the market as just a short hiccup in a beginning rebound (price declines of 2% per day in the course of the financial crisis were no longer particularly noticeable events). So, in this case, the VIX practically

Table 1.7 SPX and VIX values from 30 Mar 2009 to 17 Apr 2009

Date	VIX	SPX
30 Mar 2009	45.54	787.53
31 Mar 2009	44.14	797.87
01 Apr 2009	42.28	811.08
02 Apr 2009	42.04	834.38
03 Apr 2009	39.70	842.50
06 Apr 2009	40.93	835.48
07 Apr 2009	40.39	815.55
08 Apr 2009	38.85	825.16
09 Apr 2009	36.53	856.56
13 Apr 2009	37.81	858.73
14 April 2009	37.67	841.50
15 Apr 2009	36.17	852.06
16 Apr 2009	35.79	865.30
17 Apr 2009	33.94	869.60

Table 1.8 SPX and VIX
values from 24 Aug 2001 to
02 Oct 2001

	VIX	SPX
24 Aug 2001	19.71	1184.93
27 Aug 2001	20.56	1179.21
28 Aug 2001	22.00	1161.51
29 Aug 2001	23.03	1148.56
30 Aug 2001	25.41	1129.03
31 Aug 2001	24.92	1133.58
04 Sep 2001	25.85	1132.94
05 Sep 2001	26.35	1131.74
06 Sep 2001	28.61	1106.40
07 Sep 2001	30.99	1085.78
10 Sep 2001	31.84	1092.54
17 Sep 2001	41.76	1038.77
18 Sep 2001	38.87	1032.74
19 Sep 2001	40.56	1016.10
20 Sep 2001	43.74	984.54
21 Sep 2001	42.66	965.80
24 Sep 2001	37.75	1003.45
25 Sep 2001	35.81	1012.27
26 Sep 2001	35.26	1007.04
27 Sep 2001	34.00	1018.61
28 Sep 2001	31.93	1040.94
01 Oct 2001	32.32	1038.55
02 Oct 2001	31.18	1051.33

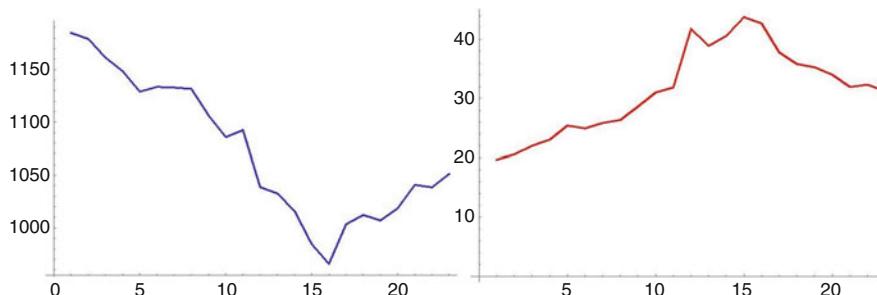


Fig. 1.41 SPX (blue) and VIX (red) from 24 August 2001 to 2 October 2001

did not respond to the price decline in the SPX, taking a “wait-and-see” stance and remaining essentially unchanged in the course of 14 April.

Table 1.8 and Fig. 1.41 shows another phase (in September 2001) in which the opposite movement of the SPX and VIX is particularly salient.

We now recall earlier chapters where, for example, we priced derivatives in a binomial model or where we attempted to determine break-even points of

various trading strategies under the assumption of volatilities that were negatively correlated with the underlying. There, we used—without giving detailed reasons—the following approach to modelling this dependence between the SPX and VIX:

$$\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t} \right)^a$$

Here, S_0 denotes the current price of the underlying asset (in this case: the SPX), S_t denotes the price of the SPX at time t , σ_0 was the volatility at time 0, and σ_t was the volatility at time t .

a was attempted to choose such that the modelling $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t} \right)^a$ matched the actual movements in volatility as closely as possible. If we choose the parameter $a = 4$ for our data example of September 2001, we get the graph shown in Fig. 1.42. The red line represents the actual volatility and the green line the modelled volatility. The modelled and real movements correspond surprisingly well.

This quite convincing sample modelling of VIX movements by a function that is dependent on the SPX price also delivers compellingly good results in almost all other phases of the VIX, in general most noticeably for periods of up to 100 trading days. In Fig. 1.43, we arbitrarily selected periods of 100 trading days each from various time ranges since 1990 and plotted the movements of the VIX (red) and the associated modelling with parameter $a = 4$ (green) in these periods.

In many cases, the modelled movements are very close to the real VIX movements. In some cases, the modellings over a period of 1000 trading days are also close to realistic, although in other cases, of course, they can be far off the mark (see Fig. 1.44 for a positive and a negative example).

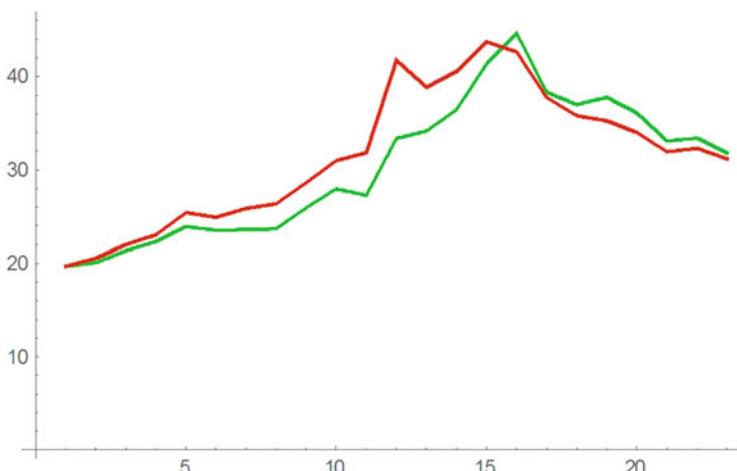


Fig. 1.42 VIX in September 2001 (red) and modelling with parameter $a = 4$ (green)

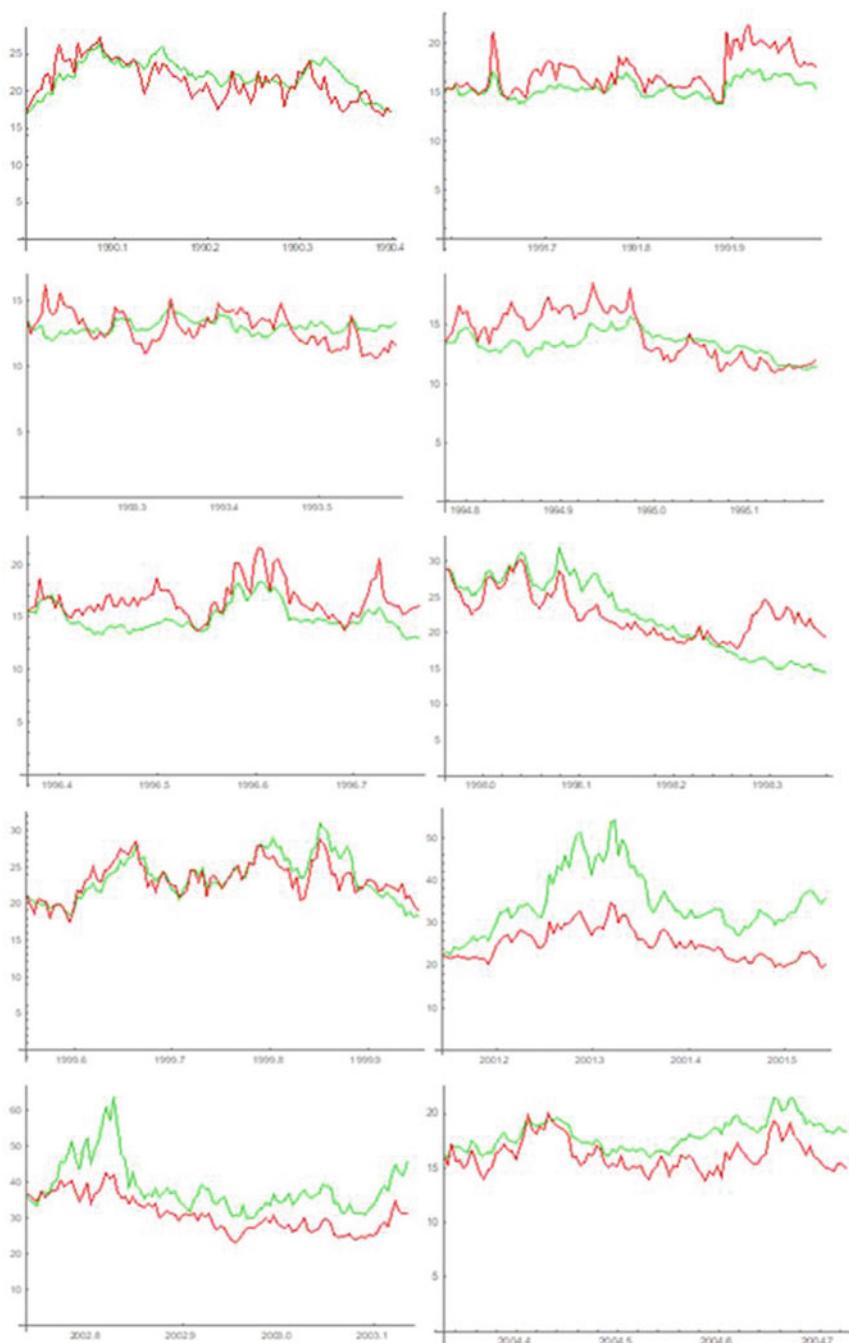


Fig. 1.43 VIX in various periods of 100 trading days (red) and modelling with parameter $a = 4$ (green)

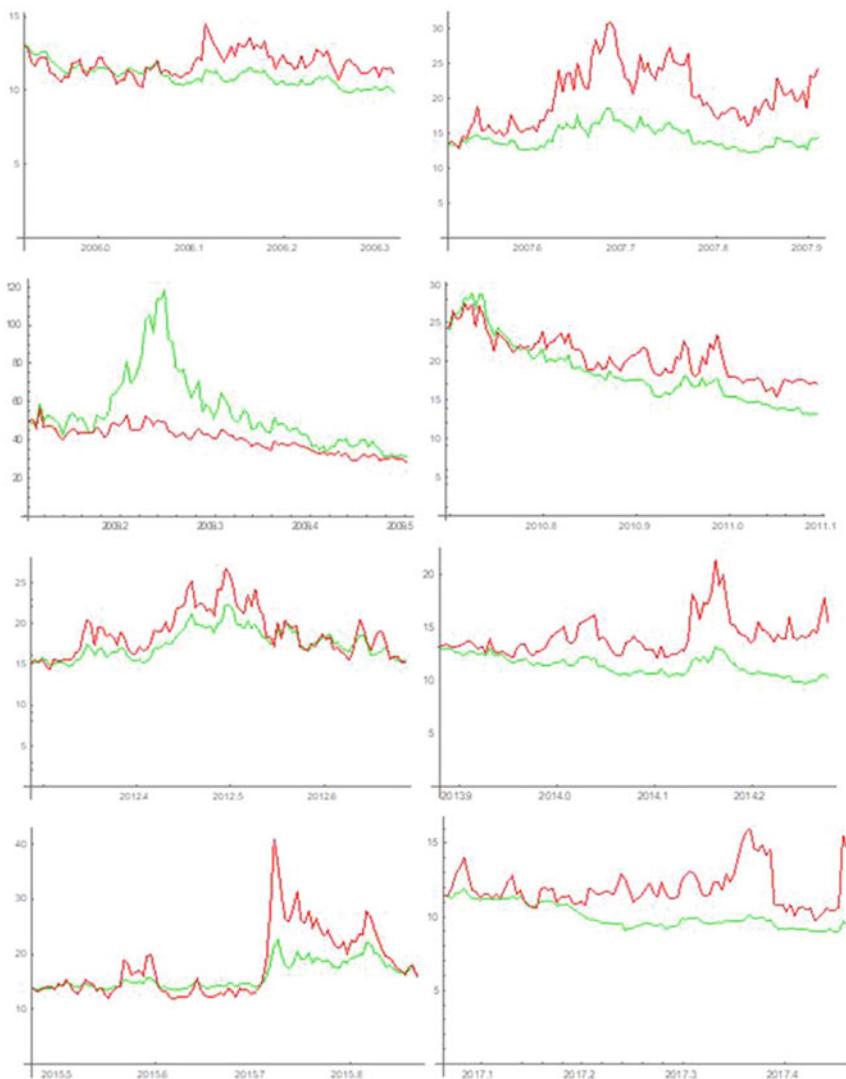


Fig. 1.43 (continued)

If we pick out the chart with the 100 trading day period in 2007 from Fig. 1.43, we see that the VIX and the modelling do not correspond particularly well. However, if we divide the 100 trading days into 5 sub-periods of 20 trading days each (= approximately one trading month each), the correspondence is strikingly good. In Fig. 1.45, the entire period is shown in the upper-left corner, followed by the five sub-periods.

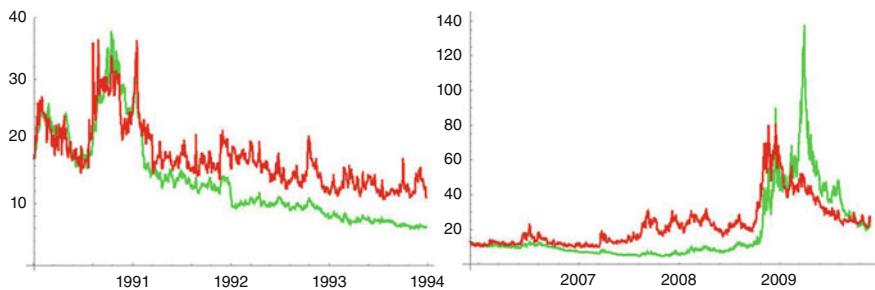


Fig. 1.44 VIX from 1990 to year-end 1993 and 2006 to year-end 2009 (red) and modelling with parameter $a = 4$ (green)

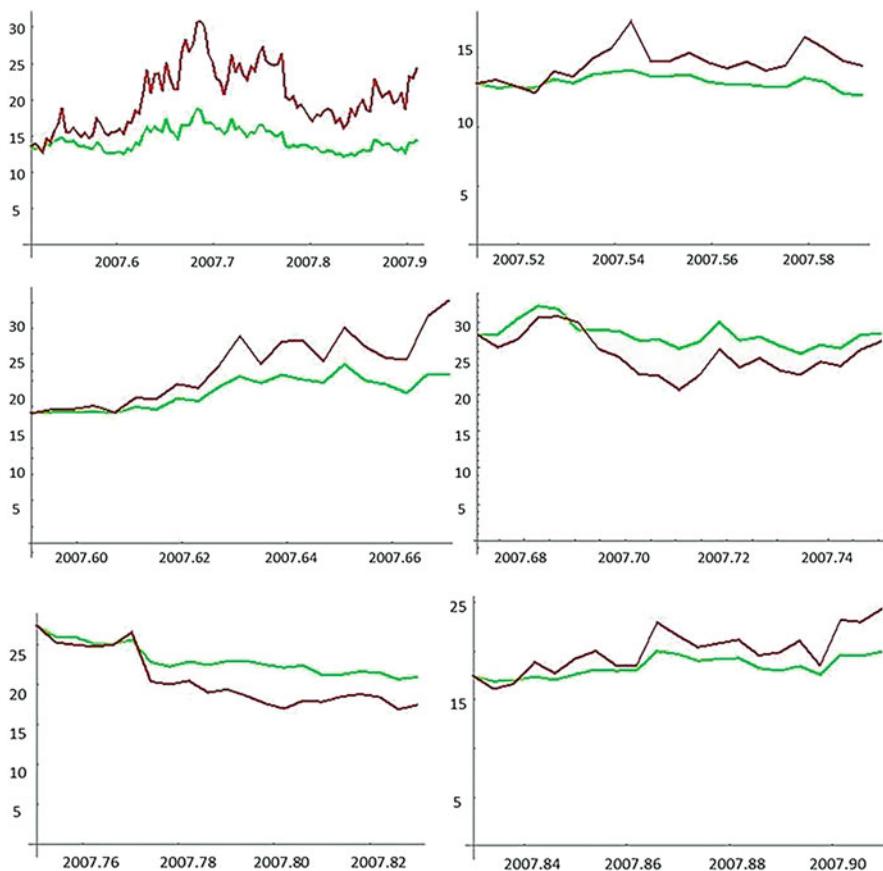


Fig. 1.45 VIX (red) over a period of about five trading months (top left), split up into five individual months (following charts), and modelling with parameter $a = 4$ (green)

The negative correlation between SPX and VIX which we noted above has quite obvious implications, of course, and is a substantial factor that must be taken into account in, for instance, analyses and risk assessments of derivative trading strategies. We have done so in earlier chapters.

In the next two sections, we are going to explicitly point out two specific effects and potential applications.

1.14 Influence of Price- and/or Time-Dependent Volatility on Delta, Gamma, and Theta

When the volatility σ as used in the Black-Scholes formula is no longer constant but depends on the time t and/or on the price of the underlying asset s , this must of course be taken into account in differentiating the Black-Scholes formula, giving us somewhat expanded versions of the relevant Greeks theta, delta, and gamma.

In the following, we will perform the explicit calculation only for the delta of a call option. Calculating gamma and theta can lead to relatively extensive formulas and is best done with mathematical software. However, we will graphically illustrate the impact of variable volatility on the delta and gamma of a call option.

We start by calculating the delta of a call option with variable volatility. In determining the delta for constant volatility, our procedure was:

Differentiate the call price $C(t, s) = s \cdot \mathcal{N}(d_1) - e^{-r(T-t)} \cdot K \cdot \mathcal{N}(d_2)$ with respect to s , where

$$d_1 = \frac{\log\left(\frac{s}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}$$

and

$$d_2 = \frac{\log\left(\frac{s}{K}\right) + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}},$$

thus $d_2 = d_1 - \sigma\sqrt{T-t}$ and hence $d_2' = d_1'$.

It follows then that

$$\begin{aligned} C'(t, s) &= \mathcal{N}(d_1) + s \cdot \frac{\partial \mathcal{N}(d_1)}{\partial d_1} \cdot d_1' - e^{-r(T-t)} \cdot K \cdot \frac{\partial \mathcal{N}(d_2)}{\partial d_2} \cdot d_2' = \\ &= \mathcal{N}(d_1) + s \cdot \frac{\partial \mathcal{N}(d_1)}{\partial d_1} \cdot d_1' - e^{-r(T-t)} \cdot K \cdot \frac{\partial \mathcal{N}(d_2)}{\partial d_2} \cdot d_1'. \end{aligned} \quad (1.1)$$

Substituting for d_2 , we find that the two last summands in this last formula cancel each other, so that only

$$C'(t, s) = \mathcal{N}(d_1)$$

is left.

(Here again, C' denotes the derivative with respect to the parameter s , of course, just like $\sigma'(t, s)$ will denote the derivative of σ with respect to the parameter s in the following.)

Now, how does this argumentation change in the case of a time- and price-dependent volatility $\sigma(t, s)$ as opposed to a constant volatility σ ? We need to substitute $\sigma(t, s)$ for each occurrence of σ in the formulas, of course. Because $d_2 = d_1 - \sigma(t, s) \cdot \sqrt{T-t}$, we now have $d_2' = d_1' - \sigma'(t, s) \cdot \sqrt{T-t}$ instead of $d_2' = d_1'$.

The formula (1.1) therefore changes to

$$\begin{aligned} C'(t, s) &= \mathcal{N}(d_1) + s \cdot \frac{\partial \mathcal{N}(d_1)}{\partial d_1} \cdot d_1' - e^{-r(T-t)} \cdot K \cdot \frac{\partial \mathcal{N}(d_2)}{\partial d_2} \cdot d_1' + \\ &\quad + e^{-r(T-t)} \cdot K \cdot \frac{\partial \mathcal{N}(d_2)}{\partial d_2} \cdot \sigma'(t, s) \cdot \sqrt{T-t}. \end{aligned}$$

As in the original case, the second and third summands in this expression cancel each other out and we are left with

$$\Delta(t, s) = C'(t, s) = \mathcal{N}(d_1) + e^{-r(T-t)} \cdot K \cdot \phi(d_2) \cdot \sigma'(t, s) \cdot \sqrt{T-t}.$$

(Here, $\phi(d_2) = \frac{\partial \mathcal{N}(d_2)}{\partial d_2}$, that is, ϕ is again the density function of the standard normal distribution.) So, in comparison with the original formula, we get an additional additive term. Since we have found that $\sigma(t, s)$ tends to fall as s increases, the derivative $\sigma'(t, s)$ tends to be negative, and so the additional additive term tends to be negative (since all other factors are positive). Consequently, it is quite possible that under certain circumstances the delta of a call option can also be negative.

The following Fig. 1.46 illustrates the behavior of the Δ of a call option with the following parameters:

$$\begin{aligned} T &= 1 \\ t &= 0 \end{aligned}$$

$$\text{Strike } K = 100$$

$$\begin{aligned} r &= 0 \\ \sigma_0 &= 0.3 \\ S_0 &= 100 \end{aligned}$$

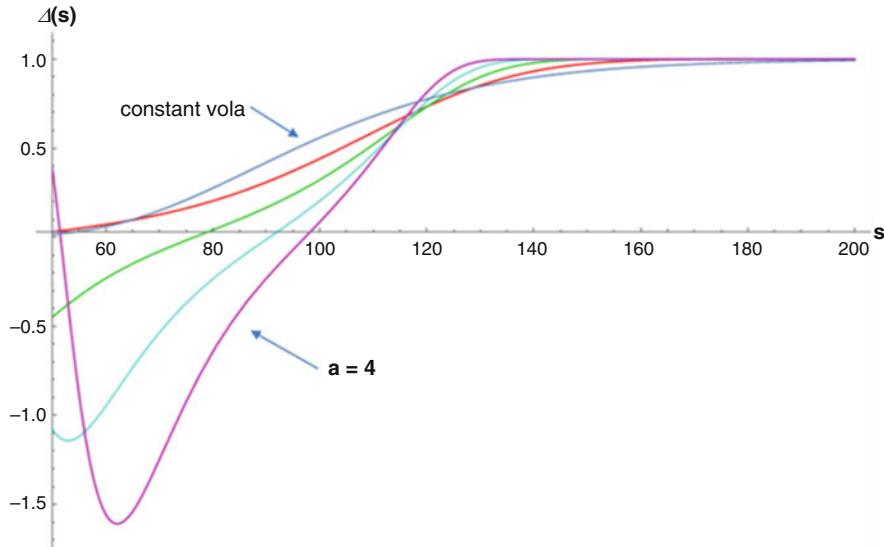


Fig. 1.46 Delta of a call option as a function of the underlying asset's price with variable volatility of $\sigma(t, s) = \sigma_0 \cdot \left(\frac{s_0}{s}\right)^a$ with $a = 0$ (blue), $a = 1$ (red), $a = 2$ (green), $a = 3$ (turquoise), and $a = 4$ (magenta)

For the variable volatility, we set $\sigma(t, s) = \sigma_0 \cdot \left(\frac{s_0}{s}\right)^a$, where $a = 0$ (constant volatility 0.3) and $a = 1, 2, 3, 4$.

We see, for example, for $a = 4$ (which we considered to be quite realistic above), that for small (but not too small) values of s , a negative delta decreases as the underlying price increases (whereas delta is always positive and monotonically increasing in the case of constant volatility). In principle, as we know, a small s leads to a smaller call price. In our setting ($a = 4$), however, a small s also leads to greatly increased volatility, and this in turn increases the call price. In the range given in our example, the influence of the increased volatility on the call price outweighs the influence of the reduced underlying price s on the call price.

The formulas for gamma and theta of call options with variable volatility are somewhat more complex and will not be given here. Figures 1.47 and 1.48, however, show how they move as a function of the underlying price for the different parameters of a .

What is perhaps particularly noticeable here: It is apparent from the way the theta moves that the variable volatility (in our modelling) increases the (negative) influence of the time to expiration on the call option's price, where the underlying price is smaller than the strike, and decreases its influence in the range, where the price is larger than the strike.

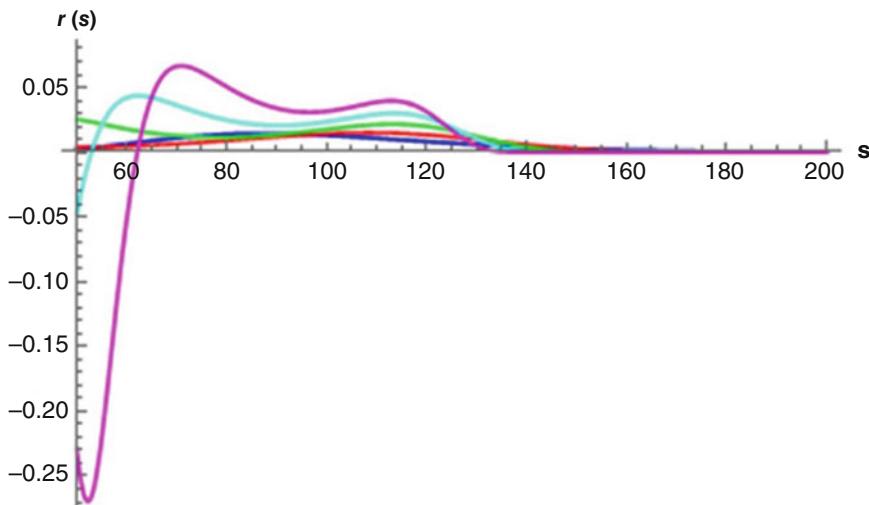


Fig. 1.47 Gamma of a call option as a function of the underlying asset's price with variable volatility of the form $\sigma(t, s) = \sigma_0 \cdot \left(\frac{s_0}{s}\right)^a$ with $a = 0$ (blue), $a = 1$ (red), $a = 2$ (green), $a = 3$ (turquoise), and $a = 4$ (magenta)

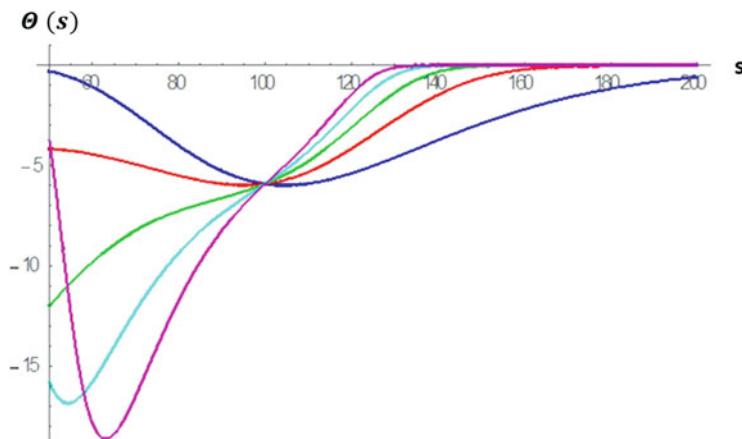


Fig. 1.48 Theta of a call option as a function of the underlying asset's price with variable volatility of $\sigma(t, s) = \sigma_0 \cdot \left(\frac{s_0}{s}\right)^a$ with $a = 0$ (blue), $a = 1$ (red), $a = 2$ (green), $a = 3$ (turquoise), and $a = 4$ (magenta)

1.15 Combined Trading of SPX and VIX for Hedging Purposes

The observed negative correlation between SPX and VIX suggests that it may be possible to reduce the risk of investing in the SPX in the following way: If an investment in the SPX is combined with an investment in the VIX (we will discuss in a later chapter how this can be done), then, because the VIX rises when the SPX falls, this will reduce potential losses of the investment. On the other hand, as the SPX rises, the VIX usually falls, reducing the potential gains of an SPX-only investment somewhat.

In any case, however, this combination should reduce the volatility (risk) of an SPX-only investment. We are going to take a brief look at the effect of such an approach using historical data from 1990 to November 2018 (purely theoretically for now, without details on the actual execution of the transactions) in the following.

To that end, we normalize both the SPX and the VIX to an initial value of 100 on 2 January 1990 and graphically represent the percentage changes in the SPX and VIX from 1990 onwards (see Fig. 1.49).

We recall that the VIX has a much higher volatility than the SPX. It is therefore questionable whether adding a VIX investment can actually balance out an investment in the SPX. As an additional risk measure, we will therefore also look at the “maximum drawdown” (MDD) of the respective strategies. MDD is the largest percentage loss in value that a strategy incurs between any two points in time over a financial product’s life.

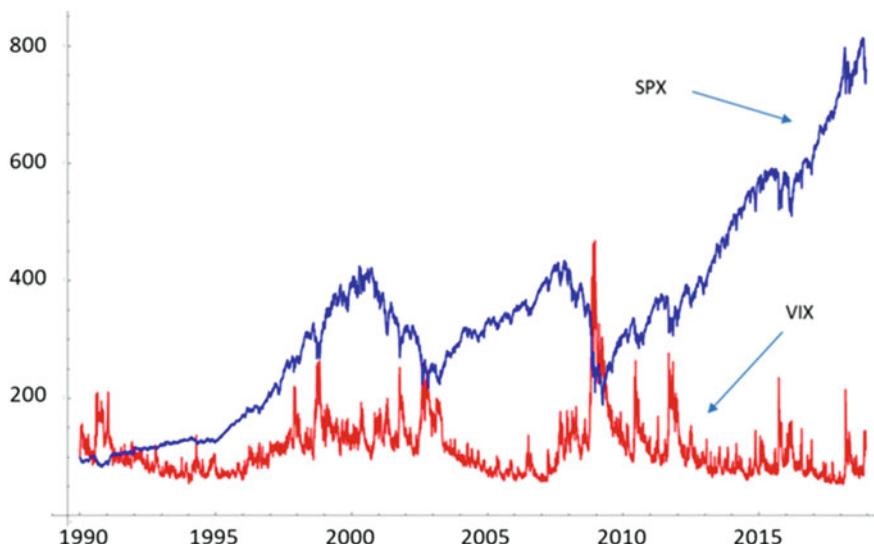


Fig. 1.49 Percentage changes in the SPX (blue) and VIX (red) from 1990 to November 2018

Excursus: Complexity of Algorithms, Calculation of Maximum Drawdown

To determine the maximum drawdown of, say, the S&P500 from 1990 to November 2018 (based on daily closing prices), the most obvious approach would be that, for any pair of days A and B over the product's life (we will always use $A < B$ in the following, meaning: A denotes an earlier date than B), we would calculate the percentage increase or decrease in the SPX price. We are looking at 7270 daily rates within this period. So, for each date A , we would need to consider $(7270 - A)$ possible dates B that come after A .

This gives us a total of $\sum_{A=1}^{7269} 7270 - A = 26,422,815$ pairs to work with.

In principle, this task is easily accomplished with a computer, of course. However, what if the number N of individual data is not 7270, but 1 million, or more? The number of pairs to be considered would then be around $\frac{N^2}{2}$. (So, for N equal to one million, there would be about 500 billion pairs to process, which would take quite some time on conventional computers.)

The computational **complexity** of this approach would be in the range of around $\frac{N^2}{2}$. We say: *The complexity of this algorithm for determining the maximum drawdown is of quadratic order.*

Illustratively, this means: Determining the maximum drawdown with this approach (i.e. calculation of the price movement from A to B for each two pairs of A and B) requires approximately $c \cdot N^2$ basic computational steps for N given data (where c is a fixed constant that does not change (materially) with increasing N).

The search for, or rather, the development and implementation of, algorithms for solving a computational problem with the lowest possible complexity is the main—and often highly challenging—objective of algorithmic mathematics.

In many cases, the persons involved would be highly satisfied with a quadratic complexity, such as in our case. As long as an algorithm for a computational problem has complexity of order N or N^2 , or N^3 , or N^4 , ..., we refer to that as *polynomial complexity of the algorithm*.

If there is a solution algorithm with polynomial complexity for a computational problem, we refer to that as a *computational problem with polynomial complexity*.

There are very prominent problems for which no solution algorithms of polynomial complexity exist. In such cases, exponential complexities of, for example, the form 2^N are not uncommon. For example: In the presence of, say, $N = 1000$ data, the algorithm requires approximately $2^{1000} \approx 10^{300}$ computational steps.

One such problem is (in all likelihood) the famous *traveling salesperson problem* (compare to Fig. 1.50).

(continued)

And it goes like this: We have a network of N cities. Each pair of cities is connected by a direct route of a certain length. We start in a city X and are supposed to find *the shortest possible route* that takes us through each city exactly once and ultimately ends up back at the starting point.

A “naive” algorithm to solve this problem would be the following:

We take every possible circuit that starts at X , visits each city exactly once and returns to X . We calculate the length of each circuit and then simply choose the shortest one. Problem solved!

But: The number of such paths is calculated as follows:

Starting from X , we have $N - 1$ cities that we can visit first. From that first city, there are $N - 2$ cities to choose from as second on our route, from each of these second cities there are $N - 3$ cities as the third place to visit, and so on. If at the end of our journey we are in the penultimate place to be visited, then the last place to be visited is well-defined, that is, we have only one place left to choose from before finally returning to X . This means that we have $(N-1) \cdot (N-2) \cdot (N-3) \cdots 1 = (N-1)!$ possible routes. Observe, however, that in this case, each route has been counted twice, since each return trip was listed as a separate route. Consequently, we actually have “only” $\frac{(N-1)!}{2}$ routes to measure. But this $\frac{(N-1)!}{2} \approx \left(\frac{N}{e}\right)^N$ even is growing super-exponentially. So, even for the moderate problem of the 50 cities shown in Figure 2.56 in Volume I Section 2.14, we are already looking at around 10^{62} possible routes. This was a (very) naive algorithm, of course, and one might think that with a much more subtle and refined approach, we could develop much faster and more efficient algorithms that could solve the problem in, say, N^2 steps.

Yet, that is not the case! We can show (under certain conditions) that the traveling salesperson problem cannot be solved faster than in exponential time. The above-mentioned caveats—“in all likelihood and under certain conditions”—have to do with the so-called *P-NP problem*. This is one of the most famous unsolved problems in mathematics and deals with the relationship between polynomial complexity and non-polynomial complexity. An exact formulation of this problem (which has been dealt with, among others, by the widely known mathematicians Kurt Gödel and John Nash) requires the conceptual world of mathematical logic and would go beyond the scope of this book.

Let us therefore get back to our much more harmless problem of determining maximum drawdown. The very naive approach to this problem confronts us with quadratic complexity. This type of complexity is much easier to handle than, say, the super-exponential complexity of the naive algorithm for the traveling salesperson problem that we described above. However, it would be even nicer if we could accomplish the task even faster than that. The dream goal, of course, would be linear complexity, where we

(continued)

would need just $c \cdot N$ computational steps. It will definitely not be possible to accomplish the task with fewer steps, since just reviewing the N data alone requires at least N “computational steps”.

It is indeed possible to find an algorithm with linear complexity for determining maximum drawdown:

- Let t be any point in time (day) within the defined period of time. The entire period has N individual points in time (in our case: days).

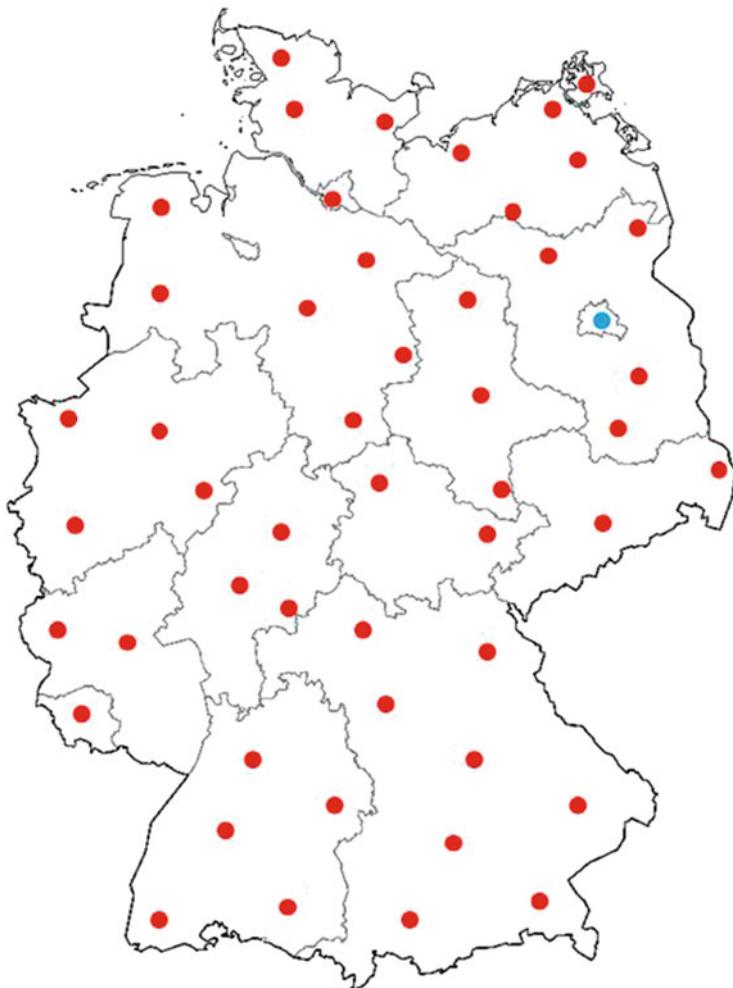


Fig. 1.50 Traveling salesperson problem: start in the blue city, visit each red city exactly once, and return to the blue city. Find the shortest possible route for this!

- The maximum price decline in percentage terms between a point in time A that's located to the left of t (earlier than t) and a point in time B located to the right of t (later than t (or equal to t)) obviously occurs from the maximum value of the price to the left of t to the minimum value of the price to the right of t .
- **Step 1:** First, we determine for each t the maximum value $U(t)$ of the price to the left of t .
- You might think that this task would again lead to a complexity of N^2 . After all, for each t , we have to examine a total of $t - 1$ earlier points in time and compare their prices with the price at time t . This leads to $\sum_{t=2}^N (t-1) = \frac{N \cdot (N-1)}{2} \approx \frac{N^2}{2}$ comparisons, that is, computational steps.
- Yet, this is not necessary. The value $U(t)$ can be derived recursively as follows:

$$U(2) = S(1)$$

Obviously: The largest price value at a time earlier than time 2 is the price value at time 1, i.e. $S(1)$.

$$U(3) = \max(U(2), S(2))$$

Obviously: The largest price value at a time earlier than time 3 is either the largest price value at a time earlier than time 2 (i.e. $U(2)$) or the new value $S(2)$, namely, if $S(2)$ is larger than the previous leader $U(2)$.

And so forth.

Once we have calculated $U(t-1)$ step by step in this way, we get (with the analogous argument to the above):

$$U(t) = \max(U(t-1), S(t-1))$$

We perform the procedure $N-1$ times in total and will then have calculated $U(t)$ for each t . For each individual step we only need one comparison, i.e. one computational step. Consequently, the complexity in calculating all $U(t)$ is only of the order of N .

- **Step 2:** Now we determine for each t the minimum value $D(t)$ of the price to the right of t (here we include t itself).
- Again, we do this recursively, as follows (we now start on the right side of the timescale):

$$D(N) = S(N)$$

$$D(N-1) = \min(D(N), S(N-1))$$

Once we have calculated $D(t+1)$ step by step in this way, we get

$$D(t) = \min(D(t+1), S(t))$$

- The complexity of determining $D(t)$ for each t is again only of order N .

(continued)



Fig. 1.51 The functions $U(t)$ (red), $D(t)$ (green), and $Draw(t)$ (magenta) for the S&P500 (blue)

- For each t , we can now calculate the maximum percentage decline in the price between a time A to the left of t (earlier than t) and a time B to the right of t (later than t (or equal to t)) by

$$Draw(t) := 100 \cdot \frac{(D(t) - U(t))}{U(t)} \quad (N \text{ computational steps}).$$
- The smallest (most negative) of the $Draw(t)$ values is then, of course, the maximum drawdown, which we obtained via a number of computational steps of the order of N .

Using the above method, we now determine the maximum drawdowns for the SPX-VIX mix strategies considered below. To illustrate the method, we first run the algorithm for the SPX only and plot the results in a graph (Figs. 1.51 and 1.52).

The maximum drawdown occurred—as was to be expected—in the period from fall 2007 to spring 2009, specifically from 9 October 2007 to 9 March 2009, from 1565.15 points to 676.53 points, representing a drop of 56.78%.

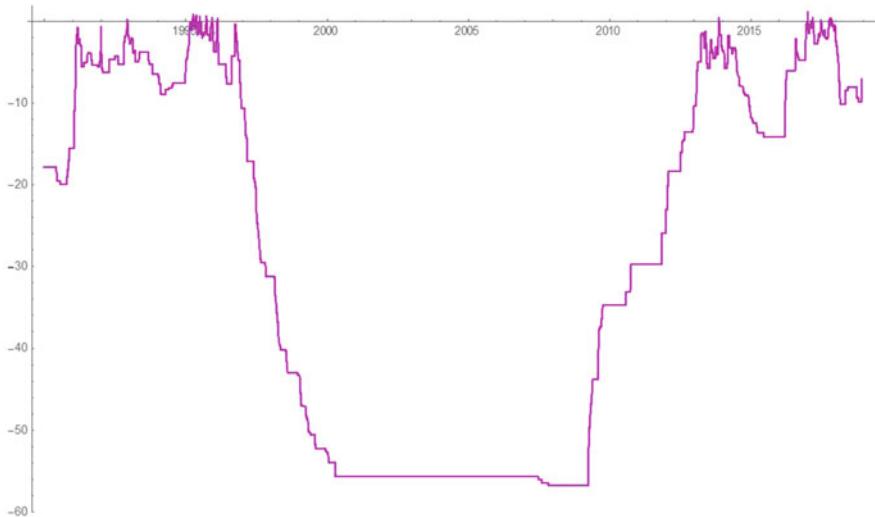


Fig. 1.52 The function $Draw(t)$ in more detail

In the following, we are going to compare some measures with respect to the SPX only and with respect to mix strategies of 80% SPX with 20% VIX, 60% SPX with 40% VIX, and 40% SPX with 60% VIX, and we will also illustrate the comparisons graphically in Fig. 1.53.

Sharpe Ratio A very important and frequently used performance metric is the Sharpe ratio for a defined performance period. It is given by the expression

$$\frac{\mu - r}{\sigma},$$

where μ denotes the annualized trend and σ denotes the annualized volatility of the analysed financial product for the relevant performance period.

r denotes the risk-free interest rate for the period.

So the Sharpe ratio measures the ratio of excess return ($\mu - r$) to risk (σ). Naturally, a risky financial product is expected to have a positive Sharpe ratio (higher trend than r). Sharpe ratios greater than 0.5 are generally considered to be very positive.

Table 1.9 and Fig. 1.53 suggest that only the mix strategy with a 20% VIX share seems to be of interest here (higher Sharpe ratio than the SPX-only strategy with smaller maximum drawdown).

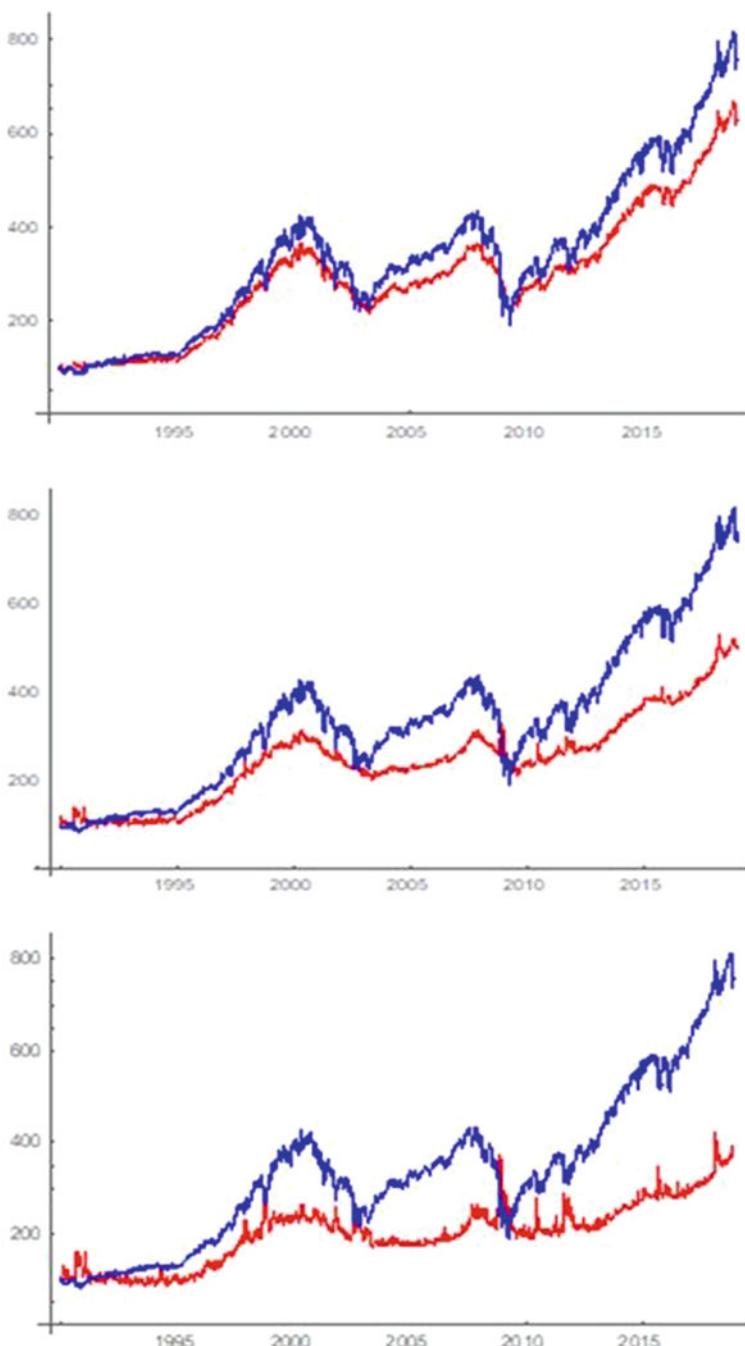


Fig. 1.53 SPX-VIX mix strategy (red) compared with SPX (blue), for a VIX share of 20%, 40%, and 60% (from top to bottom)

Table 1.9 Comparison of the SPX, VIX, and mixed strategies in terms of volatility ρ and per day as well as the sharp ratio

	Return p.a.	Volatility per day based on daily data	Volatility p.a. based on annual data	Maximum drawdown	Sharpe ratio ($r = 3\%$)
SPX	7.26%	1.10%	16.88%	-56.78%	0.25
VIX	0.43%	6.77%	40.64%	-88.70%	-0.06
20 VIX/80 SPX	6.57%	0.72%	13.63%	-43.44%	0.26
40 VIX/60 SPX	5.73%	1.20%	12.38%	-36.89%	0.22
60 VIX/40 SPX	4.64%	2.20%	14.63%	-49.80%	0.11

1.16 Relation and Correlations of VIX with Historical and Realized Volatility

The VIX, or implied volatility, of an option represents the volatility of the underlying that is implicitly contained (under certain assumptions) in the price of that option (with a given period to expiration T). Now, what is the relation between this implied volatility (for a defined period in the future) and the underlying asset's estimated historical volatility for this period to expiration, or between the implied volatility and the underlying asset's actually realized volatility as measured in retrospect?

This fundamental question will come up time and again throughout this book, as it is absolutely essential for various trading strategies. To answer it, we will start by presenting a few graphical comparisons.

Figure 1.54 shows the VIX in red from 1990 to October 2021 in the lower chart and, in comparison, the historical 20-day volatility (in green) from daily data measured in each case on the same day as the VIX (thus basically the volatility realized in the past 20 trading dates). (For reference, the price movements of the S&P500 over the same period are shown once more in the upper chart.) You could interpret the graph in the following way:

It is possible to use the historical 20-day volatility of the past 20 trading days as a forecast for the realized volatility of the next 20 trading days (which is essential for delta-hedging an option). On the other hand, the VIX is often used as a forecast for the realized volatility of the next 20 trading days. How do the two forecast values relate to each other? As it turns out, the historical volatility almost always delivers forecast values that are lower than the VIX, sometimes significantly so.

This becomes evident again in Fig. 1.55, which shows the difference between VIX and historical volatility of the previous period in absolute numbers. In Fig. 1.56, the difference is shown as a percentage with respect to the historical volatility. The difference is usually clearly positive.

Even more interesting, however, is the comparison between the VIX as a tool for forecasting subsequent realized volatility (over the next 20 trading days) and the actual realized volatility value (as determined a posteriori). This comparison

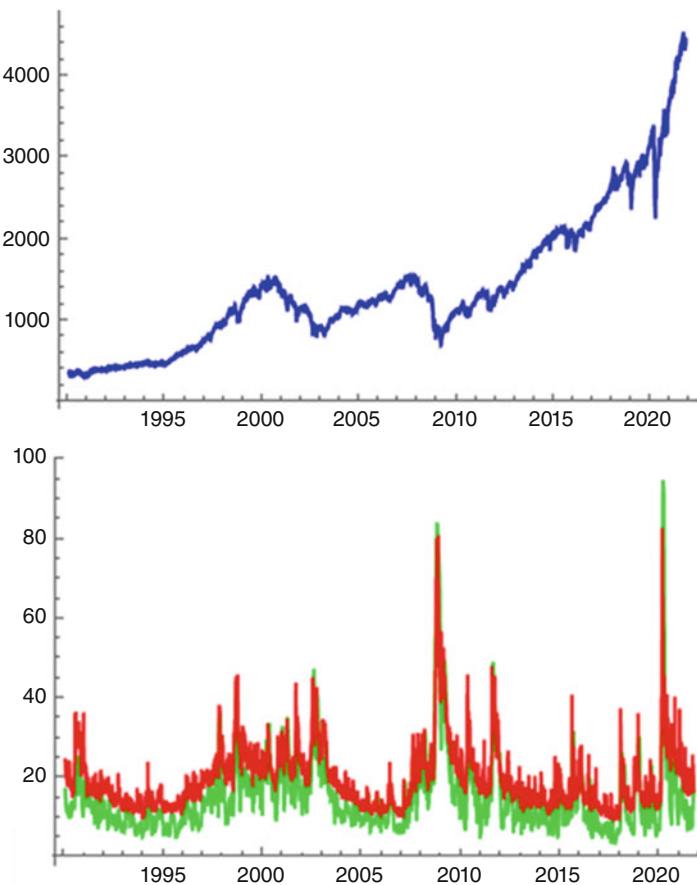


Fig. 1.54 S&P500 (blue), VIX (red), and 20-day historical volatility (green) from 1 January 1990 to 14 October 2021

is illustrated in Fig. 1.57. Here, too, we see a similar picture: mostly significantly higher values for the VIX than for the actually realized volatility.

Figures 1.58 and 1.59 again show the difference between VIX and the then actually realized 20-day volatility, first in absolute terms and then in percentage terms.

Superimposing Figs. 1.55 and 1.58 as well as Figs. 1.56 and 1.59 (to illustrate the differences between VIX and current historical volatility (blue) and between VIX and subsequently realized volatility in absolute numbers and in percentage terms (blue and red), respectively), we see that the difference in the first case is on average somewhat larger than in the latter case.

Finally, Table 1.10 shows the values for the correlations between VIX on the one hand and historical and realized volatility on the other hand. This calculation was

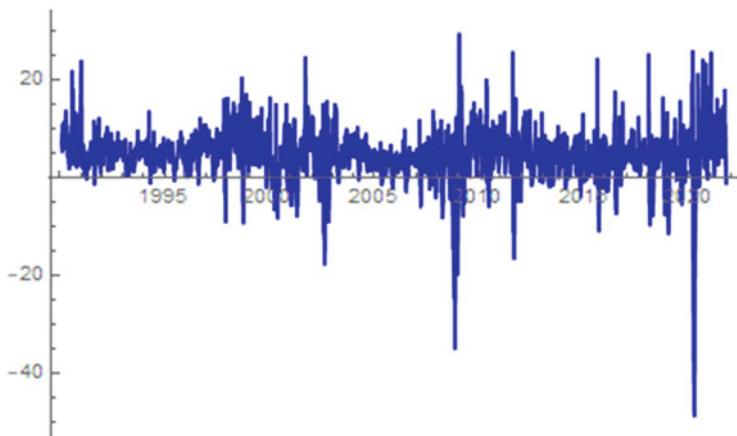


Fig. 1.55 Difference VIX minus historical 20-day volatility (same-day measurements) from 1 January 1990 to 14 October 2021

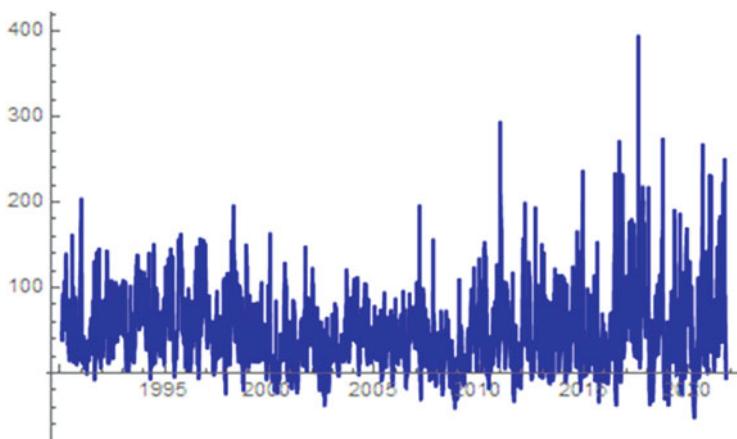


Fig. 1.56 Percentage excess of VIX over historical 20-day volatility on the same trading day

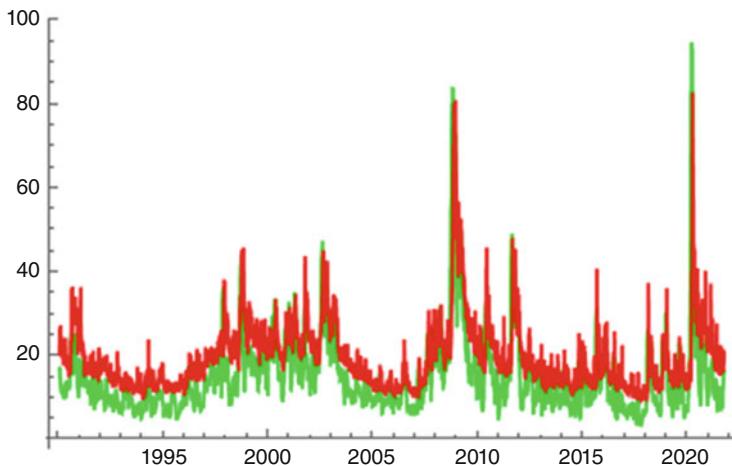


Fig. 1.57 VIX (red) and realized volatility in the immediately following 20 trading days (green) from 1 January 1990 to 14 October 2021



Fig. 1.58 Difference between VIX and realized volatility in the immediately following 20 trading days from 1 January 1990 to 14 October 2021

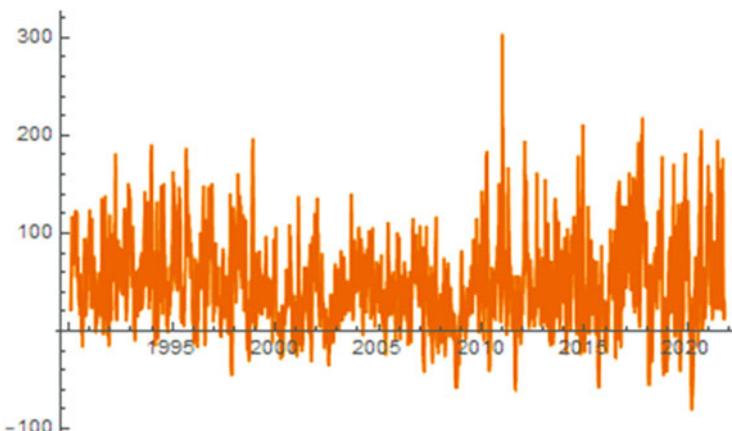


Fig. 1.59 Percentage excess of VIX over realized volatility in the subsequent 20 trading days

Table 1.10 Correlations between VIX values and historical and realized volatilities

	VIX/historical 20-day volatility same day	VIX/realized volatility on subsequent 20 trading days
1990–Oct. 2021	0.863	0.768
2000–Oct. 2021	0.873	0.776
2010–Oct. 2021	0.817	0.673

performed for different time periods to illustrate that the correlations exhibit fairly consistently high values close to 1.

Table 1.10 of correlations between VIX values and historical and realized volatilities:

Graphs 1.60 and 1.61 above seem to confirm a commonly heard observation:

“The VIX systematically overestimates subsequent realized volatility!” A large number of studies are dedicated to this observation and how to potentially exploit this difference, and a number of indices (volatility arbitrage indices) track the performance of such strategies.

Possible approaches to profit from a potentially excessive implied volatility are, for example:

- Systematically selling “overly expensive” options (options with comparatively unusually high implied volatility), in the hope of making a profit in the long run.
- Selling options with high implied volatility and concurrent delta-hedging using the underlying (or futures on the underlying). Since hypothetically, the underlying asset’s realized volatility over the life of the hedge is usually lower than the option’s implied volatility, this should usually yield a positive profit (unless trading costs get out of proportion). (See Volume III Chapter 3.13 for details.)
- Trading volatility swaps (or variance swaps).

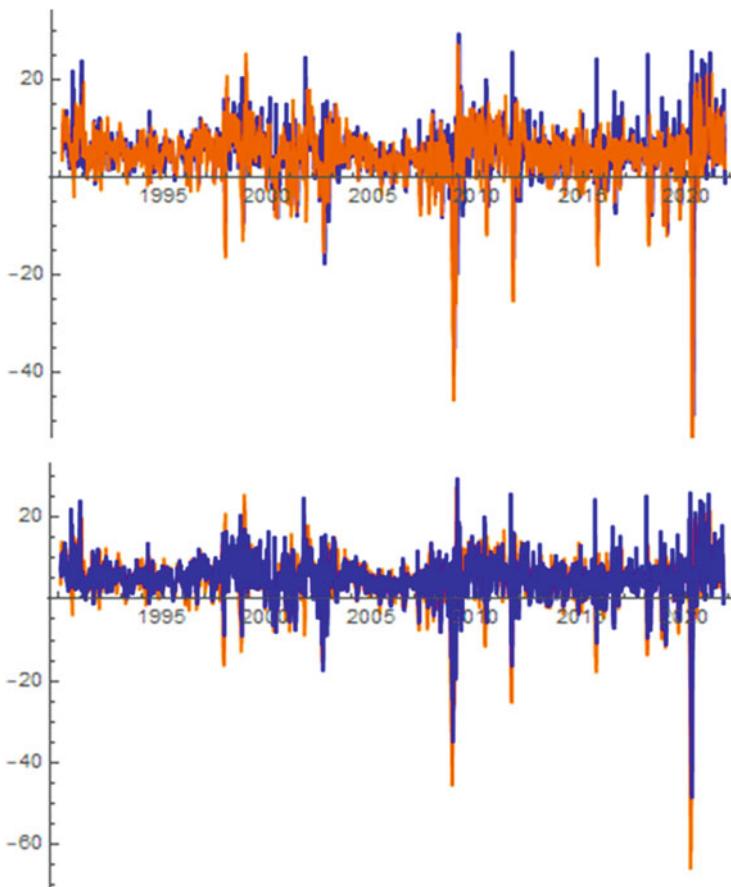


Fig. 1.60 Superimposed Figs. 1.55 and 1.58

Volatility swaps are essentially futures on an underlying asset's realized volatility up to a time T in the future. The fair strike price of such futures (i.e. the price at which the futures contract is entered into) is essentially determined by the implied volatility of the underlying asset. The payoff of such a volatility swap is therefore essentially the difference between realized volatility and implied volatility.

Volatility swaps are tracked by various volatility arbitrage indices, which serve to illustrate how realized volatility is systematically overestimated by implied volatility (based on systematically executed volatility swap trading strategies). The advantage of using volatility swaps to exploit volatility differences is definitely that these products are completely independent of the underlying asset's direction. Only the volatility itself is traded with these products. The downside is the more complex nature of the products, especially the more complex nature of the fair strike dynamics of volatility swaps.

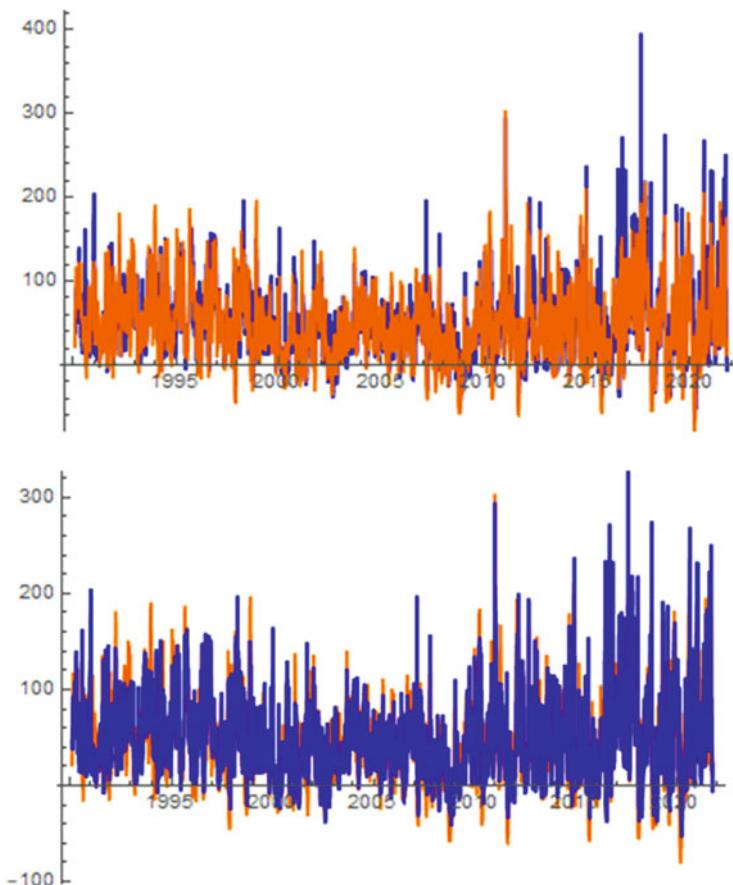


Fig. 1.61 Comparison of the graphs from Figs. 1.56 and 1.59

Here—in the following chapter—we are only going to look at the details of one such index, which is focused on “trading excessive implied volatilities”—the CBOE “S&P500 PutWrite Index”. In a later chapter, we will provide a detailed analysis of two other strategies and their characteristics.

1.17 The CBOE S&P500 Put Write Index

The CBOE S&P500 PutWrite Index tracks a continuously executed very simple options trading strategy that is based on systematically selling certain put options on the S&P500.

The concept: Regular collection of option premiums has no further activities. If there exist systematically overpriced options (i.e. options whose implied volatility is

too high compared to the subsequently realized volatility), then this approach should result in a long-term average outperformance (average excess return compared to the underlying, the S&P500).

The index was launched by the CBOE in 2007. Back-calculated data exist back to 30 June 1986. The index was normalized such that it had a normalized value of 100 on 1 June 1988.

The PutWrite Index continuously tracks the following trading strategy (we won't provide detailed technical trading information (at what time of the day exactly are transactions executed at what price, how exactly is free cash invested risk-free, etc.) here, but you can read up on them here, for example: <http://www.cboe.com/products/strategy-benchmark-indexes/putwrite-indexes/cboe-s-p-500-putwrite-index-put>):

- On every third Friday of a month, the price of the PutWrite Index has a certain value A . We interpret A as the current assets in the trading strategy.
- On this third Friday, all options held in the trading strategy expire and new options are entered into.
- Specifically, one type of put option is sold at a time with expiration on the third Friday of the following month.
- The strike K of the put options is the largest available strike less than or equal to the current S&P500 value S_0 . So we sell at-the-money put options.
- The number X of traded put options depends on the current assets in the trading strategy, i.e. the current price A of the PutWrite index, and is given by $X = \frac{A}{K}$.
- This number X is determined such that even in a worst-case scenario, the loss through the strategy is safely covered by the existing investment. The worst-case scenario is if the S&P500 were to fall to the value 0 before the option expires. In this case, we would have to pay a payoff of $X \cdot K = A$ for the short position in the put option.
- The premium received from the sale is the time value of the option.
- Premium income and other available funds are always reinvested in short-dated US Treasury bills in accordance with specific rules.

Put simply, we pocket monthly profits with this strategy when the S&P500 moves up, stagnates or falls slightly (i.e. does not fall more than the value of the collected premium, i.e. the time value of the option sold). However, the maximum profit is bounded above by the collected premium.

We make losses with the strategy if the S&P500 falls more than just slightly. The losses correspond one-to-one to the losses of the S&P500 less the collected premium.

In fact, the strategy, i.e. the performance of the CBOE S&P500 PutWrite Index, closely aligns with the performance of the S&P500 itself.

Figure 1.62 represents the index price movements (in each case as a percentage value normalized to the starting point of the time series) as from 1986, i.e. from when data for the PutWrite Index are available. We see a significant outperformance of the PutWrite Index over the S&P500. In Fig. 1.63, we present the same

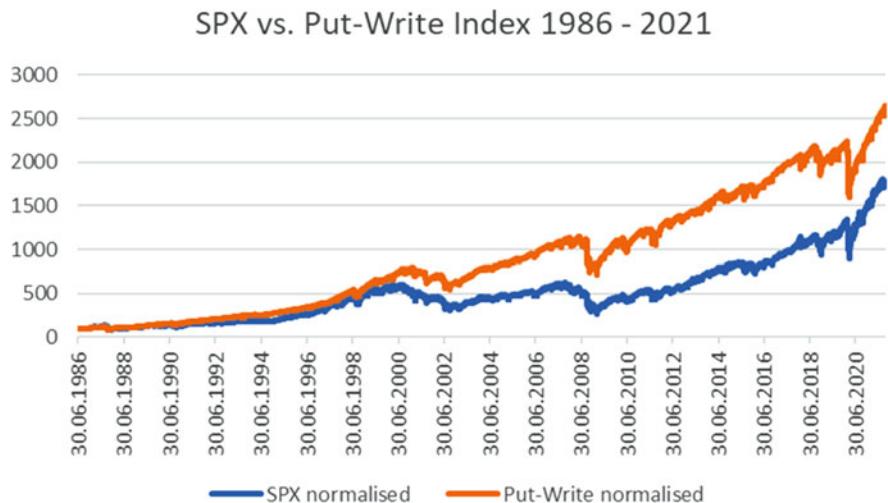


Fig. 1.62 Comparison of S&P500 performance in percent (blue) with CBOE S&P500 PutWrite Index performance in percent (orange) from June 1986 to October 2021

comparison for some other sub-periods. Except for the period from around 2010 to 2018, the graph confirms the outperformance of the PutWrite Index over the S&P500. Bear in mind that part of the performance in the PutWrite Index also results from investing cash funds in US Treasury Bills. As a result, the PutWrite Index is expected to perform significantly better in high interest rate phases than in low interest rate phases.

Particularly noticeable in Fig. 1.63 is the lower chart: In the course of the financial crisis from 2008 to around mid-2009, the price losses in the PutWrite Index were much less dramatic than in the S&P500 itself.

In the following Table 1.11, we present the performance figures of the PutWrite Index compared to the S&P500.

This immediately begs a couple of very interesting questions. In particular:

Is there any way we can optimize the strategy reflected in the PutWrite Index?

The following are the examples: by choosing different strikes, trading a different (higher) number of contracts, trading options with different expirations, applying an exit rule (closing the contracts) when a certain level of loss occurs, etc.? These questions will be the subject of one of our case studies and we will discuss them in detail in that chapter.

1.18 The VIX Calculation Methodology

As noted further above, volatility indices in the early years of the calculation and publication of such volatility indicators were determined as follows:

A certain period was specified which the volatility indexed by the index should refer to (e.g. 30 days). Then, from the array of traded options with the longest time



Fig. 1.63 Comparison of S&P0.500 performance in percent (blue) with CBOE S&P500 PutWrite Index performance in percent (orange) in different sub-periods

to expiration less than 30 days and the shortest time to expiration greater than 30 days, all implied volatilities were determined using the Black-Scholes formula. The current value of the volatility index was then obtained as a specifically weighted average of these implied individual volatilities.

Table 1.11 Performance figures of the PutWrite Index compared to the S&P500

	Return p.a.	Vol p.a.	Sharpe ratio (3%)	Max. drawdown
S&P500	7.72%	17.53%	0.27	-56.56%
S&P500 PutWrite index	9.89%	12.65%	0.54	-37.09%

From the turn of the millennium, a new calculation method was introduced for most volatility indices which is largely model-independent, that is, it does not rely on the calculation of implied volatility using the Black-Scholes formula. We will present this method in the following using the VIX as an example.

For this purpose, we will first describe how the VIX is calculated on a continuous basis (i.e. every minute). We will not go into all the technical details of the calculation, however, as that is not necessary for understanding the construction of the VIX, and will therefore take the liberty of allowing some minor simplifications. If interested in the exact calculation of the VIX, down to the smallest detail, you will find the calculation in the “CBOE VIX, White paper, CBOE Volatility Index”, which is available on the Internet.

In a second step, we will illustrate why this computation method—which may not be intuitively obvious at first but becomes extremely impressive at second glance—yields a value that replicates the implied volatility of the SPX really well.

In the following explanations, it will also become clear that on the basis of this computation method, it is possible to achieve perfect hedging of the VIX (or trade the VIX directly) using SPX options.

Computation of the VIX

The VIX is computed continuously (every minute) from the bid and ask prices of the traded SPX standard options expiring in each case upon opening of the exchange on the third Friday of a month and from the “weekly” SPX options expiring in each case upon closing of the exchange on all other Fridays (compare Fig. 1.64).

Specifically, it is computed from options with two different expirations T_1 and T_2 : those with the longest time to expiration T_1 shorter than 30 days and those with the shortest time to expiration T_2 longer than 30 days.

We are now going to look at one type of these two options. We denote its time to expiration by T (to the minute) and calculate a value σ^2 in the manner described below.

The second type will be calculated later in the exact same way as the first type.

For both types, we obtain a value σ_1^2 and σ_2^2 , respectively, in the manner described below.

The final implied volatility VIX (expressed as a percentage) is then obtained as the root of a certain weighted mean of the two individual values:

$$\text{VIX} = 100 \times \sqrt{\omega_1 \cdot \sigma_1^2 + \omega_2 \cdot \sigma_2^2}$$

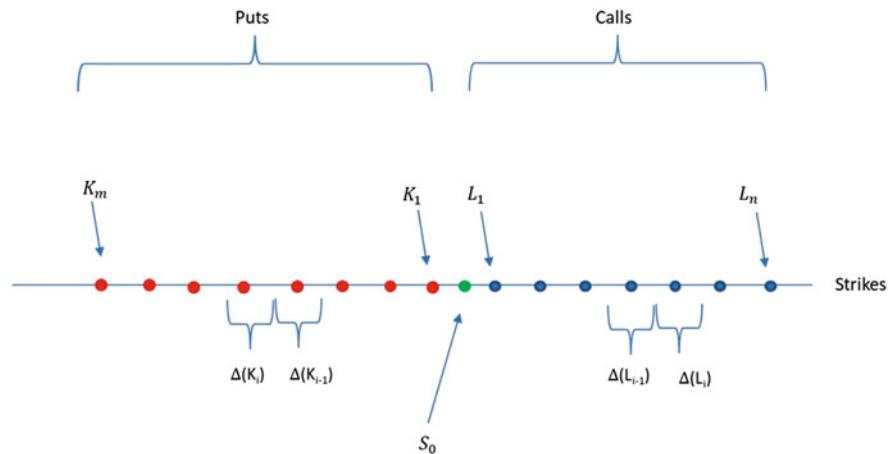


Fig. 1.64 Definition of the VIX

Here, $\omega_1 + \omega_2 = 1$ and $\omega_1 = 0$ if T_2 is exactly 30 days and $\omega_2 = 0$ if T_1 is exactly 30 days.

(The precise formulation is:

$$\omega_1 = T_1 \cdot \frac{N_{T_2} - N_{T_1}}{N_{T_2} - N_{T_1}} \cdot \frac{N_{365}}{N_{30}} \text{ and } \omega_2 = T_2 \cdot \frac{N_{T_2} - N_{T_1}}{N_{T_2} - N_{T_1}} \cdot \frac{N_{365}}{N_{30}} \text{ with}$$

$N_{30} = 43,200$ (= number of minutes of 30 days)

$N_{365} = 525,600$ (= number of minutes of 365 days)

N_{T_1} = number of minutes until T_1

N_{T_2} = number of minutes until T_2)

For each of the two T , the value σ^2 (described here in a minimally simplified way) is now calculated as follows:

Whenever we refer to “option prices” $C(K)$ for calls or $P(K)$ for puts in the following, we always refer to the mid-prices between bid and ask prices for the options with expiration T and strike K .

Let r denote the risk-free interest rate for the period $[0, T]$.

Let $S_0 > K_1 > K_2 > K_3 > \dots > K_m > K_{m+1} > K_{m+2}$ be the strikes of all put options with a strike smaller than the current S&P500 value of S_0 and for which the following holds:

m is the smallest index for which the bid prices of put options with the strikes K_{m+1} and K_{m+2} are both equal to 0.

Let $S_0 < L_1 < L_2 < L_3 \dots < L_n < L_{n+1} < L_{n+2}$ be the strikes of all call options with a strike greater than the current S&P500 value of S_0 and for which the following holds:

n is the smallest index for which the bid prices of call options with the strikes L_{n+1} and L_{n+2} are both equal to 0.

For any K_i with $i = 2, 3, \dots, m - 1$ set $\Delta(K_i) := \frac{K_{i-1} - K_{i+1}}{2}$ and similarly for any L_i with $i = 2, 3, \dots, n - 1$ set $\Delta(L_i) := \frac{L_{i+1} - L_{i-1}}{2}$.

Furthermore, $\Delta(K_1) := \frac{L_1 - K_2}{2}$, $\Delta(L_1) := \frac{L_2 - K_1}{2}$ and $\Delta(K_m) := K_{m-1} - K_m$ and $\Delta(L_n) := L_n - L_{n-1}$.

Then, we define

$$\sigma^2 := \frac{2e^{rT}}{T} \cdot \left(\sum_{i=1}^m \frac{\Delta(K_i)}{K_i^2} P(K_i) + \sum_{i=1}^n \frac{\Delta(L_i)}{L_i^2} C(L_i) \right).$$

So, the value σ^2 is defined as the current market price of a weighted portfolio D of real traded options on the S&P500. The traded options are put options with a strike smaller than the current price of the S&P500 and call options with a strike larger than the current price of the S&P500.

The illustrative meaning of the VIX:

The construction of the value σ^2 should be clear. What might be less clear, however, is what this value has got to do with the implied volatility of the underlying S&P500.

To clarify that, let us take a closer look at the portfolio D . First, we ask what the payoff of this portfolio is at time T if the value of the S&P500 at time T is equal to S_T . The payoff here is of course given by

$$\begin{aligned} \text{Payoff}(S_T) &= \frac{2e^{rT}}{T} \cdot \left(\sum_{i=1}^m \frac{\max(0, K_i - S_T)}{K_i^2} \Delta(K_i) + \right. \\ &\quad \left. + \sum_{i=1}^n \frac{\max(0, S_T - L_i)}{L_i^2} \Delta(L_i) \right) = \\ &= \begin{cases} \frac{2e^{rT}}{T} \cdot \sum_{i=1}^m \frac{\max(0, K_i - S_T)}{K_i^2} \Delta(K_i) & \text{if } S_T \leq S_0 \\ \frac{2e^{rT}}{T} \cdot \sum_{i=1}^n \frac{\max(0, S_T - L_i)}{L_i^2} \Delta(L_i) & \text{if } S_T > S_0 \end{cases}. \end{aligned}$$

The last two sums are both so-called Riemann sums of the function $f(x) := \frac{\max(0, x - S_T)}{x^2}$ if $S_T \leq S_0$ and $g(x) := \frac{\max(0, S_T - x)}{x^2}$ if $S_T > S_0$.

Since the S&P500 options market offers a very narrow band of strikes (i.e. $\Delta(K_i)$ and $\Delta(L_i)$ have small values, often value 5) and since very low strikes K_i or very high strikes L_i are also available, we can interpret

(continued)

$\sum_{i=1}^m \frac{\max(0, K_i - S_T)}{K_i^2} \Delta(K_i)$ if $S_T \leq S_0$, as a very good approximation for

$$\begin{aligned} \int_0^{S_0} f(x) dx &:= \int_0^{S_0} \frac{\max(0, x - S_T)}{x^2} dx = \int_{S_T}^{S_0} \frac{x - S_T}{x^2} dx = \\ &= \int_{S_T}^{S_0} \frac{1}{x} dx - S_T \cdot \int_{S_T}^{S_0} \frac{1}{x^2} dx = \frac{S_T}{S_0} - 1 - \log \frac{S_T}{S_0}. \end{aligned}$$

By analogy,

$\sum_{i=1}^n \frac{\max(0, S_T - L_i)}{L_i^2} \Delta(L_i)$, if $S_T > S_0$ can be interpreted as a very good approximation for

$$\begin{aligned} \int_{S_0}^{\infty} g(x) dx &:= \int_{S_0}^{\infty} \frac{\max(0, S_T - x)}{x^2} dx = \int_{S_0}^{S_T} \frac{S_T - x}{x^2} dx = \\ &= \int_{S_T}^{S_0} \frac{1}{x} dx - S_T \cdot \int_{S_T}^{S_0} \frac{1}{x^2} dx = \frac{S_T}{S_0} - 1 - \log \frac{S_T}{S_0}. \end{aligned}$$

In both cases, we get the same result. Thus, for the payoff of the portfolio D , we have:

$$\text{Payoff}(S_T) \approx \frac{2e^{rT}}{T} \cdot \left(\frac{S_T}{S_0} - 1 - \log \frac{S_T}{S_0} \right).$$

So, an approximate construction of the portfolio D could consist of the following products:

Component 1: $\frac{2e^{rT}}{T \cdot S_0}$ units of the underlying asset

Component 2: $-\frac{2e^{rT}}{T}$ units of a payment of 1 at time T

Component 3: $-\frac{2e^{rT}}{T} \log \frac{S_T}{S_0}$ units of a Derivative II which pays out at time T

If we assume that the very liquid options of the S&P500 options market are fairly priced, then we can also assume that the portfolio D is fairly priced. The current price of the portfolio D should therefore essentially correspond to the fair price of the portfolio D . In the following, we are therefore going to determine the fair price D_0 of the portfolio at time 0.

The fair price of Component 1 is $\frac{2e^{rT}}{T \cdot S_0} \times$ fair price of the underlying asset. The underlying asset's fair price is S_0 , of course, and the fair price of Component 1 is therefore $\frac{2e^{rT}}{T}$.

(continued)

The fair price of Component 2 is $-\frac{2e^{rT}}{T} \times e^{-rT} = -\frac{2}{T}$.

What is left to do now is determine the fair price of Derivative Π :

For what follows, that is, in determining the fair price Π_0 of Derivative Π , we are going to assume a Black-Scholes market. The following result regarding the fair price of Derivative Π would also apply under much milder conditions. However, since we do not yet have any other techniques than the Black-Scholes model, we can derive the following result only in the Black-Scholes model at this point. Yet, the argumentation in the Black-Scholes model can certainly serve to plausibilize the result.

In the following, E will again denote the expectation with respect to the risk-neutral measure in the Wiener model.

We can therefore assume that $S_T = S_0 \cdot e^{(r - \frac{\bar{\sigma}^2}{2}) \cdot T + \bar{\sigma} \sqrt{T} \omega}$ with a standard normally distributed random variable ω . Here, $\bar{\sigma}$ denotes the volatility of the underlying S&P500. We know that (in the Black-Scholes model) the fair price is computed as follows:

$$\begin{aligned}\Pi_0 &= e^{-rT} \cdot E(\text{payoff of Derivative } \Pi) = e^{-rT} \cdot E \left(\log \frac{S_T}{S_0} \right) = \\ &= e^{-rT} \cdot E \left(\left(r - \frac{\bar{\sigma}^2}{2} \right) \cdot T + \bar{\sigma} \sqrt{T} \omega \right) = e^{-rT} \cdot \left(r - \frac{\bar{\sigma}^2}{2} \right) \cdot T.\end{aligned}$$

Thus, Component 3 has the fair price $-\frac{2e^{rT}}{T} \times e^{-rT} \cdot \left(r - \frac{\bar{\sigma}^2}{2} \right) \cdot T = -2r + \bar{\sigma}^2$.

Adding the fair prices of the three components and approximating $e^{rT} \approx 1 + rT$ yields:

$$D_0 \approx \frac{2e^{rT}}{T} - \frac{2}{T} - 2r + \bar{\sigma}^2 \approx \bar{\sigma}^2.$$

As noted above: This result can also be derived under much milder conditions than the assumption of a Wiener model.

To summarize:

The value defined above (on which the construction of the VIX is based)

$\sigma^2 := \frac{2e^{rT}}{T} \cdot \left(\sum_{i=1}^m \frac{\Delta(K_i)}{K_i^2} P(K_i) + \sum_{i=1}^n \frac{\Delta(L_i)}{L_i^2} C(L_i) \right)$ has approximately the value $\bar{\sigma}^2$ and is the volatility in the modelling of the S&P500.

$$\sigma := \sqrt{\frac{2e^{rT}}{T} \cdot \left(\sum_{i=1}^m \frac{\Delta(K_i)}{K_i^2} P(K_i) + \sum_{i=1}^n \frac{\Delta(L_i)}{L_i^2} C(L_i) \right)}$$

is therefore an approximation of the volatility of the S&P500 for the time range $[0, T]$.

By weighting the volatilities obtained for the two option types as described above, we finally obtain the VIX, which then actually represents an approximation of the S&P500's volatility for the time range $\left[0, \frac{30}{365}\right]$, i.e. for the 30-day volatility of the S&P500.

So the VIX is based purely on a linear combination of option prices, and therefore it can be perfectly replicated and hedged by buying these options according to their respective weights.

Now, what is the relation between the VIX and the actual implied volatilities of options with a 30-day expiration?

Just as a single example: Fig. 1.65 lists a selection of options that expire exactly one month from 7 November 2018. The implied volatilities exhibit a clear skew, ending with a slight smile at the right edge. The volatility values range from about 11% to about 27%. At the money, the implied volatilities are around 14%. The actual value of the VIX at the time of the quotations in Fig. 1.65 was 16.97%. The actual value of the SPX at that time was 2774.

1.19 The Volatility Weekend Effect

Various studies point to a so-called “weekend effect” on stock prices. Many stock prices are said to—systematically—exhibit higher volatility (implied or realized) on Mondays than on other days of the week (especially Fridays).

In a slightly modified version of this effect, we can say the following: prices of financial products often go up on Fridays and fall between Friday evening and Monday evening.

A host of studies have been carried out, with detailed analyses of price data of various financial products in order to prove or disprove this weekend effect.

This weekend effect is just one of many “price anomalies” that are frequently discussed in the literature. We will not dwell on these price anomalies here, just briefly point to some of them and—since we have already

(continued)

Fig. 1.65 Select options on SPX expiring 7 December 2018 with implied volatilities, as seen on 7 November 2018

implied volatility	7 December 2018 Strike
27.1%	2400
26.5%	2425
25.4%	2450
24.7%	2475
23.9%	2500
23.1%	2525
22.2%	2550
21.3%	2575
20.6%	2600
19.8%	2625
18.9%	2650
18.1%	2675
17.3%	2700
16.4%	2725
15.6%	2750
14.9%	2775
14.2%	2800
13.4%	2825
12.8%	2850
12.3%	2875
11.8%	2900
11.5%	2925
11.4%	2950
11.6%	2975
12%	3000
12.4%	3025
12.9%	3050
13.5%	3075
14.1%	3100

dealt with the S&P500 so thoroughly—follow up with a small statistic for the S&P500 with regard to the weekend effect.

In principle, it should be noted that:

Many of the so-called anomalies that have been pointed out especially from the 1970s onwards appear to have weakened since around the turn of the millennium. Some of the anomalies are still evident but may be so weak that they cannot be exploited for trading strategies that would lead to reliable excess returns.

Some of the most popular price anomalies—besides the Weekend Effect—are the following (for a good overview, see the article “Anomalies and Market Efficiency” by G. William Schwert [2]):

January effect, weather effect, winner-loser effect (momentum effect), small-firm effect.

To conclude this short excursus, we review the daily prices of the S&P500 from 1990 to November 2018 with regard to a weekend effect and present the results in Tables 1.12, 1.13, 1.14, and 1.15 (for the entire period and for different sub-periods):

The only weekday effect that is noticeable for all the time periods in the above tables is the clearly highest VIX value on Mondays.

Table 1.12 Average returns on different weekdays and time periods

	Jan. 1990– Nov. 2018	Jan. 1990– Dec. 1999	Jan. 2000– Nov. 2018	Jan. 2008– Dec. 2009
Average return on Mon.	0.03%	0.12%	−0.02%	−0.15%
Average return on Tue.	0.06%	0.07%	0.06%	0.18%
Average return on Wed.	0.04%	0.09%	0.01%	−0.15%
Average return on Thu.	0.02%	−0.03%	0.06%	0.04%
Average return on Fri.	0.01%	0.06%	−0.02%	0.02%

Table 1.13 Average overnight returns on different weekdays and time periods

	Jan. 1990–Nov. 2018	Jan. 1990–Dec. 1999	Jan. 2000–Nov. 2018	Jan. 2008–Dec. 2009
Average return	0.003%	0.0003%	−0.005%	−0.025%
Mon. evening till Tue. morning				
Average return	−0.002%	−0.0007%	−0.002%	−0.05%
Tue. evening till Wed. morning				
Average return	−0.001%	0.0007%	0.001%	−0.006%
Wed. evening till Thu. morning				
Average return	0.002%	0%	0.003%	−0.055%
Thu. evening till Fri. morning				
Average return	0.009%	0.0004%	0.014%	−0.006%
Fri. evening till Mon. morning				

Table 1.14 Average percentage change from min to max on different weekdays and time periods

	Jan. 1990–Nov. 2018	Jan. 1990–Dec. 1999	Jan. 2000–Nov. 2018	Jan. 2008–Dec. 2009
Average percentage change from min to max on Mon.	1.24%	1.10%	1.31%	2.47%
Average percentage change from min to max on Tue.	1.27%	1.14%	1.34%	2.34%
Average percentage change from min to max on Wed.	1.26%	1.07%	1.36%	2.36%
Average percentage change from min to max on Thu.	1.29%	1.12%	1.38%	2.66%
Average percentage change from min to max on Fri.	1.24%	1.15%	1.28%	2.28%

1.20 Derivatives on the VIX: VIX Futures

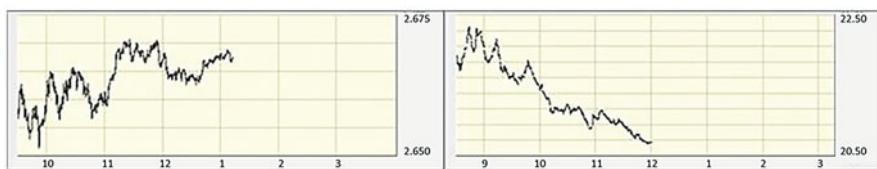
The CBOE offers futures and options as derivatives on the VIX. VIX futures are traded on the CFE, the CBOE Futures Exchange. The CBOE website states:

Introduced in 2004 on Cboe Futures Exchange (CFE), VIX futures provide market participants with the ability to trade a liquid volatility product based on the VIX Index methodology. VIX futures reflect the market's estimate of the value of the VIX Index on various expiration dates in the future. Monthly and weekly expirations are available and trade nearly 24 hours a day, five days a week. VIX futures provide market participants with a variety of opportunities to implement their view using volatility trading strategies, including risk management, alpha generation and portfolio diversification.

VIX futures (ticker symbol: VX) are indeed the most efficient method of trading the VIX. Because of its definition through SPX option prices, the VIX can of course

Table 1.15 Average VIX values on different weekdays and time periods

	Jan. 1990– Nov. 2018	Jan. 1990– Dec. 1999	Jan. 2000– Nov. 2018	Jan. 2008– Dec. 2009
Average	19.55	18.72	20.00	32.69
VIX on Mon.				
Average VIX on Tue.	19.28	18.51	19.69	31.79
Average VIX on Wed.	19.24	18.42	19.67	32.00
Average VIX on Thu.	19.16	18.42	19.56	31.73
Average VIX on Fri.	19.08	18.29	19.50	32.32

**Fig. 1.66** SPX (left) and VIX (right) on 21 November 2018

also be traded directly by trading the corresponding SPX options. However, due to the large number of different options that would need to be traded, in very specific ratios, and exactly at mid-price between bid and ask, it requires a great effort and can only be done approximately.

When you trade VIX futures contracts, just as when you trade futures on individual stocks or on stock indices, you only approximately replicate trading in the underlying asset—the VIX. At expiration, the strike price of the futures contract coincides with the then current value of the VIX. However, over the life of that contract, the strike price generally deviates somewhat from the price of the VIX. This deviation determines the inaccuracy in the replication of a VIX trade through a VIX futures trade.

Figures 1.66 and 1.67 show the performance of the SPX and the VIX and the listed quotes of the VIX futures during the trading day on 21 November 2018. At that time (when this chart was copied), the VIX was just below 20.50 points. The VIX had been falling continuously over the first few hours of the trading day—probably because the SPX had been trending upward during that time. The strikes of the futures—including those with longer-dated expirations—hovered around the price of the VIX. (Bid prices between 19.90 and 20.45, ask prices between 19.95 and 20.50) The spreads between bid and ask were mostly 0.05 dollars, i.e. the smallest possible value.

Instrument		Bid Volume	Bid	Ask	Ask volume
VIX Index	20.8				
VIX Dec19'18 @CFE		71	20.2000	20.2500	89
VIX Jan16'19 @CFE		91	20.4500	20.5000	41
VIX Feb13'19 @CFE		85	20.4000	20.4500	111
VIX Mar19'19 @CFE		38	20.2500	20.3000	62
VIX Apr17'19 @CFE		102	19.9500	20.0500	103
VIX May22'19 @CFE		49	19.9000	20.0000	50
VIX Jun19'19 @CFE		2	19.9000	19.9500	2
VIX Jul17'19 @CFE		34	19.9000	20.0000	8

Fig. 1.67 VIX futures listed on 21 November 2018 in the Interactivebrokers Trader Workstation (same time as in Fig. 1.66)

The following statement in the above quote from the CBOE website *VIX futures reflect the market's estimate of the value of the VIX Index on various expiration dates in the future* is to be questioned, however. As we know, due to the no-arbitrage principle, the fair strike price of a futures contract is well-defined, regardless of the modelling used, and provided that holding the underlying asset does not incur any costs nor generate any income. Holding the VIX obviously does not incur any costs nor generate any income. Thus, the fair strike of a futures contract is well-defined (at least theoretically), regardless of any expectations as to the performance of the VIX, and should be reflected in the quotes. How then is the above statement by the CBOE to be understood? Another page on the CBOE website sheds some light on this inconsistency: While the VIX is well-defined in theory, in practice it can only be replicated explicitly with a certain amount of effort and only by means of continuous adjustments (by trading SPX options continuously). And the longer the contract's life to expiration, the more complex this replication becomes. It is quite possible therefore that the quoted strike prices of VIX futures deviate somewhat from the fair strikes. This deviation (which cannot be excessive, of course, as otherwise arbitrage opportunities could arise) may well be due to certain expectations of market participants as to further price moves of the VIX. These expectations are often driven by the long-term "mean-reversion property" of the VIX that we mentioned in an earlier chapter. The long-term average of the VIX is 19.43% (as on November 2018). The current value of 20.50 is thus slightly above average and suggests a slight decline towards the mean in the long term. This may explain the slightly declining strikes of VIX futures in Fig. 1.67. (Due to slightly positive US interest rates in the fall of 2018, the theoretical fair strikes for longer-dated VIX futures should actually rise to a value slightly above 20.50.) In Fig. 1.68,

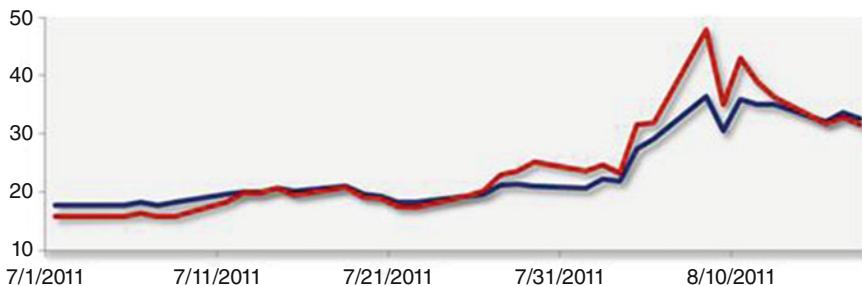


Fig. 1.68 Discrepancy between VIX (red) and strike price of a VIX futures contract (blue) in summer 2011

we see as an example the parallel movements of the VIX (blue curve) and of the strike price for one of the VIX futures contracts quoted at that time in summer 2011. We clearly see a strong deviation around 10 August 2011. What is also clear to see is that the contract's strike price is higher than the VIX, as long as the VIX value is below average (i.e. below 20), but is partly well below the VIX as soon as the VIX rises more sharply, i.e. to above-average values. The VIX futures listed in Fig. 1.67 are not all VIX futures that were actually traded at that time. Up to approximately one and a half months into the future, there generally are almost weekly expirations.

The contract multiplier for each VIX futures contract is 1000.

Let us look, for example, at the VIX futures contract in Fig. 1.67 expiring on 16 January 2019. We can **go long** on a contract with strike 20.50 (ask strike) at the price of 0 dollars.

The current margin requirements for VIX futures are available on the CBOE website at https://markets.cboe.com/us/futures/notices/margin_updates.

The actual amount of the required margin varies from broker to broker. Very roughly speaking, at Interactivebrokers, for example, the maximum margin can be expected to be approximately \$12,500 per VIX futures contract. Contract fees are low (to give you an idea, we are quoting Interactivebrokers again as an example), in the range of around \$2.50.

If the VIX rises to 22 points by the time the futures contract expires on 16 January 2019, the contract will generate a profit of $(22 - 20.50) \times 1000 = \1500 .

If the VIX drops to 18 points by the time the futures contract expires on 16 January 2019, the contract will incur a loss of $(18 - 20.50) \times 1000 = -\2500 .

Of course, the futures contract can also be closed before expiration (at price 0). Profits or losses through changes in the strike price are balanced daily in the investor's account. And as has been said: The strike price of a futures contract can deviate over its life from the value of the VIX.

So if, for example, the VIX stands at 21.00 points on 30 December 2018, the contract's strike price (bid) may well be around 20.40. In that case, due to the daily balancing policy, the investor's account will have a negative balance of

$(20.40-20.50) \times 1000 = -\100 dollars. If the investor closes the futures contract now at price 0, that investor is left with a loss of \$100 in the trading account.

We can **go short** on a contract with strike 20.45 (bid strike) at the price of 0 dollars.

Again, very roughly speaking, we can expect a maximum margin of approximately 12,500 per VIX futures contract.

If the VIX drops to 15 points by the time the futures contract expires on 16 January 2019, the futures contract will generate a profit of $(20.45-15.00) \times 1000 = \5450 dollars.

If the VIX rises to 22 points by the time the futures contract expires on 16 January 2019, the futures contract will result in a loss of $(20.45-22) \times 1000 = \1550 dollars.

By rolling over to futures with longer-dated expirations, the effect of a futures contract can be extended (at basically no cost) to periods of any length.

A Fictitious (!) Trading Experiment with VIX Futures

As a fitting conclusion to this chapter—and purely for enjoyment—we are going to conduct a simple trade experiment under very simplified conditions and simplified assumptions:

Let us assume that the long-term average of the VIX is actually around 20 points. And suppose—put fictitiously and simplistically—that we are able to use the VIX futures to actually track the VIX exactly throughout.

For a VIX futures contract, we reckon a permanent margin requirement of \$12,500 (which needs to be freely available even after daily settlement). We won't list transaction costs separately, being that they will be negligible in the following trading strategy.

The strategy has two parameters that can be chosen freely:

- A “distance” d
- a “ratio”

Distance d defines an “upper limit” $20 + d$ and a “lower limit” $20 - d$. So, for example, for $d = 4$, we get 24 as the upper limit and 16 as the lower limit.

The idea now is this: If the VIX exceeds the upper limit, we assume that in the long term, the VIX will fall below the long-term average of 20 again at some point. We therefore go short on the VIX (using VIX futures) as soon as the VIX exceeds the upper limit. This short position is held (or rolled over) until the VIX falls below 20 again for the first time.

Or conversely:

If the VIX falls below the lower limit, we assume that in the long term, the VIX will rise to a value above the long-term average of 20 again at some point. We therefore go long on the VIX (using VIX futures) as soon as the

(continued)

VIX falls below the lower limit. This long position is held (or rolled over) until the VIX moves back up to a value above 20 for the first time.

In our numerical example, this means:

As soon as the VIX rises above 24, VIX futures contracts are shorted and held (or rolled over) until the VIX falls below 20 for the first time. The futures contracts are then immediately closed out.

As soon as the VIX falls below 16, VIX futures contracts are gone long and held (or rolled over) until the VIX rises above 20 again for the first time. The futures contracts are then immediately closed out.

The “ratio” parameter, which is given in percent (e.g. ratio = 30%), determines how many futures contracts are traded in each case.

For as long as the strategy is run and futures are to be held as per strategy instructions and as long as the investment amounts to more than \$12,500, at least one contract is to be held at any one time. In principle, however, we should hold as many contracts as possible, while ensuring that no more than a maximum of “ratio” percent of the current investment is tied up as margin.

For example:

If ratio = 30% and if the current investment is \$200,000, then no more than \$60,000 at most should be tied up as margin. Since each futures contract requires a margin of \$12,500, we could trade four futures contracts in this case. If the current investment is \$30,000, then under the “ratio rule”, only \$9000 could be used for margin purposes. In this case, no futures contract could be traded. However, since the current investment amounts to more than \$12,500, we can trade one contract. Ratio = 0 means that only one futures contract is traded at a time.

We start with an initial investment of \$20,000 in all subsequent tests.

We could of course also invest that amount at the risk-free interest rate, and the subsequent returns could even be higher. But this is an aspect we are not going to pursue any further.

The “average value” of 20 at which the futures contracts are closed out in each case could also be kept variable. But in virtually all the tests we have run, this value has proven to be essentially the optimum value.

In the following Figs. 1.69, 1.70, 1.71, 1.72, 1.73, 1.74, and 1.75 and the corresponding performance data, you can see how this strategy works out for different time ranges and different “distance” and “ratio” parameters. The charts compare the performance of the S&P500 (red) in percent with the performance of the strategy (blue) in percent over the respective time period.

The given performance data are:

- The minimum excess margin for the period until expiration (excess margin = current investment – required margin). This must be positive; otherwise, the strategy would be terminated before expiration.

(continued)

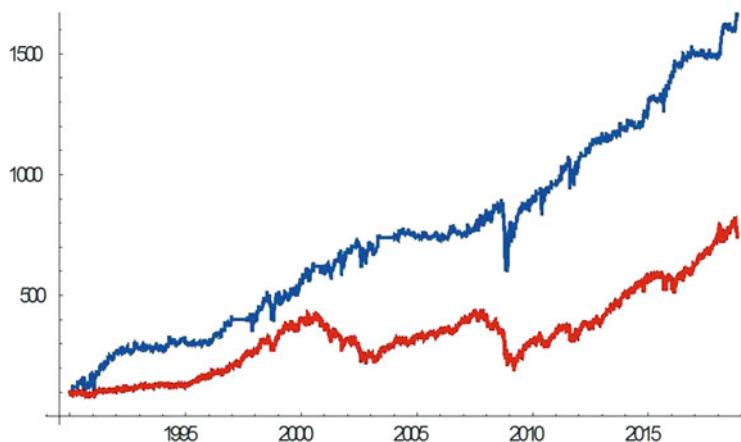


Fig. 1.69 Distance $d = 4$, ratio = 0%, time range: 1990–November 2018

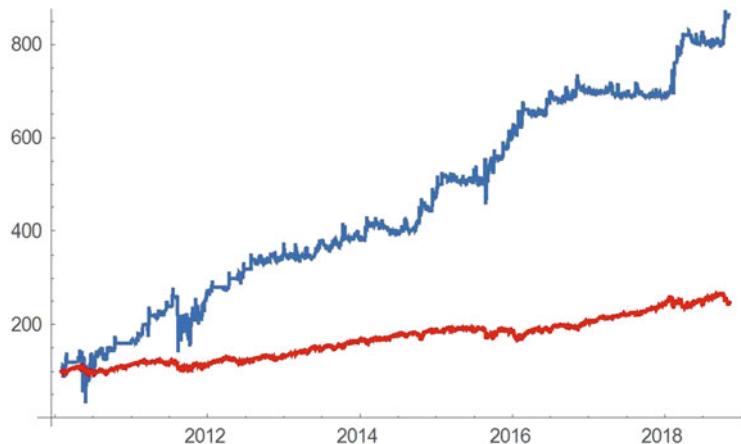


Fig. 1.70 Distance $d = 4$, ratio = 0%, time range: 2010–November 2018

- The final capital amount obtained (starting from \$20,000), in absolute terms and as a percentage
- The strategy's return p.a.
- The strategy's volatility p.a.
- The strategy's Sharpe ratio ($r = 3\%$).

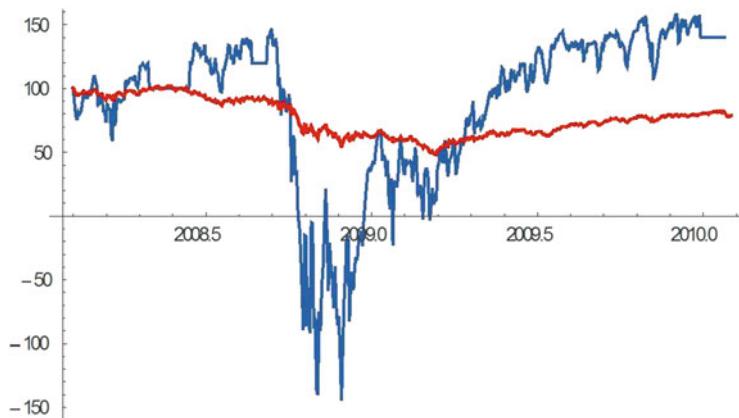


Fig. 1.71 Distance $d = 4$, ratio = 0%, time range: 2008—year-end 2009

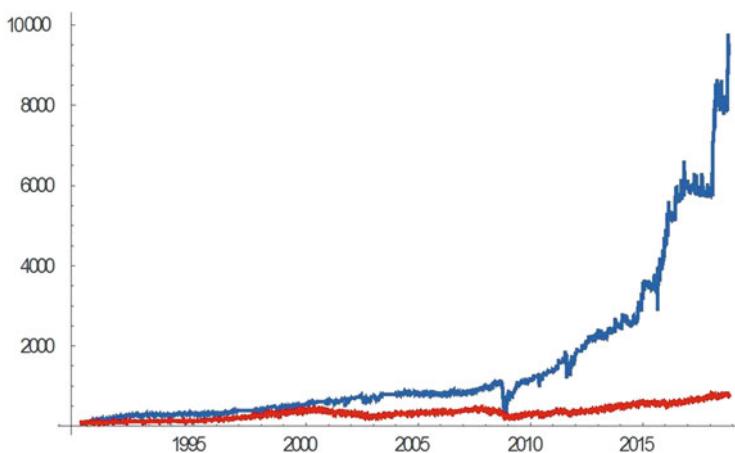


Fig. 1.72 Distance $d = 4$, ratio = 30%, time range: 1990–November 2018

Minimum excess margin (must be positive) throughout the contract's life:
12,750

Final capital amount: 332,650

Final capital as a percentage of initial capital: 1663.25

Return per annum: 10.23%

Volatility of the strategy p.a. in percent: 25.67

Sharpe ratio ($r = 3\%$) of the strategy: 0.28

(continued)

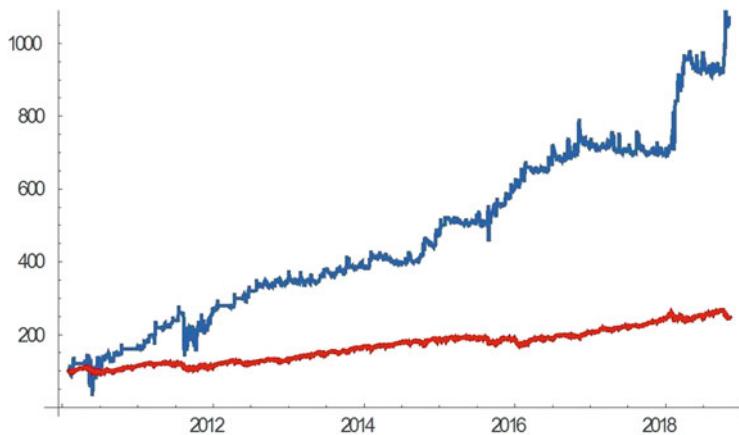


Fig. 1.73 Distance $d = 4$, ratio = 30%, time range: 2010–November 2018

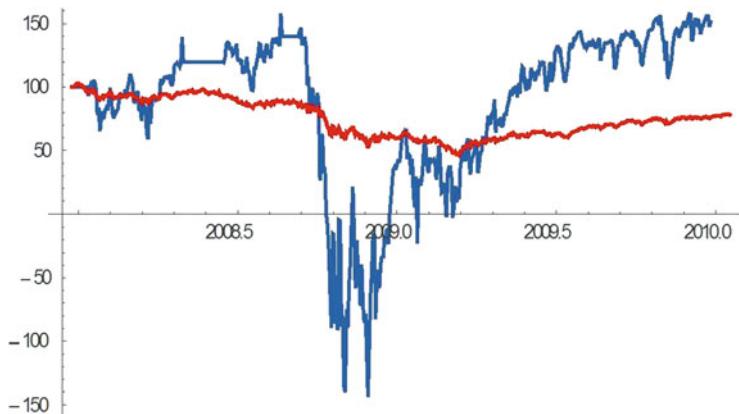


Fig. 1.74 Distance $d = 4$, ratio = 30%, time range: 2008–year-end 2009

Minimum excess margin (must be positive) throughout the contract's life:
2210
Final capital amount: 172,650
Final capital as a percentage of initial capital: 863.25
Return per annum: 27.87
Volatility of the strategy p.a. in percent: 76.05
Sharpe ratio ($r = 3\%$) of the strategy: 0.32

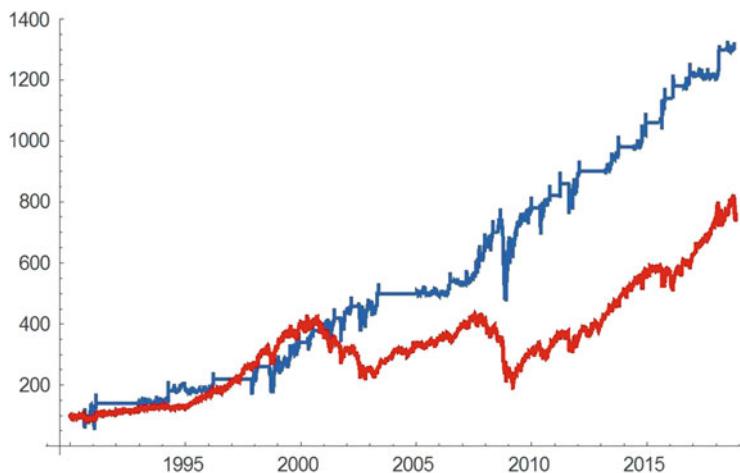


Fig. 1.75 Distance $d = 8$, ratio = 0%, time range: 1990–November 2018

Minimum excess margin (must be positive) throughout the contract's life:
-32,860
 Minimum excess margin (must be positive) throughout the contract's life:
12,750
 Final capital amount: 1,902,200
 Final capital as a percentage of initial capital: 9511
 Return per annum: 17.10
 Volatility of the strategy p.a. in percent: 46.55
 Sharpe ratio ($r = 3\%$) of the strategy: 0.30
 Minimum excess margin (must be positive) throughout the contract's life:
2210
 Final capital amount: 213,950
 Final capital as a percentage of initial capital: 1069.75
 Return per annum: 31.04
 Volatility of the strategy p.a. in percent: 76.84
 Sharpe ratio ($r = 3\%$) of the strategy: 0.36
 Minimum excess margin (must be positive) throughout the contract's life:
-32,860
 Minimum excess margin (must be positive) throughout the contract's life:
3530
 Final capital amount: 263,690
 Final capital as a percentage of initial capital: 1318.45
 Return per annum: 9.35

(continued)

Volatility of the strategy p.a. in percent: 34.92

Sharpe ratio ($r = 3\%$) of the strategy: 0.18

We see a consistently good performance when running this strategy. The choice of $d = 4$ gives the best results in all the tests. It is only if the strategy had been launched at the beginning of the financial crisis, i.e. in about early 2008, that margin requirements would not have been met, resulting in termination of the strategy.

Note that while these experiments are theoretical simulations under simplified assumptions, taking a closer look at such strategies will certainly be worth the effort.

1.21 VIX Options

The CBOE offers options on the VIX for trading. The **contract size for a VIX option is 100\$** and is therefore relatively small. The **VIX options**—like the SPX options—are **European style**.

Details on margin requirements and transaction costs can be found on the CBOE website. We will not go into detailed specifications here. However, to give you a rough idea for the following discussion:

When trading a short position via IB, expect a margin of approximately 2000 dollars plus the current price of the option.

When trading a VIX option contract, expect transaction costs of approximately 1.50 dollars. Since a change in the VIX by one point means that only around 100 dollars are moved per contract, the transaction fee of around 1.50 dollars per contract can add up to a substantially higher amount than, for instance, in the case of SPX options.

Figure 1.76 shows a small sample of VIX options quoted on 21 November 2018.

For example: The **VIX call option** from Fig. 1.76 with expiration on 11 December 2018 and strike 22 is quoted at 1.05 // 1.60. **Buying the call** will likely be possible at around 1.40. Thus, the purchase of a contract costs 140 dollars plus fees of approximately 1.50 dollars.

If the VIX rises to 24 by 11 December 2018, you pocket a payoff of $100 \times (24 - 22) = 200$ dollars on 11 December 2018. The profit in this case is $200 - 140 = 60$ dollars less expenses of about 1.50 dollars. If the VIX falls to 20 by 11 December 2018, the option expires worthless. The loss in this case is 140 dollars (+ fees).

Selling the call will likely be possible at around 1.30. So, by selling a contract, you collect 130 dollars less fees of approximately 1.50 dollars.

If the VIX rises to 24 by 11 December 2018, you pay a payoff of $100 \times (24 - 22) = 200$ dollars on 11 December 2018. The loss in this case is $200 - 130 = 70$ dollars plus fees of about 1.50 dollars. If the VIX falls to 20 by 11 December 2018, the option expires worthless. The profit in this case is 130 dollars (less fees).

The **VIX put option** from Fig. 1.76 with expiration on 11 December 2018 and strike 22 is quoted at 3.10 // 3.90.

Bid x Ask VIX Calls Dec 11'18	impl. vola.	Strike
3.60 x 4.70	83.8%	16
2.95 x 3.90	89.7%	17
2.80 x 3.20	109.3%	18
1.90 x 2.65	100.1%	19
1.55 x 2.20	103.5%	20
1.25 x 1.90	109.1%	21
1.05 x 1.60	112.4%	22
0.85 x 1.40	116.8%	23
1.00 x 1.20	129.2%	24
0.55 x 1.05	122.9%	25

Fig. 1.76 Quotation of VIX options as of 21 November 2018 (excerpt), with VIX at 20.31 (source: Interactivebrokers)

Buying the put will likely be possible at around 3.60. Thus, the purchase of a contract costs 360 dollars plus fees of approximately 1.50 dollars.

If the VIX falls to 18 by 11 December 2018, you pocket a payoff of $100 \times (22 - 18) = 400$ dollars on 11 December 2018. The profit in this case is $400 - 360 = 40$ dollars less expenses of about 1.50 dollars. If the VIX rises to 24 by 11 December 2018, the option expires worthless. The loss in this case is 360 dollars (+ fees).

Selling the put will likely be possible at around 3.40. So, by selling a contract you get 340 dollars less fees of approximately 1.50 dollars.

If the VIX falls to 18 by 11 December 2018, you pay a payoff of $100 \times (22 - 18) = 400$ dollars on 11 December 2018. The loss in this case is $400 - 340 = 60$ dollars plus fees of about 1.50 dollars.

If the VIX rises to 24 by 11 December 2018, the option expires worthless. The profit in this case is 340 dollars (less fees).

Of course, these basic options can again be used to create a wide variety of combinations (in analogy to the strategies discussed earlier). Also interesting in this context are of course strategy considerations with regard to **combinations of VIX options with SPX options**.

To enable more precise analyses in this regard, we obviously need a way to price VIX options. Now, if we were able to model the VIX using a Wiener model and had estimates for the volatility of the VIX, we could use the Black-Scholes formulas for pricing the VIX options.

For estimations as to the volatility of the VIX, we can draw on the VVIX. This index, calculated and published by the CBOE, measures the implied volatility of the VIX, which is in turn calculated from VIX options. Just like the VIX, the

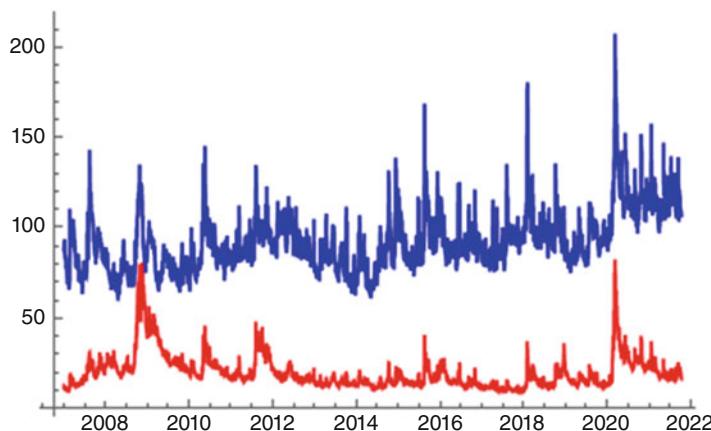


Fig. 1.77 VVIX (blue) and VIX (red) from 2007 to October 2021

VVIX is calculated directly from the prices of VIX options (and not from their implied volatilities). Remember: This calculation is independent of the choice of any particular model!

Figure 1.77 shows the price movements of the VVIX compared to the VIX. We are not going to discuss properties of the VVIX at this point. Interesting information about the VVIX can be found, for example, on: <http://www.cboe.com/products/vix-index-volatility/volatility-on-stock-indexes/the-cboe-vvix-index/vvix-whitepaper>

We see that for the most part, the VIX exhibits rather high volatilities between 60 and 150 and occasionally even higher values than that.

However, the Wiener model is not a suitable tool for modelling the VIX!

For instance, the mean-reversion property of the VIX (i.e. the reversion to a long-term mean) is not replicated in the Wiener model. Usually more appropriate models are therefore used to model the VIX and subsequently price VIX options, such as the Heston model. For an adequate discussion of this model and relevant pricing tasks in this model, we would, however, need certain tools from stochastic analysis, which we will only learn about later. Nevertheless, if we were to disregard for a moment that it is not permissible and were to use the Black-Scholes formulas over the Wiener model to calculate fair prices for, say, the VIX options on 21 November 2018 expiring on 11 December 2018 from the Fig. 1.76, we would use the following parameters and get the following results:

Parameters

Strikes K: from 16 to 25.

S₀ current VIX price: 20.31.

Risk-free interest rate r: 2.1%.

Volatility σ of the VIX: The VVIX values on 21 November were between 99 and 106. We choose the mean at 102.5.

Table 1.16 Results for call and put and the real quotes

Strike	BS price call	Quotes call	BS price put	Quotes put
16	4.69	3.60×4.70	0.37	0.05×0.45
17	3.92	2.95×3.90	0.59	0.30×0.75
18	3.23	2.80×3.20	0.90	0.70×1.10
19	2.63	1.90×2.65	1.30	1.15×1.60
20	2.11	1.55×2.20	1.77	1.75×2.35
21	1.67	1.25×1.90	2.34	2.40×3.10
22	1.31	1.05×1.60	2.97	3.10×3.90
23	1.01	0.85×1.40	3.68	4.00×4.70
24	0.78	1.00×1.20	4.44	4.60×5.60
25	0.59	0.50×1.05	5.25	5.50×6.40

The subsequent Table 1.16 displays the results and real quotes (BS prices within quotes are shown in blue, otherwise in red):

Using the—inadequate—Black-Scholes model, we still get indicative VIX option prices, which—at least in this example—do not deviate much from the actual quotes.

For more accurate analyses of VIX option strategies, however, we have to use more suitable volatility models, as noted above.

But even at this point, we can at least estimate payoff and profit functions of combinations of SPX and VIX options with respect to the expiration date, provided we can confirm certain dependencies between SPX and VIX movements. We will illustrate this briefly with an example in the next section.

1.22 Payoff and Profit Functions of a Trading Strategy for Combinations of SPX and VIX Options

As noted above: We will look just briefly at **one** example of a trading strategy's payoff and profit functions from a combination of SPX and VIX options at expiration. For this purpose, we assume real option price data and that the short-term dependence of the VIX on the SPX is indeed roughly reflected by the formula $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^a$ with an exponent a approximately in the range of $a \approx 4$ (see Sect. 1.13).

All real data we use below refer to the closing prices on Friday, 23 November 2018.

The **SPX value** was **2632.40**.

The **VIX value** was **21.52**.

To illustrate the genesis of this particular strategy, we are going to use a specific VIX call option, namely, the

VIX call option with expiration 11 December 2018 and strike $L = 20$

Quotes: Bid // Ask 1.80 // 2.55 (see Fig. 1.78)

Bid x Ask VIX Calls Dec 11'18	impl. vola.	Strike
3.60 x 4.70	83.8%	16
2.95 x 3.90	89.7%	17
2.80 x 3.20	109.3%	18
1.90 x 2.65	100.1%	19
1.55 x 2.20	103.5%	20
1.25 x 1.90	109.1%	21
1.05 x 1.60	112.4%	22
0.85 x 1.40	116.8%	23
1.00 x 1.20	129.2%	24
0.55 x 1.05	122.9%	25

Fig. 1.78 Quotes on 23 November 2018 of VIX call options expiring 11 December 2018 (excerpt)

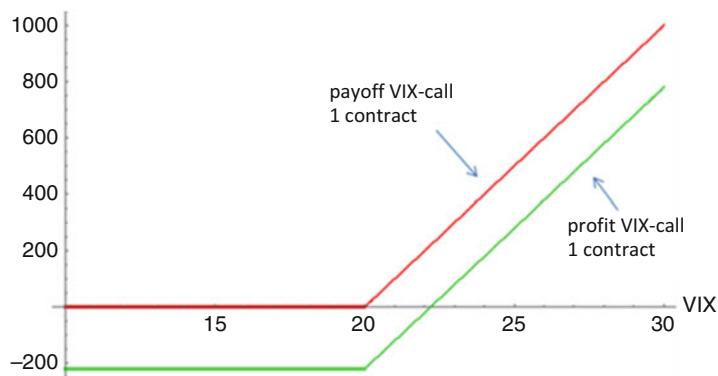


Fig. 1.79 Payoff function (red) and profit function (green) of the VIX call option long position used in the strategy as a function of the VIX at expiration

(Here and in the following, we denote the VIX option strikes by L and the SPX option strikes by K .)

The payoff/profit function of a long position in a contract of this call option as a function of the VIX price on 11 December 2018 is obviously as shown in Fig. 1.79. We are assuming an actual purchase price of \$2.20. If we further assume that the performance of the VIX in the short term (the option's time to expiration being 16 days) is strictly determined by the performance of the SPX, then the payoff function and the profit function of the VIX call option can also be represented as functions of the SPX price on 11 December 2018.

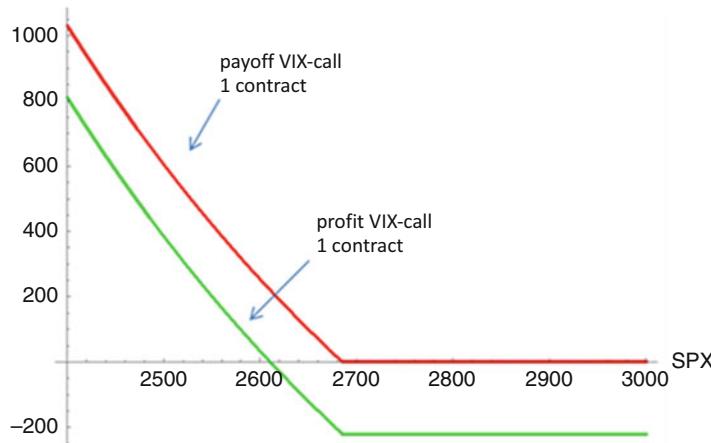


Fig. 1.80 Payoff function (red) and profit function (green) of the long VIX call option as a function of the SPX at expiration assuming dependence $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^4$

For now, as announced above, let us assume the strict dependence $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^4$. This—rather restrictive—assumption will be substantially attenuated at a later point!

The payoff of the VIX call option, namely, $\max(\sigma_t - L, 0)$, thus becomes a function of S_t , namely, $\max\left(\sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^4 - L, 0\right)$. The values σ_0 , S_0 , and L are given.

So, in our specific case, the payoff has the form

$$\max\left(21.52 \cdot \left(\frac{2632.40}{S_t}\right)^4 - 200, 0\right)$$

and the payoff and profit functions of the VIX call option as a function of the SPX price on 11 December 2018 have the form as shown in Fig. 1.80. Since volatility is assumed to be inversely proportional to the SPX, the payoff is now increasing as the SPX falls. The left branch of the payoff function is slightly left-skewed.

The profit function of the VIX option now exhibits (assuming strict dependence of the VIX on the SPX) great similarity to the profit function of an SPX put option with a strike K close to 2680 points.

The obvious thing to do now is to compare this with an SPX put option with strike 2680 and the same expiration as the VIX option. Unfortunately, we don't have the exact same expiration dates available for SPX options as for VIX options. We will therefore use the SPX options with expiration date 10 December 2018 in the following. This deviation in the expiration date should not have any material impact on the following results.

Puts SPX, 10 Dec. 2018 Strike		Bid x Ask
2670	c54.40	60.70 x 68.10
2675	61.50	+4.50 64.10 x 71.60
2680	c59.70	67.20 x 74.50
2685	c62.50	70.30 x 77.80
2690	c65.50	74.10 x 81.10
2695	c68.50	77.40 x 84.50
2700	c71.70	81.00 x 88.10

Fig. 1.81 Quotes on 23 November 2018 of SPX options expiring 10 December 2018 (excerpt)

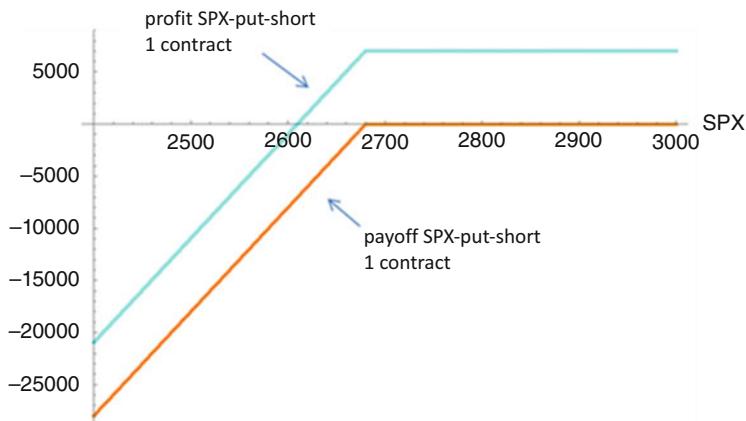


Fig. 1.82 Payoff function (orange) and profit function (green) of the shorted SPX put option used in the strategy as a function of the SPX at expiration

We are thus going to use the following SPX option:

SPX put option with expiration 10 December 2018 and strike $K = 2680$

Quotes: Bid // Ask 67.20 // 74.50 (see Fig. 1.81)

The way we are going to compare the VIX option with the SPX option is by combining a certain number of long positions in the VIX call option with a short position in the SPX put option. Based on the above bid // ask quotes, we expect to be able to short the SPX put at a price of \$70.50. The payoff/profit function for a contract of the short put is shown in Fig. 1.82.

A cursory inspection of the profit functions of the long VIX call and the short SPX put reveals that a combination of about 30 contracts of the VIX option with one contract of the SPX option to the right of strike $K = 2680$ could cause the profit function to be nearly wiped out to a value close to 0. What happens—with

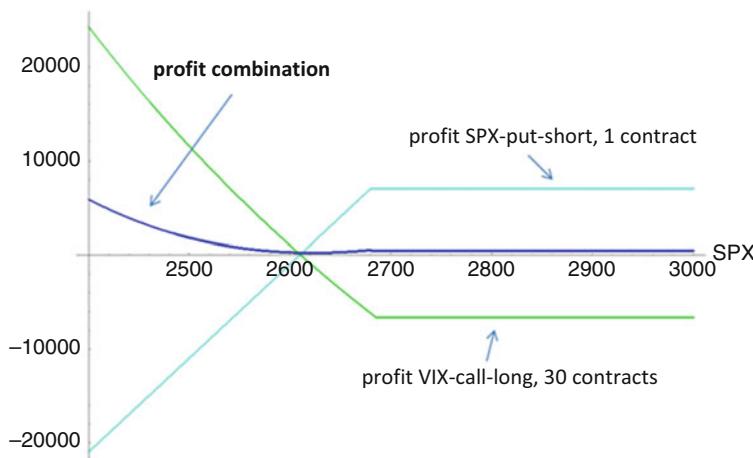


Fig. 1.83 Profit functions of 30 contracts VIX call long (green), one contract SPX put-short (turquoise), and the combination of both (blue) as a function of the SPX at expiration and assuming a relation of the form $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^4$ between SPX and VIX

this combination—to the left of strike $K = 2680$ is what we want to look at in the graph in Fig. 1.83, which plots the profit function of a combination of 30 long VIX call options contracts with one short SPX put options contract. All while keeping in mind that in this representation, we assume dependence of the VIX on the SPX in the form $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^4$.

In the section of this profit function that shows the “critical range” between 2500 points and 2800 points (Fig. 1.84), we can clearly see that the values of the profit function are consistently well above 0. The smallest value is around 220 dollars. For SPX values greater than 2680, the profit is 450 dollars. This means: The above combination represents a strategy which, in case the VIX actually takes the form $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^4$ as a function of the SPX, will always be profitable.

Let us look at the profit function and, in principle, the impact of the options strategy in some more detail:

- If the SPX is at more than 2680 points at expiration, then the SPX put option expires worthless. The VIX call option might yield a profit. But in any case, the premium income of \$450 remains.
- If, at expiration, the SPX is below 2680 points and the VIX value is at least $\sigma_T = \sigma_0 \cdot \left(\frac{S_0}{S_T}\right)^4$ or greater, then the VIX call option yields a profit that is at least as large as the profit in the case of equality $\sigma_T = \sigma_0 \cdot \left(\frac{S_0}{S_T}\right)^4$. The behavior of the SPX put option and the amount of premiums received is not affected by this. As a result, we make a profit that is at or above the blue profit function in Fig. 1.83.

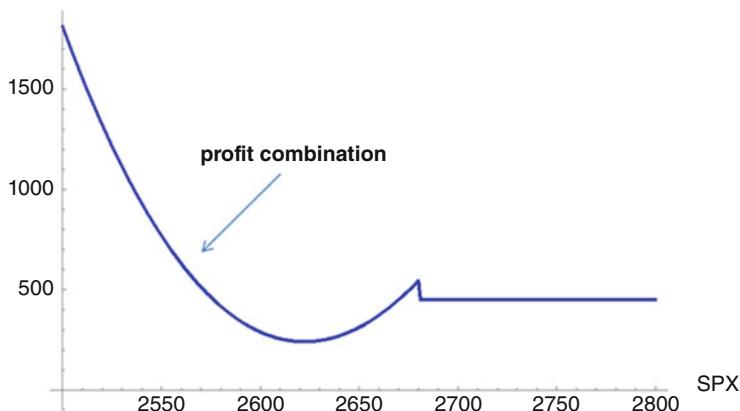


Fig. 1.84 Section from Fig. 1.83

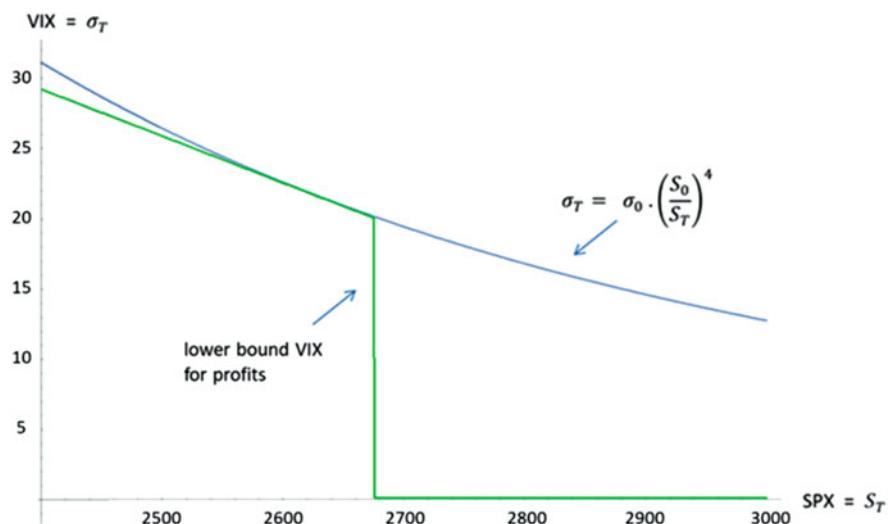


Fig. 1.85 Dependence of the VIX on the SPX of the form $σ_T = σ_0 \cdot \left(\frac{S_0}{S_T}\right)^4$ (blue curve) and lower bound for the value of the VIX as a function of the SPX, so that the options combination (green curve) yields a positive profit

- Losses through the strategy can only be incurred if at expiration, the SPX is below 2680 points and the VIX value is significantly below $σ_T = σ_0 \cdot \left(\frac{S_0}{S_T}\right)^4$.

We have now calculated, for each possible value S_T of the SPX at expiration, how large the VIX would have to be at expiration, so that the strategy would not cause a loss. The blue curve in Fig. 1.85 represents the performance of the VIX as

Table 1.17 SPX and minimum VIX values for a positive profit from the combination strategy

SPX value	2400	2425	2450	2475	2500	2525	2550	2575	2600	2625	2650	2675
Minimum VIX value	29.2	28.4	27.6	26.7	25.9	25.1	24.2	23.4	22.6	21.7	20.9	20.1

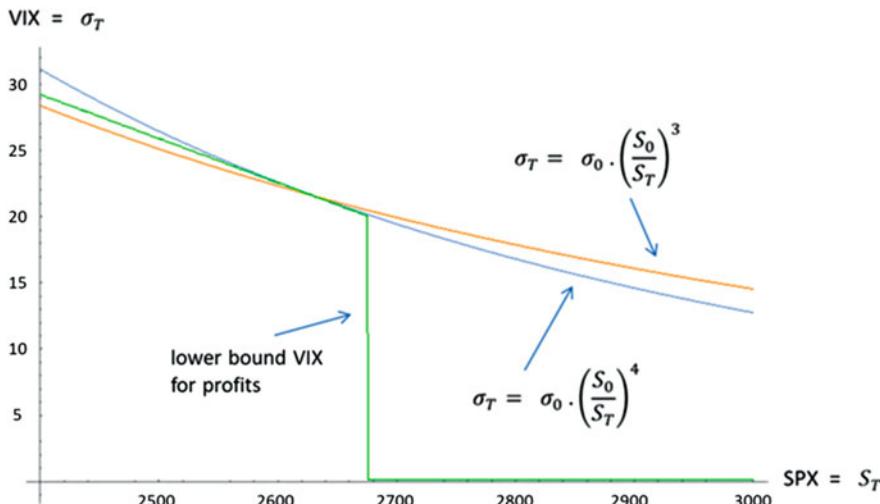


Fig. 1.86 Dependence of the VIX on the SPX of the form $\sigma_T = \sigma_0 \cdot \left(\frac{S_0}{S_T}\right)^4$ (blue curve), of the form $\sigma_T = \sigma_0 \cdot \left(\frac{S_0}{S_T}\right)^3$ (orange curve), and lower bound for the value of the VIX as a function of the SPX, so that the options combination yields a positive profit (green curve)

a function of the SPX, of the form $\sigma_T = \sigma_0 \cdot \left(\frac{S_0}{S_T}\right)^4$. The green curve depicts the lower bound that the VIX would need to have in relation to the respective SPX price in order for the strategy to yield a positive profit.

Table 1.17 lists some more SPX values and the minimum VIX values required to ensure a positive profit from the combination strategy.

For comparison, to illustrate what can happen in the case of a weaker VIX growth and falling SPX, see Figs. 1.86 and 1.87, where we added the performance of the VIX according to the formula $\sigma_T = \sigma_0 \cdot \left(\frac{S_0}{S_T}\right)^3$ (i.e. exponent $a = 3$ instead of exponent $a = 4$) and the profit function of the strategy for this type of dependence and where we then set the results in relation to the case $a = 4$.

In all cases, it should be noted of course, that it may be possible to take profits by closing all positions before expiration of the strategy if the performance of the individual options would suggest that.

This may be particularly interesting if the SPX suddenly starts to fall. Especially shortly after such a decline, the VIX may rise sharply for a short time—due to

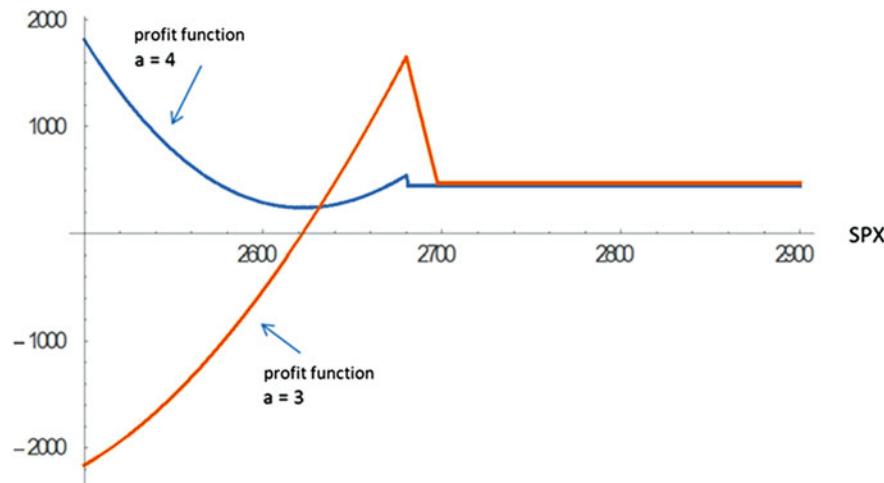


Fig. 1.87 Profit functions of the options combination as a function of the SPX at expiration and assuming a relation of $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^4$ between SPX and VIX (blue curve), and a relation of the form $\sigma_t = \sigma_0 \cdot \left(\frac{S_0}{S_t}\right)^3$ (orange curve)

a potential overreaction of the market—opening up a window of opportunity for taking profits. On the other hand, in such a setting, it is necessary to precisely calculate the impact of this increase in the VIX on the price of the SPX option, of course. A more detailed analysis of these correlations can only be done once we have the adequate techniques for pricing VIX options.

Finally, it should also be noted that in order to execute the strategy, we need an investment of 6600 dollars for buying the VIX call options and an initial margin for the SPX put option in the amount of approximately 15% of the strike, i.e. of approximately 40,000 dollars.

As said before, there would be many other highly interesting strategies related to VIX options (and their combination with other products). As an example, we could also discuss exploiting the presumed long-term “reversion” of the VIX to the long-term mean of around 20 points by systematically shorting VIX options, i.e. shorting both a put and a call option with strike 20 in each case.

However, in order to actually analyse these (and other) strategies in a robust manner, we would need adequate techniques for reliable VIX option pricing, or we would need to conduct studies and backtests over longer periods using historical option data.

References

1. John Hull. *Options, Futures and other Derivatives*. Pearson, 2018.
2. G William Schwert. Anomalies and market efficiency. *Handbook of the Economics of Finance*, Elsevier, 1(Part B):939–974, 2003.



Extensions of the Black-Scholes Theory to Other Types of Options (Futures Options, Currency Options, American Options, Path-Dependent Options, Multi-asset Options)

2

Keywords

Currency options · Futures options · Valuation of American options · Path-dependent options · Barrier options · Monte Carlo methods · Quasi-Monte Carlo methods · Variance reduction methods · Multi-asset options

2.1 Introduction and Discussions So Far

In terms of options pricing, we have so far done, and thoroughly discussed, the following:

We have looked at options where the underlying asset follows a Wiener model (a geometric Brownian motion) and where holding that underlying asset either doesn't incur costs or generate payments or where we can estimate the amount of such costs or payments reliably in advance.

Examples of underlying assets for which modelling with the Wiener model is justifiable:

- Stocks (without dividends or with dividends that are known or can be well estimated)
- Stock indexes (price indexes (dividends not included) or performance-based indexes where good estimates of average dividends are possible)
- Currencies (see later in this chapter)
- Futures on the above products (see later in this chapter)
- Commodities (with no costs incurred or costs that can be well estimated)

Where these options are European-style, have this one underlying asset only, and are non-path-dependent (i.e. the option's payoff is determined solely by the

underlying asset's price at expiration with no regard for previous prices), for these options, we have derived and discussed the general Black-Scholes formula for calculating their fair prices.

Let us recap the Black-Scholes formula once again here for cases without and with costs/payments:

Black-Scholes Formula (for Underlying Assets *without* Payments/Cost)

Let D be a European, non-path-dependent derivative with expiration T and payoff function Φ on an underlying asset with price $S(t)$ that follows a Wiener model with parameters μ and σ in the time range $[0, T]$.

(It is assumed that no payments or costs are incurred through the underlying asset.) The fair price $F(0)$ of D at time 0 is then given by

$$F(0) = e^{-rT} \cdot E(\Phi(\tilde{S}(T)))$$

where the price movement of \tilde{S} over time is

$$\tilde{S}(T) = S(0) \cdot e^{(r - \frac{\sigma^2}{2})T + \sigma\sqrt{T}w}$$

with a standard normally distributed random variable w . “E” in this equation denotes the expectation, and r is the risk-free interest rate $f_{0,T}$.

Black-Scholes Formula (for Underlying Assets *with* Payments/Cost)

Let D be a European, non-path-dependent derivative with expiration T and payoff function Φ on an underlying asset with price $S(t)$ that follows a Wiener model with parameters μ and σ in the time range $[0, T]$.

It is assumed that payments or costs amounting to a constant percentage of $q\%$ per annum will be incurred through the underlying over the option's life (positive q means a payment; negative q means costs).

The fair price $F(0)$ of D at time 0 is then given by

$$F(0) = e^{-rT} \cdot E(\Phi(\tilde{S}(T)))$$

where the price movement of \tilde{S} over time is

$$\tilde{S}(T) = S(0) \cdot e^{(r - q - \frac{\sigma^2}{2})T + \sigma\sqrt{T}w}$$

with a standard normally distributed random variable w . “E” in this equation denotes the expectation, and r is the risk-free interest rate $f_{0,T}$.

The only critical parameters in the Black-Scholes formulas are the volatility σ , which was discussed in detail in the previous chapter, and—in the case of costs or payments—the payment rate q .

The result (in both cases) is stated as an expected value. In a few cases, this expectation can be calculated explicitly, leading to explicit formulas for the price of the option in question. This is possible, for example, in the case of traditional call and put options. In many cases, the expectation has to be approximated. A very useful technique to quickly and easily obtain good approximations of such expectations is the Monte Carlo method, which we will get to know and use in this chapter. Monte Carlo will be our method of choice especially when it comes to path-dependent options.

The aim of this chapter is to provide the tools for pricing most of the options that we have not been able to handle so far, for reasons of not yet having the appropriate methods. More specifically, we will cover the following topics:

- Pricing **currency options**
- Pricing and trading **options on futures**
- Pricing and hedging **American options**
- Pricing (European) **path-dependent options**
- The **Monte Carlo method** for pricing options
- Pricing **multi-asset options** (options on several underlying assets)

We will, of course, illustrate the theoretical explanations again with detailed examples.

Topics that, for the time being, will still be outstanding at the end of this chapter are:

- pricing **American path-dependent options**
- the big chapter on pricing **options on interest rate products**

2.2 Currency Options

The underlying asset of a currency option (also known as a forex or FX option) is always a specific exchange rate (currency pair). The basic version of a European call option or a European put option on an exchange rate looks like this:

Let the foreign currency XXX have a current exchange rate of $X(0)$ in EUR. This means: One XXX currently costs $X(0)$ EUR.

A call option on XXX with strike K and expiration T gives us the right to buy one XXX at the time T at the price of K EUR. The payoff is therefore $\max(0, X(T) - K)$ EUR. Here, $X(T)$ denotes the price of one unit (a fixed amount) of XXX at time T in EUR.

A put option on XXX with strike K and expiration T gives us the right to sell one XXX at time T at the price of K EUR. The payoff is therefore $\max(0, K - X(T))$ EUR.

By buying (going long on) a call option, we are betting that the price $X(t)$ is going to increase, i.e. that XXX will become stronger. The XXX will cost more euros than before.

By buying (going long on) a put option, we bet on a falling price $X(t)$, i.e. on a weakening XXX. The XXX will cost less euros than before.

For example:

On 3 December 2018 at 8 p.m., the rate $X(0)$ of the US dollar expressed in euro was 0.8812 EUR (1 USD costs 0.8812 EUR).

A European put option on the US dollar with strike 0.91 EUR, expiration 16 January 2019, and a ratio of 0.01:1 was quoted at 3.39 // 3.44 EUR.

The ratio indicates that one option is quoted on the basis of 100 US dollars. (More precisely: 0.01 options are required to trade 1 dollar.)

So, by buying (going long on) this put option, we have acquired the right to sell a total of 100 USD at 91 EUR on 16 January 2019.

The payoff for such an option is therefore $100 \cdot \max(0, 0.91 - X(T))$.

Now, how can we price such a European FX option?

Let us again assume that the currency rate $X(t)$ can be modelled within a Wiener model. And we ask that question not only specifically for calls and puts but for options with any payoff $\Phi(X(T))$. But then we already know how to determine the fair price of this option. The only question is whether holding the underlying asset (in this case, the foreign currency) will incur costs or payments over the option's life. The answer to that determines which of the two Black-Scholes formula versions—as summarized in the previous paragraph—should be used.

Holding the foreign currency automatically implies (as agreed on above) that we always invest that foreign currency at its risk-free interest rate. Thus, we receive continuous payments as a percentage of the foreign currency interest rate, which we denote by r_f . We continue to use r to denote the risk-free euro interest rate. It is therefore necessary to use the second version of the Black-Scholes formula (with payments), with $q = r_f$, which gives us:

Theorem 2.1 *Let D be a European, non-path-dependent derivative with expiration T and payoff function Φ on the exchange rate $X(t)$ of a foreign currency XXX where that $X(t)$ follows a Wiener model with parameters μ and σ in the time range $[0, T]$. $X(t)$ denotes the price of 1 XXX in EUR at time t . The fair price $F(0)$ of D at time 0 is then given by*

$$F(0) = e^{-rT} \cdot E(\Phi(\tilde{S}(T)))$$

where the price movement of \tilde{S} over time is

$$\tilde{S}(T) = S(0) \cdot e^{(r-r_f-\frac{\sigma^2}{2})T+\sigma\sqrt{T}w}$$

with a standard normally distributed random variable w . “E” denotes the expected value. r is the risk-free euro interest rate for the period $[0, T]$, and r_f is the risk-free interest rate of the foreign currency XXX for the period $[0, T]$.

Specifically for foreign currency call options and foreign currency put options, we have the following explicit formulas:

Theorem 2.2 *For the fair price $C(t)$ of a call option with expiration T and strike K at time $t \in [0, T]$ on a foreign currency XXX with a price $X(t)$ that follows a Wiener model with parameters μ and σ ,*

$$C(t) = e^{-r_f \cdot (T-t)} \cdot X(t) \cdot \mathcal{N}(\tilde{d}_1) - e^{-r(T-t)} \cdot K \cdot \mathcal{N}(-\tilde{d}_2)$$

with

$$\tilde{d}_1 = \frac{\log\left(\frac{X(t)}{K}\right) + \left(r - r_f + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}$$

and

$$\tilde{d}_2 = \frac{\log\left(\frac{X(t)}{K}\right) + \left(r - r_f - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}$$

and \mathcal{N} being the distribution function of the standard normal distribution.

Theorem 2.3 *For the fair price $P(t)$ of a put option with expiration T and strike K at time $t \in [0, T]$ on a foreign currency XXX with a price $X(t)$ that follows a Wiener model with parameters μ and σ ,*

$$P(t) = e^{-r(T-t)} \cdot K \cdot \mathcal{N}(-\tilde{d}_2) - e^{r_f \cdot (T-t)} X(t) \cdot \mathcal{N}(-\tilde{d}_1),$$

with \tilde{d}_1 and \tilde{d}_2 as in the theorem above and \mathcal{N} being the distribution function of the standard normal distribution.

Example 2.4 We use the explicit put price formula to determine the implied volatility (in terms of bid and ask prices) for the above put option as on 3 December 2018 on the dollar. To do this, we need to solve the equations

$$\begin{aligned} & e^{-rT} \cdot K \cdot \mathcal{N}\left(-\frac{\log\left(\frac{X(0)}{K}\right) + \left(r - r_f - \frac{\sigma^2}{2}\right) \cdot T}{\sigma\sqrt{T}}\right) - \\ & - e^{-r_f T} \cdot X(0) \cdot \mathcal{N}\left(-\frac{\log\left(\frac{X(0)}{K}\right) + \left(r - r_f + \frac{\sigma^2}{2}\right) \cdot T}{\sigma\sqrt{T}}\right) = 3.63 \end{aligned}$$

(= bid price)

and

$$e^{-rT} \cdot K \cdot \mathcal{N} \left(-\frac{\log \left(\frac{X(0)}{K} \right) + \left(r - r_f - \frac{\sigma^2}{2} \right) \cdot T}{\sigma \sqrt{T}} \right) - e^{-r_f T} X(0) \cdot \mathcal{N} \left(-\frac{\log \left(\frac{X(0)}{K} \right) + \left(r - r_f + \frac{\sigma^2}{2} \right) \cdot T}{\sigma \sqrt{T}} \right) = 3.66$$

(= ask price)

for the parameter σ .

The other parameters are as specified in the above example:

$$K = 0.91$$

$$X(0) = 0.8812$$

$$T = \frac{44}{365}$$

Further, we can determine the shortest-term euro interest rates and dollar interest rates for 3 December 2018 through financial data providers (at no cost, e.g. via www.finanzen.net), as follows:

$$r = -0.004613$$

$$r_f = 0.021775$$

By plugging in these values, we obtain the two above equations with only one single parameter σ , and solving the equations with a mathematical software yields the implied volatilities

$$\sigma_{bid} = 0.0970 \sim 9.70\%$$

$$\sigma_{ask} = 0.1038 \sim 10.38\%$$

Currency options can however be specified in different ways. See, for instance, the following **real example** of currency options, offered by the German Commerzbank. Note that Commerzbank's certificate business was recently taken over by Société Générale.

On 4 December 2018, the option offered by Commerzbank had the following parameters:

Type: call option

Strike price: 1.10

Expiring: 16 January 2019

Exercise ratio 0.01:1

Quotation (bid/ask): 3.99 // 4.03 USD

Current underlying price: 1.1388 USD

Note that the underlying asset's strike price and current price are given in dollars.

The value 1.1388 is the price of 1 euro in dollars (and not, as in the above “standard example”, the price of 1 USD in EUR), and consequently, the strike price of 1.10 is the critical value for the price of 1 euro in dollars.

So, this call option gives us the right to buy 100 EUR at time T in exchange for 110 USD (or, put differently, to sell 110 USD in exchange for 100 EUR). We will exercise this right if, at time T , the price of 1 EUR in USD is more than 1.10 USD (or, put differently, if the price of 1 USD in EUR is below $\frac{1}{1.10} = 0.909$ EUR).

So, given how we basically understand options, the above specifications would result in a payoff of $100 \times \max(0, Y(T) - 1.10)$ USD! Here, $Y(T)$ denotes the price of 1 EUR in USD at time T . (In contrast, we still use $X(T)$ to denote the price of 1 USD in EUR. $Y(T) = \frac{1}{X(T)}$ of course.)

Observe that the payoff is stated in dollars. However, it is quite confusing that the payout is to be made in dollars. We therefore refer to the option's product specifications, which were also available on the Commerzbank website. The relevant excerpt from the product specifications is reproduced in Fig. 2.1, with the definition of “payout amount” in the second paragraph.

As we can see, our interpretation of the payoff as “ $100 \times \max(0, Y(T) - 1.10)$ dollars” was indeed correct. However, this amount (see the second line in the definition of the “payout amount”) is not paid out in dollars but first “converted into EUR ...” before it is actually paid out. This conversion is done at the time the payout is effected, i.e. at the exchange rate at expiration. The actual payoff is therefore $100 \times \max(0, Y(T) - 1.10) \times X(T)$ EUR!

\$4 Redemption

1. The options grant the holder of securities the right (the “Option Right”) to receive from the Issuer, in accordance with these Terms and Conditions, payment of a Payout Amount (rounded up or down, as the case may be, to the nearest EUR 0.01).

The “**Payout Amount**” per option is the amount, expressed in USD, multiplied by the Multiplier and converted into EUR, by which the Reference Price determined on the Valuation Date is higher (for call warrants) or lower (for put warrants) than the Strike Price.

The “**Strike Price**” is the value stated in the product's Static Data.

The “**Multiplier**” is expressed as a decimal number and corresponds to the ratio specified in the product's Static Data.

Fig. 2.1 English translation of an excerpt from the Commerzbank product specifications for EUR/USD currency options

We rewrite this expression somewhat (factoring out 1.10 and multiplying $X(T)$ into the parentheses and observing that $X(T) \times Y(T) = 1$), which gives us

$$100 \times \max(0, Y(T) - 1.10) \times X(T) \text{ Euro} = 100 \times 1.10 \times \max\left(0, \frac{1}{1.10} - X(T)\right) \text{ EUR.}$$

However, this last expression also denotes precisely 1.10 times the payoff of a put option (in our original conventional sense) with strike $\frac{1}{1.10} = 0.909$ on the price of 1 USD in EUR.

Remember: The quotes for this conventional put option in our example were 3.39 // 3.44. The quotes of the Commerzbank calls would therefore likely be somewhere in the range of $1.10 \times 3.39 // 3.44 = 3.73 // 3.78$, which, of course, is still slightly different from the actual quotes for the Commerzbank option, i.e. 3.99 // 4.03.

This is because the Commerzbank option was of the American type, which would explain the slightly higher quotes.

So when pricing and trading currency options, pay very close attention to the actual specifications.

2.3 Futures Options

Derivatives can also be based on other derivatives as their underlying asset. This is true of, for example, futures options. While they are often perceived as dauntingly complex constructs at first glance, futures options are in fact traded quite heavily in certain areas. For illustration, let us look at options on the SPX mini futures.

In Fig. 2.2, we have again taken just a very small excerpt of quotes for existing options on the SPX mini futures that can be traded via the IB Trader Workstation.

Note that these specific options are American-style.

SPX-futures options Calls Mar' 19/Mar' 19	impl. vola.	Strike
123.25 x 124.25	16.6%	2725
105.75 x 107.25	16%	2750
103.00 x 103.75	15.9%	2755
99.75 x 100.50	15.8%	2760
96.50 x 97.25	15.6%	2765
93.25 x 94.00	15.6%	2770
90.25 x 91.00	15.5%	2775
75.25 x 76.00	14.8%	2800

Fig. 2.2 Quotes for SPX futures options (excerpt) on 4 December 2018

Nevertheless, for the time being, we will only look at European-style futures options.

We now have two different expirations to take into account, the expiration date of the futures contract, which we denote by T_F , and the expiration date of the option, which we denote by T_O , needless to say that $T_O \leq T_F$. In many cases, T_O is very close to T_F or coincides with T_F .

Where there are several futures contracts on the underlying asset (the SPX in our example), then an option with expiration date T_O usually relates to the futures contract with the shortest time to expiration T_F that is greater than or equal to T_O .

The table in Fig. 2.2 shows futures options with three different expirations, namely:

- 21 December 2018 (this option type refers to the contract expiring 21 December 2018 (see first data row in Fig. 2.2)),
- 15 March 2019 (this option type refers to the contract expiring 15 March 2019 (see second data row in Fig. 2.2)),
- 21 June 2019 (this option type refers to the contract expiring 21 June 2019 (see third data row in Fig. 2.2)).

To illustrate how this works, we will again examine a specific example, as follows:

Type: call option

Underlying: SPX Mini Future ES March 15, 2019

Contract size: 50 (i.e. one options contract relates to one mini futures contract)

Expiration: 15 March 2019 (101 days)

Strike: 2770

Quotes (Bid // Ask): 93.25 // 94.00

Underlying price: 2777.25 (this is the contract's current strike price, as the futures price is always 0)

This option can be purchased (long position) at around 94.00 dollars.

If at the option's expiration, the price (strike price) of the futures contract is $F(T_O)$ and $F(T_O) < K$, then the option expires worthless.

If at the option's expiration, the price (strike price) of the futures contract is $F(T_O)$ and $F(T_O) > K$, then the holder of the option will receive a payoff of $F(T_O) - K$.

Whether a payoff is achieved or not is determined solely by how the strike price of the futures moves. Since a futures strike price moves pretty much in lockstep with the underlying asset (the S&P500 in our case), the Wiener model is certainly a suitable model for simulating the strike price movement.

Now we need to decide if holding the futures contract results in a payout on the investment used for purchasing it. The answer to that determines whether we should choose the first version of the Black-Scholes formula (without payments) or the second version of the Black-Scholes formula (with payments) to price the contract.

Recall, however, that purchasing a futures contract requires an investment in the amount of 0. (The price of a futures contract is 0!) It is difficult to decide therefore whether (and in what amount) holding a futures contract will lead to a fixed percentage payment on the investment made. For this reason, to get an idea, we will once again use a one-step binomial model for pricing a futures option, as follows:

The payoff of a futures option depends on that futures strike price (not on its price, of course, which is always 0).

We will therefore apply the conventional binomial model (compare Fig. 2.3) to the price path of the futures strike price. We denote the strike price by K and assume that it will move either up to $u \cdot K$ or down to $d \cdot K$. As usual, we denote the option's respective payoffs by f_u or f_d .

Here, T denotes T_O , that is, the option's expiry.

r is again the risk-free interest rate for the period $[0, T]$.

The futures price at time 0 (or, actually, at **any** time) is equal to 0.

If the futures strike price rises to $u \cdot K$ by T , a payoff of $u \cdot K - K = K \cdot (u - 1)$ is received from the futures contract at time T .

if the futures strike price falls to $d \cdot K$ by T , a payoff of $d \cdot K - K = K \cdot (d - 1)$ is received from the futures contract at time T .

This applies if the option's expiration T is exactly equal to the futures contract's expiration T_E . But it also holds, however, when $T < T_E$:

Because profits and losses from futures are settled on a continuous basis, an amount of $u \cdot K - K$ or of $d \cdot K - K$ has in fact already been credited to the investor's

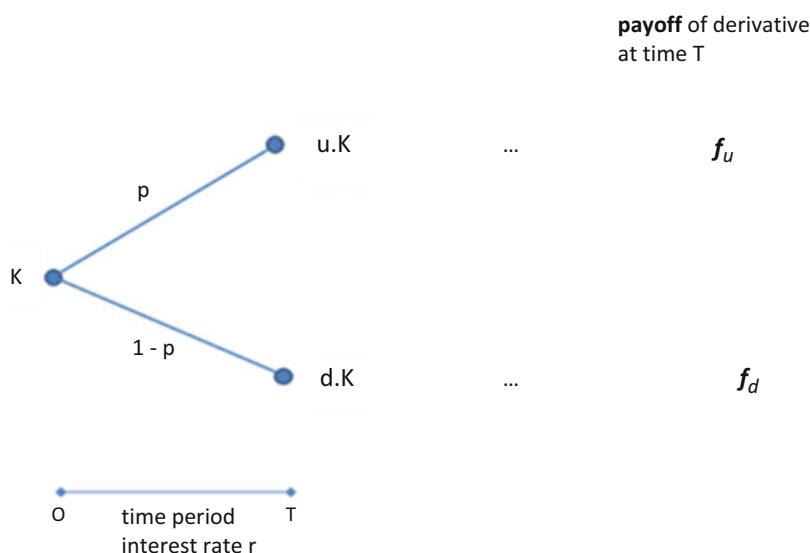


Fig. 2.3 One-step binomial model for futures options

account by time T . The futures contract itself can then simply be sold at the price of 0 at time T .

Let us again attempt to replicate the option's payoff by investing in a portfolio of x futures and of y euros in cash. This is achieved by solving the system of equations

$$\begin{aligned}x \cdot K \cdot (u - 1) + y \cdot e^{rT} &= f_u \\x \cdot K \cdot (d - 1) + y \cdot e^{rT} &= f_d\end{aligned}$$

Solving the system of equations yields the two solutions

$$x = \frac{f_u - f_d}{K \cdot (u - d)} \text{ and } y = e^{-rT} \cdot \left(f_u - \frac{(u - 1) \cdot (f_u - f_d)}{u - d} \right).$$

The price of this replicating portfolio at time 0 (since the price of the futures contract at time 0 equals 0) is simply $x \cdot 0 + y = y = e^{-rT} \cdot \left(f_u - \frac{(u - 1) \cdot (f_u - f_d)}{u - d} \right)$.

The option's price f_0 at time 0 is now equal to the replicating portfolio's price at time 0; thus (after some rearrangement),

$$\begin{aligned}f_0 &= e^{-rT} \cdot \left(f_u - \frac{(u - 1) \cdot (f_u - f_d)}{u - d} \right) = \\&= e^{-rT} \cdot \left(f_u \cdot \frac{1 - d}{u - d} + f_d \cdot \left(1 - \frac{1 - d}{u - d} \right) \right).\end{aligned}$$

The artificial probability p' with respect to which the option's price is the discounted expected payoff is now given by $p' = \frac{1-d}{u-d}$.

So, instead of using the risk-neutral probability $p' = \frac{e^{rT}-d}{u-d}$ (in the case without payments), we now work with an artificial probability that can be given as $p' = \frac{e^{(r-q)\cdot T}-d}{u-d}$ (in the case with payments) with $q = r$.

To summarize: In the binomial model, an option on a futures contract is priced in the same way as an option on an underlying asset with continuous payments at a per annum interest rate of $q = r$. Consequently, we know that such an option is priced in the same way in a Wiener model. So we get:

Theorem 2.5 *Let D be a European, non-path-dependent derivative with expiration T and payoff function Φ on a futures contract F with a strike price $K(t)$ that follows a Wiener model with parameters μ and σ in the time range $[0, T]$.*

The fair price $F(0)$ of D at time 0 is then given by

$$F(0) = e^{-rT} \cdot E(\Phi(\tilde{K}(T)))$$

where the price movement of \tilde{K} over time is

$$\tilde{K}(T) = K(0) \cdot e^{-\frac{\sigma^2}{2}T + \sigma\sqrt{T}w}$$

with a standard normally distributed random variable w . “E” in this equation denotes the expected value, and r is the risk-free interest rate for the time range $[0, T]$.

Specifically for futures call options and futures put options, we have the following explicit formulas:

Theorem 2.6 For the fair price $C(t)$ of a call option with expiration T and strike L at time $t \in [0, T]$ on a futures contract with strike price $K(t)$ that follows a Wiener model with parameters μ and σ ,

$$C(t) = e^{-r(T-t)} \cdot (K(t) \cdot \mathcal{N}(\tilde{d}_1) - L \cdot \mathcal{N}(-\tilde{d}_2))$$

where

$$\tilde{d}_1 = \frac{\log\left(\frac{K(t)}{L}\right) + \frac{\sigma^2}{2} \cdot (T-t)}{\sigma\sqrt{T-t}}$$

and

$$\tilde{d}_2 = \frac{\log\left(\frac{K(t)}{L}\right) - \frac{\sigma^2}{2} \cdot (T-t)}{\sigma\sqrt{T-t}}$$

and \mathcal{N} being the distribution function of the standard normal distribution.

Theorem 2.7 For the fair price $P(t)$ of a put option with expiration T and strike L at time $t \in [0, T]$ on a futures contract with strike price $K(t)$ that follows a Wiener model with parameters μ and σ ,

$$P(t) = e^{-r(T-t)} \cdot (L \cdot \mathcal{N}(-\tilde{d}_2) - K(t) \cdot \mathcal{N}(-\tilde{d}_1)),$$

with \tilde{d}_1 and \tilde{d}_2 as in the theorem above and \mathcal{N} being the distribution function of the standard normal distribution.

Comment 2.8 We know that for reasons of no-arbitrage bounds, the fair strike price of a futures contract is always $K(t) = S(t) \cdot e^{r \cdot (T_f - t)}$. Here, $S(t)$ denotes

the price curve of the underlying asset. If we substitute this expression for $K(0)$ in $\tilde{K}(T) = K(0) \cdot e^{-\frac{\sigma^2}{2}T + \sigma\sqrt{T}w}$, we get the dynamics $\tilde{K}(T) = S(0) \cdot e^{rT_f - \frac{\sigma^2}{2}T + \sigma\sqrt{T}w}$ for $\tilde{K}(T)$.

In the specific case that $T_f = T_0 = T$, we thus get $\tilde{K}(T) = S(0) \cdot e^{(r - \frac{\sigma^2}{2})T + \sigma\sqrt{T}w}$.

And with that, we are right back to the case of pricing options on the underlying asset itself. The value of an option on a futures contract when $T_f = T_0$ is thus exactly equal to the value of an option on the underlying asset itself (with all other parameters remaining the same).

Futures options are also frequently offered on interest rate futures. One example of that is German Bund futures options. However, in the case of such options on interest rate futures, the above considerations are not applicable, since the Wiener model is not suitable for modelling price movements of interest rate futures.

2.4 Valuation of American Options and of Bermudan Options Through Backwardation (the Algorithm)

In general, there are no explicit formulas for pricing American options on underlying assets in the Wiener model. However, we can price American options (assuming they are not path dependent) quite well using a so-called “backwardation” procedure in the binomial model and in this way obtain arbitrarily accurate approximations through the Wiener model.

Let us begin this chapter on the valuation of American options by noting that, under certain circumstances, it is indeed possible to explicitly price American call and put options.

For this purpose, let us first assume a positive risk-free interest rate r and consider an American call option with expiration T and strike K on an arbitrarily chosen underlying without payments and costs. We denote the fair price of this **American** call option at time t by $C_a(t)$. The fair price of its **European** counterpart with the same parameters is denoted by $C_e(t)$.

It will always hold that $C_a(t) \geq C_e(t)$, of course; the American option is always worth at least as much as its European counterpart. “Otherwise, we would...”, we leave this for the reader to complete. With this, and from Volume I Section 3.2, we thus know that the following relationship follows from the put-call parity equation:

$$C_a(t) \geq C_e(t) = P_e(t) + S(t) - K \cdot e^{-r(T-t)} > S(t) - K$$

For the last inequality, we also used that $r > 0$. This means that the call option’s value $C_a(t)$ is always greater than $S(t) - K$. Further, $C_a(t) > 0$ always, of course. Hence, $C_a(t) > \max(S(t) - K, 0)$. However, this value $\max(S(t) - K, 0)$ is precisely the payoff that we would receive if the option were exercised immediately. From that follows that nobody would ever utilize the right to exercise early, since selling the option at the price of $C_a(t)$ would always yield a higher payoff than exercising

the option. Therefore, the American option has no added value compared to the European option of the same type. Thus, we have $C_a(t) = C_e(t)$!

In the case that $r < 0$, we cannot argue for the call option in this way. For the put option, however, we can proceed analogously, as follows:

$$P_a(t) \geq P_e(t) = C_e(t) + K \cdot e^{-r(T-t)} - S(t) > K - S(t).$$

$\max(0, K - S(t))$ is precisely the instantaneous payoff of the American put at time t , and using the same reasoning as above, it follows that in this case $P_a(t) = P_e(t)$. Consequently:

Theorem 2.9

- (a) If the risk-free interest rate r for a time range $[0, T]$ is **greater than or equal to 0**, then for each point in time t in $[0, T]$, the following holds for the prices $C_a(t)$ and $C_e(t)$ of an American and a European call option, respectively, on an underlying asset without payments and costs and with expiration in T and also otherwise the same parameters:

$$C_a(t) = C_e(t)$$

and the American option should never be exercised early.

- (b) If the risk-free interest rate r for a time range $[0, T]$ is **less than or equal to 0**, then for each point in time t in $[0, T]$, the following holds for the prices $P_a(t)$ and $P_e(t)$ of an American and a European put option, respectively, on an underlying asset without payments and costs and with expiration in T and also otherwise the same parameters:

$$P_a(t) = P_e(t)$$

and the American option should never be exercised early.

Note that the theorem only tells us that—in an optimal approach—the American option should not be exercised in that particular case, yet it does not provide any information on whether, at any point in time, we should continue to hold the option or close it out or what kind of strategy should be pursued going forward. Information about whether to exercise an American option, and the arguments to support that information, will in all subsequent results also be:

Either:

Exercise the option now at the current price of the underlying asset, because: If you don't, then there would be a strategy with exercise that would definitely yield better results than without exercise!

Or:

Do not exercise the option now at the current price of the underlying asset, because: If you do, then there would be a strategy without exercise that would definitely yield better results than with exercise!

As previously announced, we will now derive and prove the valuation of American options through backwardation in an N-step binomial model. By approximating a Wiener model to arbitrary accuracy using an N-step binomial model (assuming suitable parameters in that binomial model), we can price American options on underlying assets in the Wiener model with arbitrary accuracy (just like we did for European options). The corresponding program on our website allows you to do exactly that. See <https://app.lsqf.org/binomial-model/derivative-valuation>.

We are now going to proceed by first explaining and proving backwardation in the two-step binomial model. This can then be extended to the N-step model in an obvious way and could be proved exactly by induction. For this purpose, we assume a very general setting in the two-step model. At first glance, the two-step model in Fig. 2.4, which we will use to explain backwardation, seems identical to the two-step model we used to price the European options. However, we now have a value denoted by \tilde{f} and the corresponding index at every single node of the tree. This value is the instantaneous payoff if the American (!) option were exercised at the respective node.

For example, \tilde{f}_u denotes the payoff if the American option were exercised at time dt and an underlying price of $u \cdot S_0$.

Further notations that we are going to introduce at each node of the tree are the values $f^{(a)}$ and $f^{(e)}$ with the corresponding index in each case. Here, $f^{(a)}$ represents the fair price of the American option in the respective node of the tree, and $f^{(e)}$ represents the price of the same option but European-style.

So, what we are looking for is $f_0^{(a)}$, the price of the American option at time 0, and for each node, we are going to decide whether or not to exercise the American option at that node.

- Evidently, for all indices at time T (i.e. for uu , du , dd), we always have $\tilde{f} = f^{(a)} = f^{(e)}$.
- We also know the following about the European option prices at time dt :

$$f_u^{(e)} = e^{-rdt} \cdot \left(p' \cdot \tilde{f}_{uu} + (1 - p') \cdot \tilde{f}_{ud} \right)$$

$$f_d^{(e)} = e^{-rdt} \cdot \left(p' \cdot \tilde{f}_{ud} + (1 - p') \cdot \tilde{f}_{dd} \right)$$

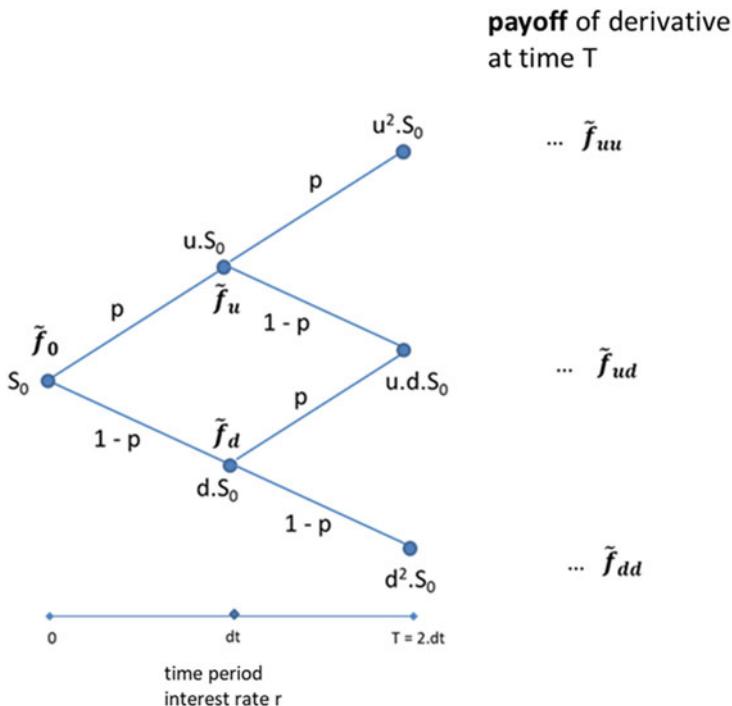


Fig. 2.4 Two-step binomial model for pricing an American option

Because of $\tilde{f} = f^{(a)}$ at time T , this is equivalent to

$$f_u^{(e)} = e^{-rdt} \cdot \left(p' \cdot f_{uu}^{(a)} + (1 - p') \cdot f_{ud}^{(a)} \right)$$

$$f_d^{(e)} = e^{-rdt} \cdot \left(p' \cdot f_{ud}^{(a)} + (1 - p') \cdot f_{dd}^{(a)} \right)$$

- Let us assume that we are at the node $u \cdot S_0$ at time dt . We now have two possible courses of action:

We can exercise the American option and receive a payoff of \tilde{f}_u . Or we don't exercise and keep the option. In that case however, the option is in fact a European option (since at time $T = 2 \cdot dt$, an early exercise is no longer possible), and its value is therefore $f_u^{(e)}$. Of the two possible courses of action, it would be logical to choose the one that gives us the greater value, i.e.:

If $\tilde{f}_u > f_u^{(e)}$, we exercise,
if $\tilde{f}_u < f_u^{(e)}$, we don't exercise.
(If $\tilde{f}_u = f_u^{(e)}$, it doesn't matter which we choose.)

And in this case, with this kind of approach, the value of the American option in the node $u \cdot S_0$ would be equal to $\max(\tilde{f}_u, f_u^{(e)})$.

- Later, we will **prove** (*) that the above approach is not only logical and sensible but actually optimal in a rigorous financial-mathematical sense. Similarly, we would exercise at the node $d \cdot S_0$ if and only if $\tilde{f}_d > f_d^{(e)}$ and the value of the American option at the node $d \cdot S_0$ is equal to $\max(\tilde{f}_d, f_d^{(e)})$.
- So, $f_u^{(a)} = \max(\tilde{f}_u, f_u^{(e)})$ and $f_d^{(a)} = \max(\tilde{f}_d, f_d^{(e)})$.
- Now let us turn to our objective, the computation of $f_0^{(a)}$.

At time 0, we again have two possibilities:

We can exercise the American option and receive a payoff of \tilde{f}_0 . Or we don't exercise, in which case we keep the option at least until dt . At time dt , the option then has either the (already calculated) value $f_u^{(a)}$ (if the underlying moves up to $u \cdot S_0$) or the (already calculated) value $f_d^{(a)}$ (if the underlying moves down to $d \cdot S_0$). So, if we do not exercise, then we keep a product that in the following one-step model will give us a "payoff" of $f_u^{(a)}$ or of $f_d^{(a)}$ after a time dt , depending on how the underlying price moves. But we already know the value of such a product in a one-step model. It is $e^{-r \cdot dt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})$.

Of the two possibilities, it would be logical to choose the one that gives us the greater value, i.e.:

If $\tilde{f}_0 > e^{-r \cdot dt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})$, we exercise,
 if $\tilde{f}_0 < e^{-r \cdot dt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})$, we don't exercise.

And in this case, with this kind of approach, the value of the American option in the node S_0 would be equal to $\max(\tilde{f}_0, e^{-r \cdot dt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)}))$.

- Later, we will **prove** (**) that the above approach (which is not exactly comparable to the earlier one at time dt) is also not only logical and sensible but actually optimal in a rigorous financial-mathematical sense.
- Thus:

$$f_0^{(a)} = \max\left(\tilde{f}_0, e^{-r \cdot dt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})\right).$$

It is clear how this algorithm can be generalized to arbitrary N-step models. We define the procedure recursively: Starting with the end time $T = N \cdot dt$, we compute the fair values of the American option step by step for the earlier times and state the exercise decisions until we have worked our way back to time 0:

Algorithm

- $f^{(a)}_{u^i d^{N-i}} := \tilde{f}_{u^i d^{N-i}}$ for $i = 0, 1, \dots, N$
- Let $f^{(a)}_{u^i d^{n-i}}$ be already defined for $i = 0, 1, \dots, n$ and for $n = N, N - 1, \dots, M + 1$; then

$$f^{(a)}_{u^i d^{M-i}} := \max \left(\tilde{f}_{u^i d^{M-i}}, e^{-rdt} \cdot \left(p' \cdot f^{(a)}_{u^{i+1} d^{M-i}} + (1 - p') \cdot f^{(a)}_{u^i d^{M+1-i}} \right) \right)$$

for $i = 0, 1, \dots, M$, and then in $u^i d^{M-i} \cdot S_0$, we exercise if and only if

$$\tilde{f}_{u^i d^{M-i}} > e^{rdt} \cdot \left(p' \cdot f^{(a)}_{u^{i+1} d^{M-i}} + (1 - p') \cdot f^{(a)}_{u^i d^{M+1-i}} \right).$$

- Here, $p' = \frac{e^{rdt} - d}{u - d}$ for underlying assets with no payments and costs, and $p' = \frac{e^{(r-q)dt} - d}{u - d}$ for underlying assets with continuous payment at interest rate q .

Theorem 2.10 *The above algorithm yields the fair value of an American option in each node of the N -step binomial model as well as the optimal exercise decision in each node.*

Proof We give the proof explicitly for the two-step binomial model. For the general N -step model, the proof (with the obvious adaptations) is completely analogous.

After the above preliminary observations, we now only have to provide the two partial proofs (*) and (**) mentioned above. \square

ad (*):

The proof consists of two parts.

First part:

Suppose that $\tilde{f}_u > f_u^{(e)}$ and we do not exercise.

We are now going to show that in this case, there would be a strategy *with* exercise that would *surely* (!) give a better result than the strategy without exercise. This means: Exercising is certainly better in this case.

If we do not exercise, then we have an option that we can either sell now or hold until T . A definitely better approach would be to exercise and receive \tilde{f}_u and at the same time buy a replicating portfolio for the payoff values \tilde{f}_{uu} and \tilde{f}_{ud} at time T . This replicating portfolio costs $f_u^{(e)}$. We get to keep $\tilde{f}_u - f_u^{(e)} > 0$ in cash. The replicating portfolio has the exact same characteristics as the option (that we would hold if we do not exercise). However, given the additional positive cash amount, we are now in a better position (regardless of how the option (= the replicating portfolio) performs from then on). So this strategy *with* exercise is definitely better.

Second part:

Suppose that $\tilde{f}_u < f_u^{(e)}$ and we exercise.

The most obvious counterargument would be: It would definitely be better to sell the option at the price $f_u^{(e)}$, since by doing so, we will definitely make more than \tilde{f}_u ,

which is what we would get if we exercised. However, we need to be careful here, since it is possible, in principle, that the option has the fair price $f_u^{(e)}$ but cannot be traded momentarily at that price (for whatever reason). So, to be on the safe side, we will phrase the argument as follows:

Definitely better than exercising the option is to keep it and sell the replicating portfolio for the option. This definitely gives us $f_u^{(e)}$ (which is greater than \tilde{f}_u). We then just keep the option and the shorted replicating portfolio until time T , as the two cancel each other out at time T anyway.

ad (**):

In principle, the proof in this case is practically analogous to the above; we only need the argument to be slightly more precise in detail.

The proof again consists of two parts.

First part:

Suppose that $\tilde{f}_0 > e^{-rdt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})$ and we don't exercise.

We are now going to show that in this case, there would be a strategy *with* exercise that would *surely* (!) give a better result than the strategy without exercise. This means: Exercising is certainly better in this case.

If we do not exercise, then we have an option that we can either sell now or by time dt or hold until time $T = 2 \cdot dt$. A definitely better approach would be the following: We exercise and receive \tilde{f}_0 , and at the same time, we buy a replicating portfolio for the payoff values $f_u^{(a)}$ and $f_d^{(a)}$ at time dt .

This replicating portfolio costs $e^{-r \cdot dt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})$.

We get to keep $\tilde{f}_0 - e^{-rdt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)}) > 0$ in cash!

Until time dt , the replicating portfolio has the exact same characteristics as the option (which we would hold if we do not exercise).

In particular, at time dt , the replicating portfolio has a value of either $f_u^{(a)}$ or $f_d^{(a)}$ (depending on where the underlying is going), and these two values are at least as large as the potential payoffs \tilde{f}_u and \tilde{f}_d (whichever applies).

We can exercise the option (which we would still hold at time dt if we do not exercise) at time dt or keep it until time T . If we exercise at time dt , we receive \tilde{f}_u or \tilde{f}_d . However, selling the replicating portfolio at time dt yields $f_u^{(a)}$ or $f_d^{(a)}$, thus at least the same amount.

If we do not exercise the option (which we would still hold at time dt if we do not exercise) at time dt and are, for example, in $u \cdot S_0$ at time dt (the case " $d \cdot S_0$ " applies analogously), then we sell the replicating portfolio we have held so far at time dt and receive $f_u^{(a)}$ in return. We use that to buy the replicating portfolio for the values \tilde{f}_{uu} and \tilde{f}_{ud} at time T . According to the

(continued)

definition of $f_u^{(a)}$, the cost of this replicating portfolio is $f_u^{(a)}$ at most. At time T , this replicating portfolio will certainly have the same value as the option held until expiration. However, given the additional positive cash amount that we made at the beginning, we are now in a better position (regardless of how the option (= the replicating portfolio) moves from then on). So this strategy *with exercise* is definitely better.

Second part:

Suppose that $\tilde{f}_0 < e^{-rdt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})$ and we exercise.

Definitely better than exercising the option is to keep it and sell the replicating portfolio for the payoff values $f_u^{(a)}$ and $f_d^{(a)}$ at time dt . This definitely gives us $e^{-rdt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)})$ (which is more than \tilde{f}_0).

We keep the option until time T and keep the shorted replicating portfolio until time dt . At time dt , we close the replicating portfolio (cost of $f_u^{(a)}$ or $f_d^{(a)}$) and go short on the replicating portfolio for \tilde{f}_{uu} and \tilde{f}_{ud} at time T (if we are in $u \cdot S_0$ at time dt) or for \tilde{f}_{ud} and \tilde{f}_{dd} at time T (if we are in $d \cdot S_0$ at time dt) (payoff $f_u^{(a)}$ or $f_d^{(a)}$). At time T , the option and the replicating portfolio cancel each other out. So we are definitely better off than if we had exercised. \square

To ensure full understanding of the pricing process, we are going to run a numerical example in the next subsection.

A **Bermudan option** is a hybrid between a European and an American option. Early exercise of a Bermudan option is possible at certain predefined—but not all—times. The approach for pricing a Bermudan option is of course analogous to the above algorithm. The decision as to whether or not an option should be exercised (and any adjustment to the option price in the respective node as a consequence of that decision) only takes place at the times at which the option can actually be exercised.

2.5 Valuation Examples for American Options in the Binomial and in the Wiener Model

Example 2.11 We are looking at an American option in a two-step binomial model (the underlying asset's price path is shown in black) the payoffs of which are plotted in red in Fig. 2.5. So the parameters are:

$$\begin{aligned}S_0 &= 4, \\u &= 1.5, \\d &= 0.5,\end{aligned}$$

and thus $u \cdot S_0 = 6$, $d \cdot S_0 = 2$, $u^2 \cdot S_0 = 9$, $u \cdot d \cdot S_0 = 3$, and $d^2 \cdot S_0 = 1$

Furthermore, $r = 0$ and thus $e^{rdt} = 1$ and $p' = \frac{1-d}{u-d} = 0.5$.

The payoffs are given by

$$f_{uu}^{(a)} = \tilde{f}_{uu} = 100, f_{ud}^{(a)} = \tilde{f}_{ud} = 50, f_{dd}^{(a)} = \tilde{f}_{dd} = 20,$$

$$\tilde{f}_u = 70, \tilde{f}_d = 40, \text{ and } \tilde{f}_0 = 50.$$

We now compute

$$f_u^{(e)} = e^{-rdt} \cdot (p' \cdot \tilde{f}_{uu} + (1 - p') \cdot \tilde{f}_{ud}) = 0.5 \times 100 + 0.5 \times 50 = 75$$

$$f_d^{(e)} = e^{-rdt} \cdot (p' \cdot \tilde{f}_{ud} + (1 - p') \cdot \tilde{f}_{dd}) = 0.5 \times 50 + 0.5 \times 20 = 35$$

The comparison $\tilde{f}_u = 70 < f_u^{(e)} = 75$ shows that the option is **not exercised in $u \cdot S_0$** and that therefore $f_u^{(a)} = f_u^{(e)} = 75$.

The comparison $\tilde{f}_d = 40 > f_d^{(e)} = 35$ shows that the option is **exercised in $d \cdot S_0$** and that therefore $f_d^{(a)} = 40$.

We now compute

$$e^{-rdt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)}) = 0.5 \times 75 + 0.5 \times 40 = 57.5.$$

The comparison $\tilde{f}_0 = 50 < 57.5$ shows that the option is **not exercised in S_0** and that therefore $f_0^{(a)} = 57.5$.

The nodes in which the American option is exercised are shown in red in Fig. 2.5.

The price of the corresponding European option is given by

$$f_0^{(e)} = e^{-rdt} \cdot (p' \cdot f_u^{(e)} + (1 - p') \cdot f_d^{(e)}) = 0.5 \times 75 + 0.5 \times 35 = 55.$$

The binomial model can of course again be used to approximate the Wiener model to arbitrary accuracy and thus to value American options in the Wiener model to arbitrary accuracy.

The corresponding program can be found on our website (see <https://app.lsqf.org/binomial-model/derivative-valuation>).

You only need to enter the relevant parameters S_0 , r and σ for the underlying and the expiration T as well as the payoff function $\Phi(t, S_t)$ as a function of the time and the value of the underlying. In addition, you can select the number N of steps for the binomial model with which the Wiener model is to be approximated.

The program then automatically computes the appropriate parameters u , d , and p' of the binomial model, performs backwardation, outputs the fair value of the American option (in comparison to the value of the corresponding European option), and graphically illustrates where the American option should be exercised.

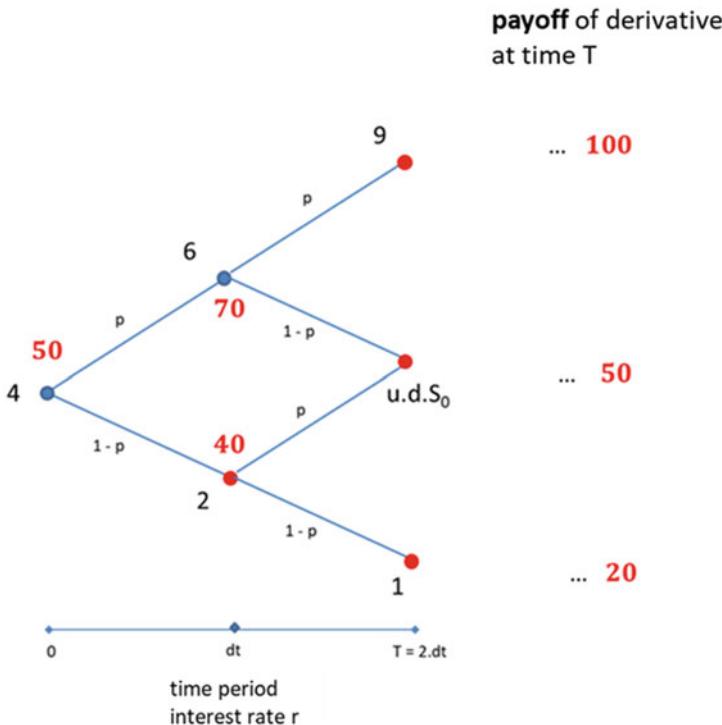


Fig. 2.5 Example of an American option in a two-step binomial model

This of course takes us to the question as to how well these approximations actually approximate the correct fair values of the option. For European options, we have the exact Black-Scholes price as the benchmark. In the above example of Fig. 2.6, this exact price is 231.765. For the American option, we do not have such an exact benchmark. However, the following Table 2.1 and the following graph (Fig. 2.7) provide some information about the quality of the approximation.

In Table 2.1, the values of the European option and the American option are those obtained from valuations in the respective N -step model. These values are also entered in the graph of Fig. 2.7. We can see a very good approximation to the exact value in the European case and that the value in the American case stabilizes at approximately 236.3. But even the values for $N = 25$ are already good approximations to the values that will emerge later. Moreover, we are looking at an option with a relatively long time to expiration and relatively high volatility.

It is also interesting to note how the approximation behaves when N is small. For illustration, we created the same graph as in Fig. 2.7 for the range $N = 1, 2, \dots, 20$ (see Fig. 2.8).

Again, it is clear how **Bermudan options** can be approximately priced in the Wiener model through pricing in the binomial model. You approximate the Wiener

Valuation of a derivative in a binomial N-step-model

Parameters of the derivative

Underlying Value 2,620	volatility (%) 26
risk-free interest rate (%) 2.1	

Derivative

Type put	maturity today 10.12.2018
Strike 2,600	10.12.2019

european american

Parameters of binomial N-step model

Number of steps 250

Valuation

Fair value of derivative 231.735

Fig. 2.6 Pricing American options in the LSQF software

Table 2.1 Values of European option and American option obtained from the respective N-step model

N	25	50	75	100	125	150	175	200	225	250	275	300
European	232.8	232.9	231.0	232.4	231.7	232.1	231.9	231.9	232.0	231.7	231.9	231.8
American	237.8	237.5	235.9	237.0	236.5	236.7	236.5	236.5	236.5	236.3	236.5	236.2

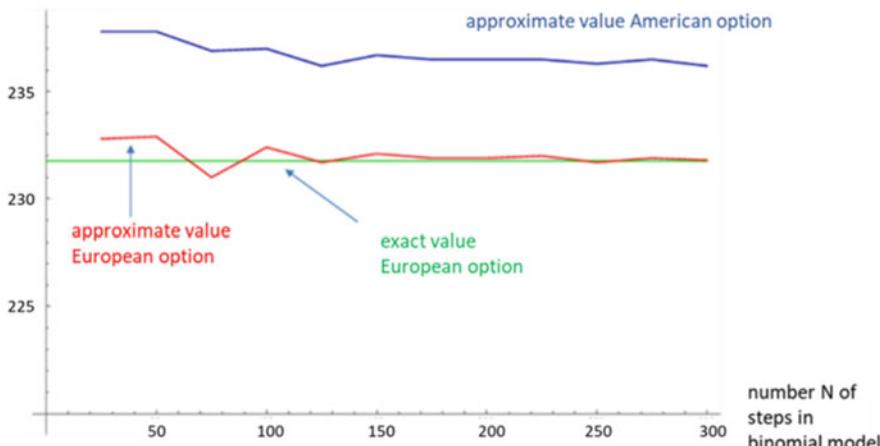


Fig. 2.7 Convergence of approximate values when priced in the N-step binomial model, European (red) and American (blue)

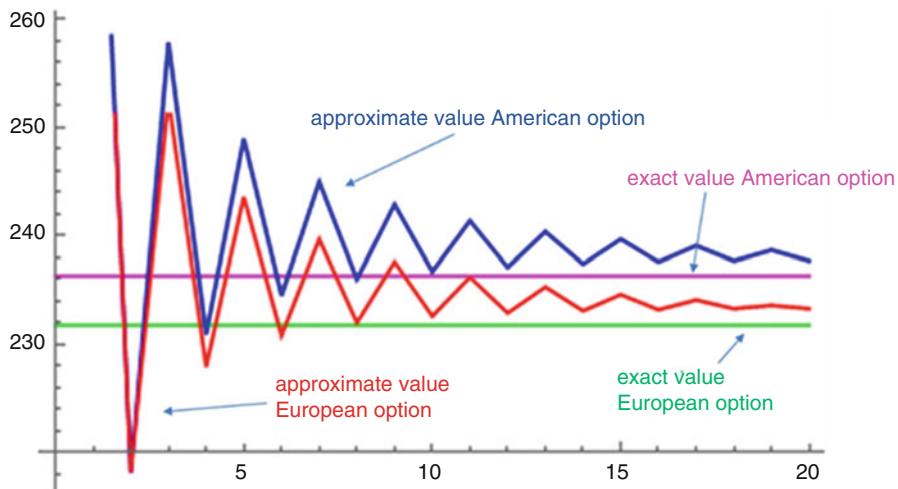


Fig. 2.8 Convergence of approximate values when priced in the N -step binomial model, European (red) and American (blue) $N = 1, 2, \dots, 20$

model through an N -step binomial model. In this N -step model, you allow early exercise only at the times closest to the times at which an exercise is possible in the Wiener model.

2.6 Hedging American Options

An American option can be hedged in basically the same way as a European option. The only difference is that in the American case, the replicating strategy needs to be executed in a dynamical step-by-step manner based on the option's (interim) values.

Incidentally, hedging opens up additional opportunities in the case of American options, specifically if the holder of the American option, contrary to theory, chooses to not exercise the option at certain points in time when early exercise would be possible. We will illustrate this below using the two-step model from the previous section as an example.

If we have an underlying asset that is to be modelled in a Wiener model, then we must first choose a binomial model to approximate the Wiener model and to perform hedging in the binomial model.

Hedging the American option in the binomial model is, as has already been said, analogous to hedging a European option. For illustration, the procedure is carried out explicitly using the above example. The details of this example are given again below.

$$\begin{aligned} S_0 &= 4, \\ u &= 1.5, \\ d &= 0.5, \end{aligned}$$

and thus $u \cdot S_0 = 6$, $d \cdot S_0 = 2$, $u^2 \cdot S_0 = 9$, $u \cdot d \cdot S_0 = 3$, and $d^2 \cdot S_0 = 1$

$$r = 0$$

$$p' = 0.5.$$

The payoffs are given by

$f_{uu}^{(a)} = \tilde{f}_{uu} = 100$, $f_{ud}^{(a)} = \tilde{f}_{ud} = 50$, $f_{dd}^{(a)} = \tilde{f}_{dd} = 20$, $\tilde{f}_u = 70$, $\tilde{f}_d = 40$, and $\tilde{f}_0 = 50$.

$$f_u^{(e)} = e^{-rdt} \cdot (p' \cdot \tilde{f}_{uu} + (1 - p') \cdot \tilde{f}_{ud}) = 75$$

$$f_d^{(e)} = e^{-rdt} \cdot (p' \cdot \tilde{f}_{ud} + (1 - p') \cdot \tilde{f}_{dd}) = 35$$

In $u \cdot S_0$, the option is not exercised and thus $f_u^{(a)} = f_u^{(e)} = 75$.

In $d \cdot S_0$, the option is exercised and thus $f_d^{(a)} = 40$.

$$e^{-rdt} \cdot (p' \cdot f_u^{(a)} + (1 - p') \cdot f_d^{(a)}) = 57.5.$$

In S_0 , the option is not exercised and thus $f_0^{(a)} = 57.5$.

$$f_0^{(e)} = e^{-rdt} \cdot (p' \cdot f_u^{(e)} + (1 - p') \cdot f_d^{(e)}) = 55.$$

The situation at the outset is illustrated in Fig. 2.5.

- We are now going to assume that we sold the American option at the fair price of $f_0^{(a)} = 57.5$.
- The buyer of the option will of course not exercise it immediately, since the achievable payoff is only 55 (i.e. lower than the price that was just paid for the option).
- The collected amount of 57.5 has been calculated so that we have the minimum we need to open a replicating portfolio that, at time dt , attains the values $f_u^{(a)} = 75$ (if the underlying rises to $u \cdot S_0$) or $f_d^{(a)} = 40$ (if the underlying falls to $d \cdot S_0$).
- If the underlying moves up to $u \cdot S_0$ by time dt ,
then the value of our replicating portfolio will be 75.

If the option holder acts contrary to expectations (and contrary to what theory dictates) and chooses to exercise, we liquidate the replicating portfolio at 75 and can thus afford to pay the payoff of $\tilde{f}_u = 70$. We pocket a profit of 5.

If the option holder acts as expected (and in line with theory) and does not exercise, we also liquidate the replicating portfolio and receive proceeds of 75. We use this amount to buy a replicating portfolio for the payoffs $\tilde{f}_{uu} = 100$ and $\tilde{f}_{ud} = 50$ at time T . At time T , the option and the replicating portfolio then cancel each other out.

- If the underlying drops to $d \cdot S_0$ by time dt ,
then the value of our replicating portfolio will be 40.

If the option holder acts as expected (and in line with theory) and chooses to exercise, we sell the replicating portfolio at 40 and use that amount to pay the payoff to the option holder.

If the option holder acts contrary to expectations (and contrary to what theory dictates) and does not exercise, we liquidate the replicating portfolio and collect 40. At the same time, we buy a replicating portfolio for the payoffs $\tilde{f}_{ud} = 50$ and $\tilde{f}_{dd} = 20$ at time T . The cost of this is only 35.

At time T , the option and the replicating portfolio cancel each other out, and we pocket a profit of 5.

Hedging in any N-step model is done analogously.

Our software offers the possibility to perform and test hedging of American options.

2.7 Path-Dependent (Exotic) Derivatives, Definition and Examples

So far, we have dealt exclusively with derivatives whose payoff function Φ was a function of the underlying price at time T . Examples included the payoff function of a call option $\Phi(S_T) = \max(S_T - K, 0)$ or a put option $\Phi(S_T) = \max(K - S_T, 0)$.

In the case of **path-dependent (or exotic) options**, the payoff depends not only on the price of the underlying asset at expiration T but on its entire trajectory or at least parts of its trajectory.

The payoff is therefore a function $\Phi((S_t)_{t \in [0, T]})$.

Path-dependent derivatives can be both European- and American-style.

In the following, we present some examples of path-dependent options used in real trades:

Asian Calls and Puts

In the case of Asian options, the payoff depends on an average of the underlying asset's price at certain points in time $0 \leq t_1 < t_2 \dots < t_M \leq T$, i.e. from $\frac{1}{M} \sum_{i=1}^M S(t_i)$ (arithmetic mean) or from $\sqrt[M]{\prod_{i=1}^M S(t_i)}$ (geometric mean). For example, $S(t_i)$; $i = 1, 2, \dots, M$ can be the underlying asset's daily closing prices over the option's life.

An arithmetic Asian call option would then have the payoff

$$\Phi(S(t_1), S(t_2), \dots, S(t_M)) = \max \left(0, \frac{1}{M} \sum_{i=1}^M S(t_i) - K \right)$$

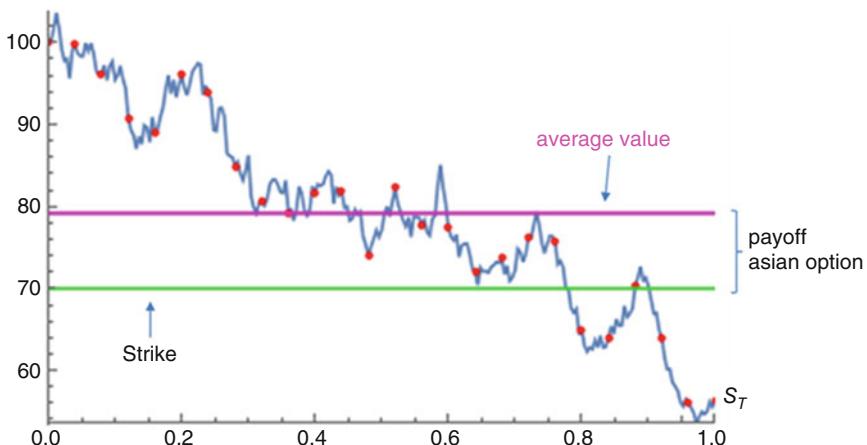


Fig. 2.9 Illustration of how an Asian call option works

and an arithmetic Asian put option would have the payoff

$$\Phi(S(t_1), S(t_2), \dots, S(t_M)) = \max\left(0, K - \frac{1}{M} \sum_{i=1}^M S(t_i)\right)$$

Examples of such—often (exchange-) traded—options are, for example, the “TAPOs” traded on the LME (London Metal Exchange) (see <https://www.lme.com/en-GB/Trading/Contract-types/TAPOs#tabIndex=0>).

These are Asian options, for example, on the aluminium price or on the copper price.

Figure 2.9 illustrates how an Asian call option works. The red dots mark the prices, measured at equidistant points in time, that are used for computing the average. The magenta line shows the average value, and the green line shows the call option’s strike price. Although the price S_T at expiration is lower than the strike, the Asian call does generate a payoff because the average price is higher than the strike.

If the option illustrated in Fig. 2.9 were a put option instead of a call, then the averaging method would have the exact opposite effect: Although the price S_T at expiration is lower than the strike, the Asian put does not generate a payoff because the computed average is higher than the strike.

Lookback Options

The payoff in the case of a lookback option always depends on the underlying asset’s minimum or maximum price over the option’s life (or over a certain part of the option’s life).

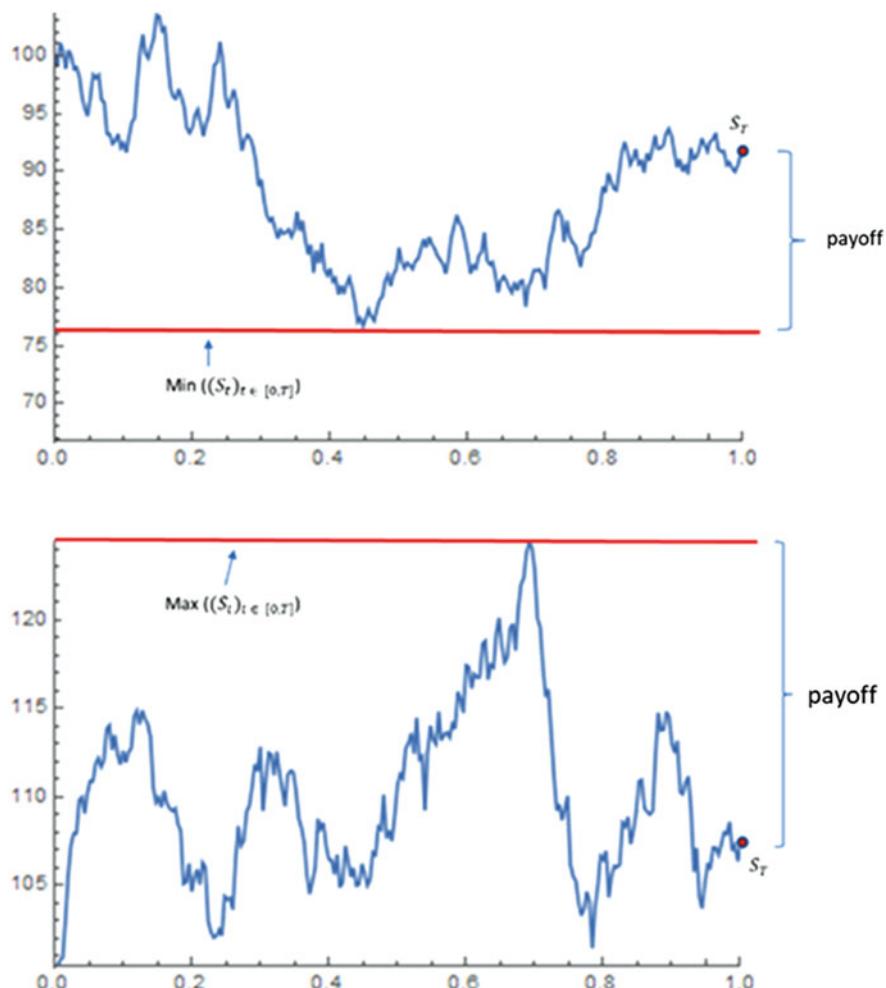


Fig. 2.10 Lookback options with payoffs $S_T - \min\{(S_t)_{t \in [0,T]}\}$ and $\max\{(S_t)_{t \in [0,T]}\} - S_T$

The payoff function of a lookback option can be of various types. The most essential versions are the following payoff functions:

$$\Phi((S_t)_{t \in [0,T]}) = S_T - \min\{(S_t)_{t \in [0,T]}\}$$

$$\Phi((S_t)_{t \in [0,T]}) = \max\{(S_t)_{t \in [0,T]}\} - S_T$$

(Figure 2.10 illustrates how these two types work.)

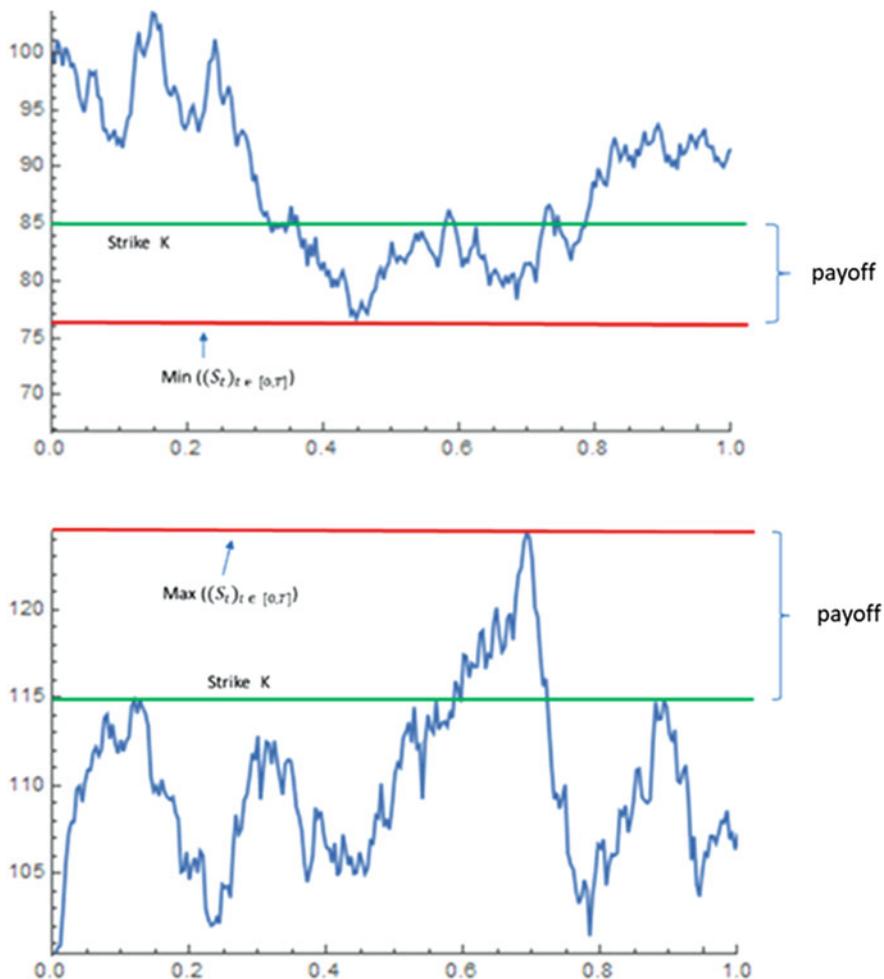


Fig. 2.11 Lookback options with payoffs $K - \min((S_t)_{t \in [0, T]})$ and $\max((S_t)_{t \in [0, T]}) - K$

Or

$$\Phi((S_t)_{t \in [0, T]}) = \max((S_t)_{t \in [0, T]}) - K$$

$$\Phi((S_t)_{t \in [0, T]}) = K - \min((S_t)_{t \in [0, T]})$$

where K is a fixed predefined strike price. (Figure 2.11 illustrates how these two types work.)

Barrier Options

Barrier options normally have a payoff function that principally depends only on the underlying asset's price S_T at expiration. However, these options become invalid as soon as the price of the underlying once exceeds or falls below a certain barrier L during the option's life, or they become valid only if the price once exceeds or falls below a certain barrier L during the option's life.

Standard types of such barrier options are, for example:

- Up-and-out barrier puts
- Down-and-out barrier calls
- Up-and-in barrier calls
- Down-and-in barrier puts

The payoff of an **up-and-out barrier put** with strike K and barrier L is

$$\Phi((S_t)_{t \in [0, T]}) = \begin{cases} \max(0, K - S_T) & \text{if } S_t < L \text{ for all } t \in [0, T] \\ 0 & \text{if } S_t > L \text{ for one } t \in [0, T] \end{cases}$$

Figure 2.12 illustrates how this type of option works:

As we can see, this is an ordinary put option, which however expires immediately if and as soon as the price exceeds the barrier at any one point over its life.

The payoff of a **down-and-out barrier call** with strike K and barrier L is

$$\Phi((S_t)_{t \in [0, T]}) = \begin{cases} \max(0, S_T - K) & \text{if } S_t > L \text{ for all } t \in [0, T] \\ 0 & \text{if } S_t < L \text{ for one } t \in [0, T] \end{cases}$$

Figure 2.13 illustrates how this type of option works:

As we can see, this is an ordinary call option, which however expires immediately if and as soon as the price falls below the barrier at any one point over its life.

The payoff of an **up-and-in barrier call** with strike K and barrier L is

$$\Phi((S_t)_{t \in [0, T]}) = \begin{cases} \max(0, S_T - K) & \text{if } S_t > L \text{ for at least one } t \in [0, T] \\ 0 & \text{if } S_t < L \text{ for all } t \in [0, T] \end{cases}$$

Figure 2.14 illustrates how this type of option works:

As we can see, this is an ordinary call option, which however becomes valid only if and as soon as the price exceeds the barrier at any one point over its life.

The payoff of a **down-and-in barrier put** with strike K and barrier L is

$$\Phi((S_t)_{t \in [0, T]}) = \begin{cases} \max(0, K - S_T) & \text{if } S_t < L \text{ for at least one } t \in [0, T] \\ 0 & \text{if } S_t > L \text{ for all } t \in [0, T] \end{cases}$$

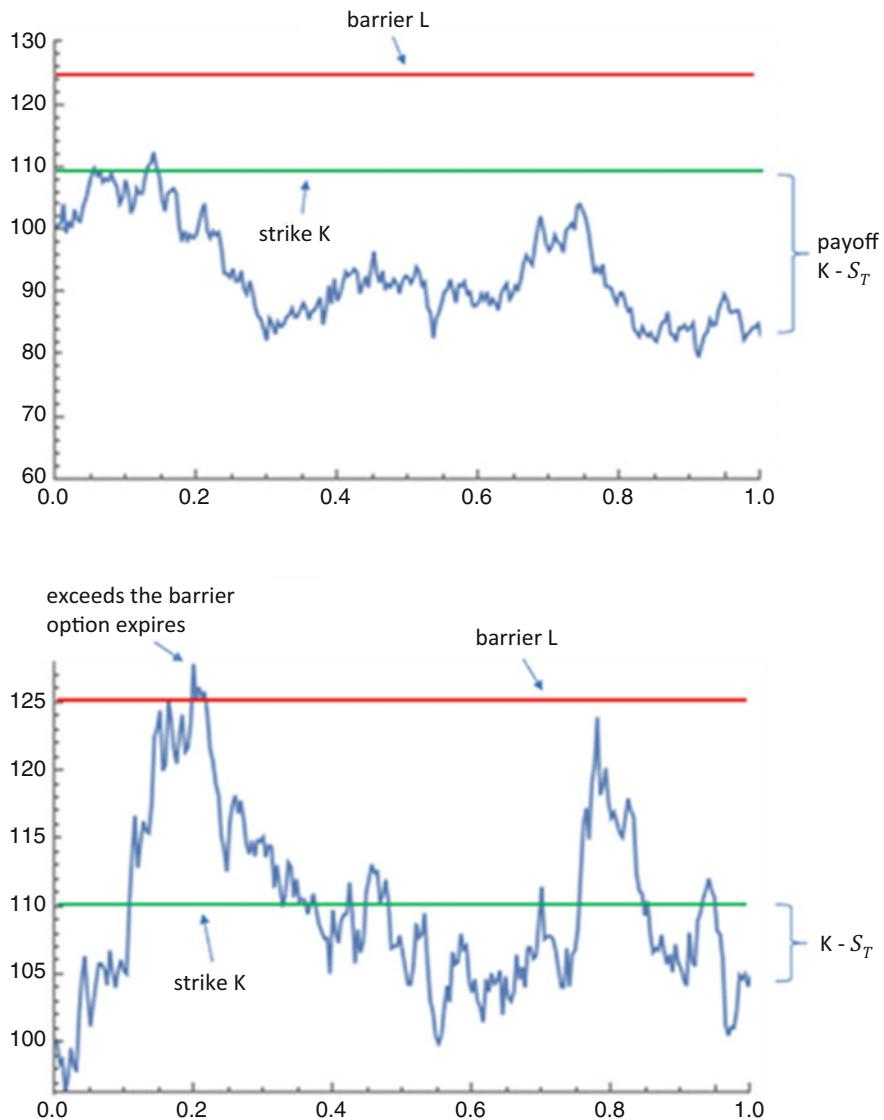


Fig. 2.12 Illustration of how an up-and-out barrier put option works

Figure 2.15 illustrates how this type of option works:

As we can see, this is an ordinary put option, which however becomes valid only if and as soon as the price falls below the barrier at any one point over its life.

As stated earlier, the above are just a few examples from a wide range of path-dependent options traded on the financial markets.

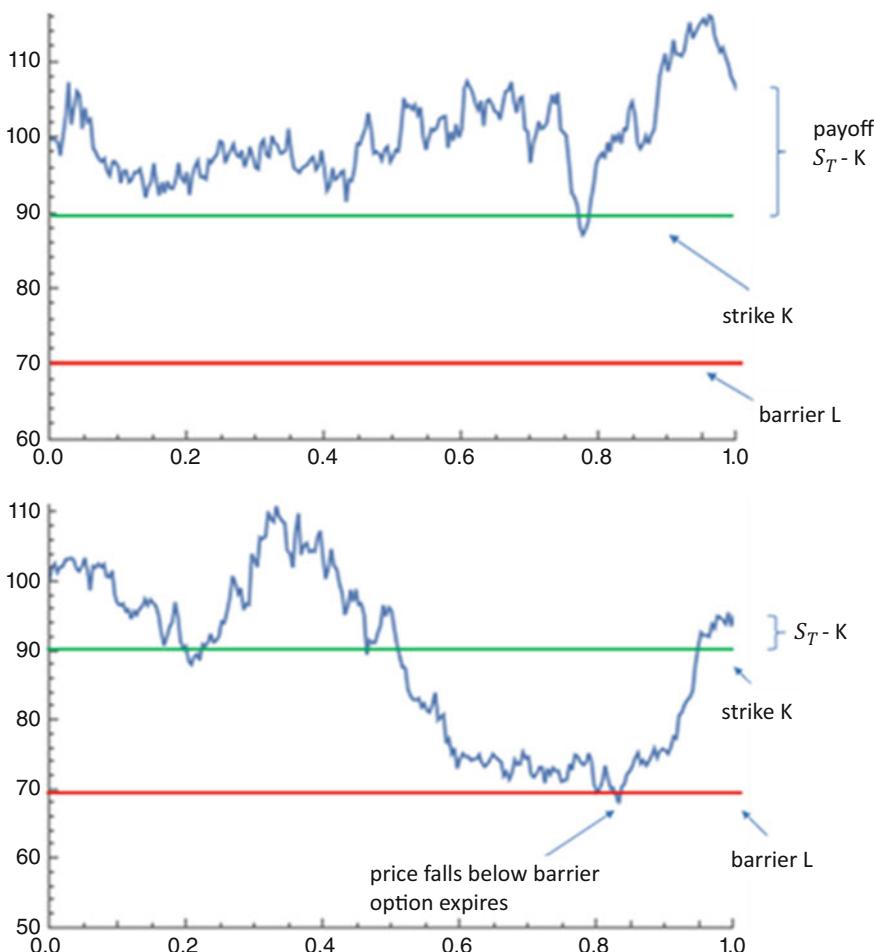


Fig. 2.13 Illustration of how a down-and-out barrier call option works

2.8 Valuation of Path-Dependent Options, the Black-Scholes Formula for Path-Dependent Options

For the fair price of path-dependent options on an underlying asset that follows a Wiener model, we use a counterpart to the Black-Scholes formula, in the obvious form. We are now going to formulate this Black-Scholes formula for path-dependent options and will then plausibilize it by pricing path-dependent options in the binomial model. An exact proof of the formula is beyond the scope of this monograph.

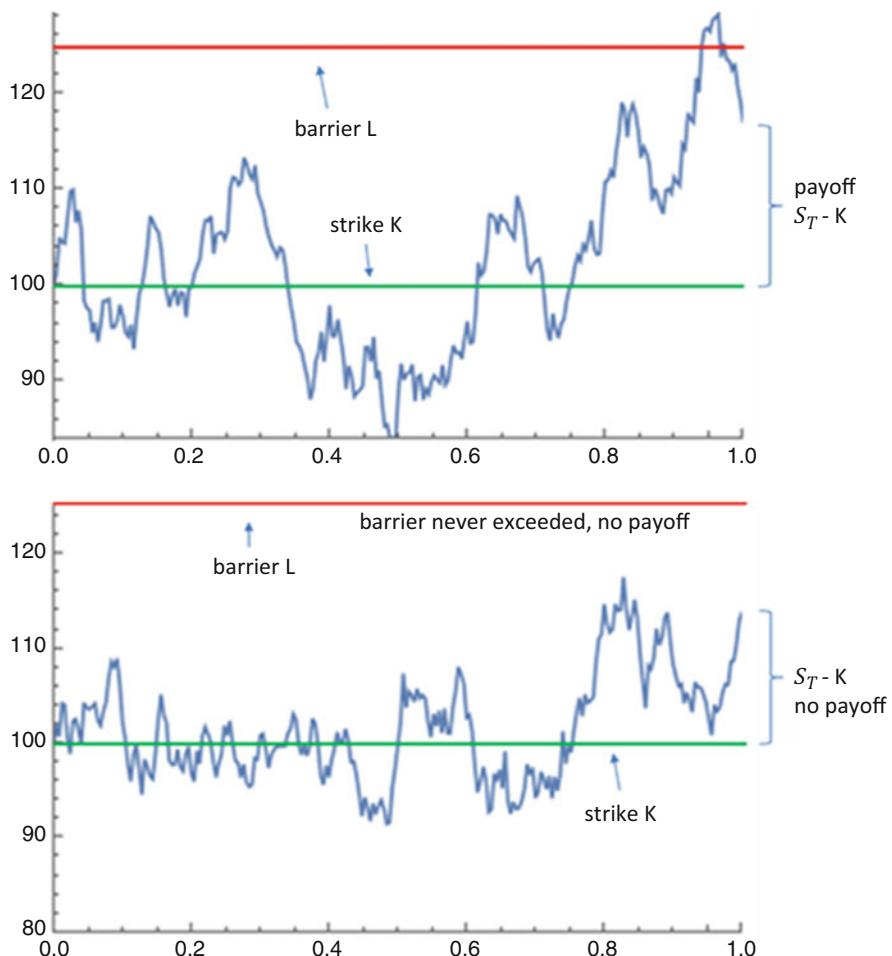


Fig. 2.14 Illustration of how an up-and-in barrier call option works

Theorem 2.12 (Black-Scholes Formula for Path-Dependent Options) Let D be a European, path-dependent derivative with expiration T and payoff function Φ on an underlying asset with price $S(t)$ that follows a Wiener model with parameters μ and σ in the time range $[0, T]$. (It is assumed that no payments or costs are incurred through the underlying asset.) The fair price $F(0)$ of D at time 0 is then given by

$$F(0) = e^{-rT} \cdot E \left(\Phi \left((\tilde{S}_t)_{t \in [0, T]} \right) \right)$$

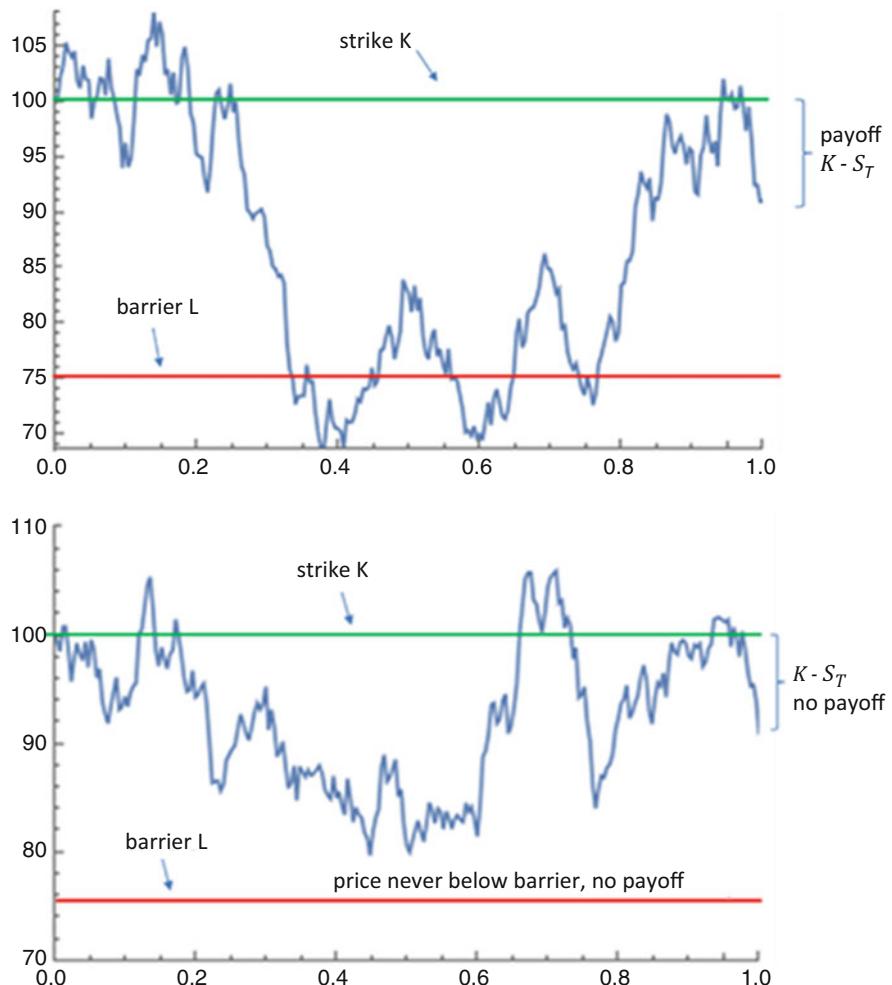


Fig. 2.15 Illustration of how a down-and-in barrier put option works

where the price movement of \tilde{S} over time is

$$\tilde{S}(t) = S(0) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)t + \sigma \sqrt{t} w}$$

for each $t \in [0, T]$ with a standard normally distributed random variable w .

“E” in this equation denotes the expected value, and r is the risk-free interest rate $f_{0,T}$.

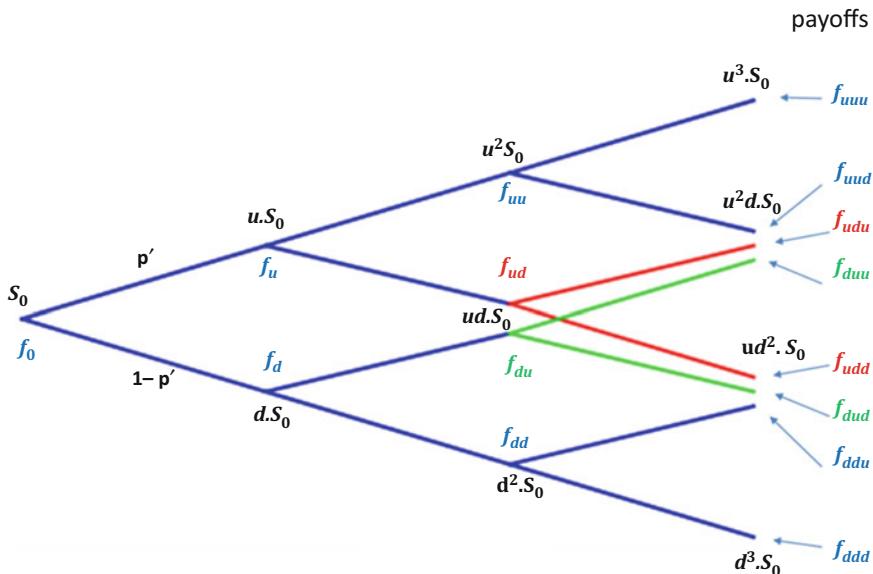


Fig. 2.16 Path-dependent option in a three-step binomial model

How to calculate that expected value $E\left(\Phi\left(\left(\tilde{S}_t\right)_{t \in [0, T]}\right)\right)$, which in this case does not only depend on the end price of the underlying asset but on its entire performance over time, will be discussed in the next chapter.

Here, we want to plausibilize the result, and to that end, we will again proceed by first considering how to price a path-dependent option in a binomial three-step model. The main difference in the representation of the model in the case of a path-dependent option is that the option's payoffs f at time T (or the option's fair prices f at earlier times) no longer depend only on the value of the underlying at the respective point in time, but rather on how (by which path) this value is reached.

In the above example (see Fig. 2.16), the values $u^2d \cdot S_0$ and $ud^2 \cdot S_0$ can each be reached via three different paths (we denote them as uud , udu and duu , and udd , dud , and ddu , respectively). This means that three different payoffs f_{uud} , f_{udu} , or f_{duu} (or f_{udd} , f_{dud} , or f_{ddu}) can occur at each of the two points.

Furthermore, in the above example, the value $ud \cdot S_0$ at time $2 \cdot dt$ can be reached via two paths. Accordingly, we will price the option with f_{ud} if $ud \cdot S_0$ was reached via $u \cdot S_0$ (i.e. via the path ud), and we will price it with f_{du} if $ud \cdot S_0$ was reached via $d \cdot S_0$ (i.e. via the path du).

How the option is priced at time $2 \cdot dt$ (depending on the path that the underlying takes to get there) naturally follows from our earlier results in the one-step binomial

model:

$$f_{uu} = e^{-rdt} \cdot (p' \cdot f_{uuu} + (1 - p') \cdot f_{uud})$$

$$f_{ud} = e^{-rdt} \cdot (p' \cdot f_{udu} + (1 - p') \cdot f_{udd})$$

$$f_{du} = e^{-rdt} \cdot (p' \cdot f_{duu} + (1 - p') \cdot f_{dud})$$

$$f_{dd} = e^{-rdt} \cdot (p' \cdot f_{ddu} + (1 - p') \cdot f_{ddd})$$

Likewise, it is also clear (from our observations in the one-step binomial model) how the option is priced at time dt :

$$f_u = e^{-rdt} \cdot (p' \cdot f_{uu} + (1 - p') \cdot f_{ud})$$

$$f_d = e^{-rdt} \cdot (p' \cdot f_{du} + (1 - p') \cdot f_{dd})$$

And finally we get

$$f_0 = e^{-rdt} \cdot (p' \cdot f_u + (1 - p') \cdot f_d)$$

If we now plug into the last formula for f_0 the formulas for f_u and for f_d from the last-but-one formula block and then the formulas from the first formula block for f_{uu} , f_{ud} , f_{du} , and f_{dd} , we get

$$\begin{aligned} f_0 = & e^{-rdt} \cdot (p' \cdot p' \cdot p' \cdot f_{uuu} + p' \cdot p' \cdot (1 - p') \cdot f_{uud} + p' \cdot (1 - p') \cdot p' \cdot f_{udu} + \\ & + (1 - p') \cdot p' \cdot p' \cdot f_{duu} + p' \cdot (1 - p') \cdot (1 - p') \cdot f_{udd} + \\ & + (1 - p') \cdot p' \cdot (1 - p') \cdot f_{dud} + (1 - p') \cdot (1 - p') \cdot p' \cdot f_{ddu} + \\ & + (1 - p') \cdot (1 - p') \cdot (1 - p') \cdot f_{ddd}) \end{aligned}$$

Taking a closer look at the formula for f_0 , we see that we can also write the formula as follows:

$$\begin{aligned} f_0 = & e^{-rdt} \cdot (W(uuu) \cdot \Phi(uuu) + W(uud) \cdot \Phi(uud) + W(udu) \cdot \Phi(udu) + \\ & + W(duu) \cdot \Phi(duu) + W(udd) \cdot \Phi(udd) + W(dud) \cdot \Phi(dud) + \\ & + W(ddu) \cdot \Phi(ddu) + W(ddd) \cdot \Phi(ddd)) \end{aligned}$$

Here, $W(uuu), \dots$ denotes the artificial (!) probability (i.e. with respect to p') for the price to take the path uuu, \dots , and $\Phi(uuu), \dots$ denotes the payoff if it takes the path uuu, \dots . We therefore sum over each possible path that the underlying price can take from time 0 to time T , determine the payoff for each of these paths, and weight it by the (artificial) probability of its occurrence.

So, f_0 is again nothing other than the discounted expected payoff under the risk-neutral measure, although in this case, of course, the value of the payoff depends on the actual path taken. Thus, in the three-step binomial model:

$$f_0 = e^{-r dt} \cdot E \left(\Phi \left((\tilde{S}_t)_{t \in [0, T]} \right) \right)$$

where \tilde{S}_t denotes the underlying asset's price path under the risk-neutral measure.

It is relatively obvious (and easily proved by induction) that this also holds in any N-step binomial model.

For this reason, it should be plausible that the corresponding formula—just as in the case of non-path-dependent plain vanilla options—remains valid when passing to the Wiener model and that Theorem 2.12, the Black-Scholes formula for path-dependent options, therefore also holds true in the Wiener model.

Similarly, it is plausible (and easily proved, just as in the non-path-dependent case) that in the **binomial model**, it is also possible to price **American path-dependent** options by **backwardation**. In this case, it is again necessary that, in addition to the usual valuation process in each node, the instantaneous payoff is compared with the interim value established in the subsequent one-step model (European-mode).

In the following section, we are going to run a valuation example of a path-dependent option in the three-step binomial model, in both European- and American-style.

2.9 Numerical Valuation Example of a Path-Dependent Option in a Three-Step Binomial Model (European and American)

For illustration, we are going to price the arithmetic Asian put option with strike $K = 15$ in the binomial three-step model of Fig. 2.17. The black values represent the price movements of the underlying asset; the possible payoffs for each path up to each of these values are plotted in blue.

The development of the underlying asset is given by the values $S_0 = 8$, $u = \frac{3}{2}$, and $d = \frac{1}{2}$. In addition, we choose $r = 0$, resulting in $p' = \frac{1}{2}$.

The blue payoff values result in each case from

$$\max(0, K - \text{average value of the underlying in the path so far})$$

This value is always equal to “15 – average value of the underlying in the path so far” except at time T at the final value of 27. There, the payoff is 0.

For example, the payoff at time $2 \cdot dt$ at the underlying value of 6, if reached on the path with value 12 at time dt , is given by

$$\max(0, 15 - \text{average value of the underlying in the path so far}).$$

The average value of the underlying in the path so far is $\frac{(8+12+6)}{3} = \frac{26}{3}$.

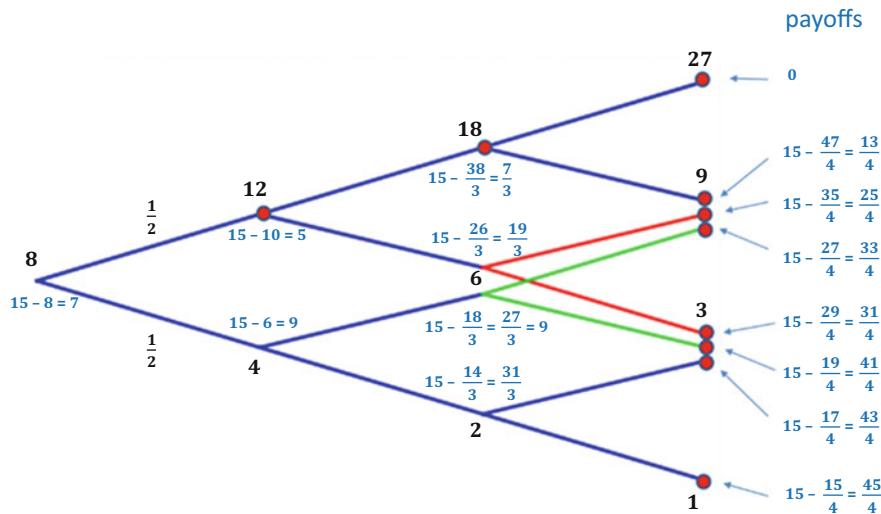


Fig. 2.17 Example of Asian put option (nodes at which the American version is exercised are plotted in red)

The payoff is therefore equal to $\max\left(0, 15 - \frac{26}{3}\right) = \max\left(0, \frac{19}{3}\right) = \frac{19}{3}$. We are going to price the European version of the option first and then the American version.

For the values $f^{(e)}$ of the European option, we get

$$f_{uu}^{(e)} = \frac{1}{2} \left(0 + \frac{13}{4} \right) = \frac{13}{8} = 1.625$$

$$f_{ud}^{(e)} = \frac{1}{2} \left(\frac{25}{4} + \frac{31}{4} \right) = 7$$

$$f_{du}^{(e)} = \frac{1}{2} \left(\frac{33}{4} + \frac{41}{4} \right) = \frac{37}{4} = 9.25$$

$$f_{dd}^{(e)} = \frac{1}{2} \left(\frac{43}{4} + \frac{45}{4} \right) = 11$$

$$f_u^{(e)} = \frac{1}{2} \left(\frac{13}{8} + 7 \right) = \frac{69}{16} = 4.3125$$

$$f_d^{(e)} = \frac{1}{2} \left(\frac{37}{4} + 11 \right) = \frac{81}{8} = 10.125$$

$$f_0^{(e)} = \frac{1}{2} \left(\frac{69}{16} + \frac{81}{8} \right) = \frac{231}{32} = \mathbf{7.21875}$$

For the values $f^{(a)}$ of the American option, we get

$$f_{uu}^{(a)} = \max\left(\frac{7}{3}, f_{uu}^{(e)}\right) = \max\left(\frac{7}{3}, \frac{13}{8}\right) = \frac{7}{3} = 2.333$$

in node “uu”, the option is **exercised**

$$f_{ud}^{(a)} = \max\left(\frac{19}{3}, f_{ud}^{(e)}\right) = \max\left(\frac{19}{3}, 7\right) = 7$$

in node “ud”, the option is **not exercised**

$$f_{du}^{(a)} = \max\left(\frac{7}{3}, f_{du}^{(e)}\right) = \max\left(9, \frac{37}{4}\right) = \frac{37}{4} = 9.25$$

in node “du”, the option is **not exercised**

$$f_{dd}^{(a)} = \max\left(\frac{31}{3}, f_{dd}^{(e)}\right) = \max\left(\frac{31}{3}, 11\right) = 11$$

in node “dd”, the option is **not exercised**

$$f_u^{(a)} = \max\left(5, \frac{1}{2}(f_{uu}^{(a)} + f_{ud}^{(a)})\right) = \max\left(5, \frac{1}{2}\left(\frac{7}{3} + 7\right)\right) = \max\left(5, \frac{14}{3}\right) = 5$$

in node “u”, the option is **exercised**

$$f_d^{(a)} = \max\left(9, \frac{1}{2}(f_{du}^{(a)} + f_{dd}^{(a)})\right) = \max\left(9, \frac{1}{2}\left(\frac{37}{4} + 11\right)\right) = \max\left(9, \frac{81}{8}\right) = \frac{81}{8} = 10.125$$

in node “d”, the option is **not exercised**

$$f_0^{(a)} = \max\left(7, \frac{1}{2}(f_u^{(a)} + f_d^{(a)})\right) = \max\left(7, \frac{1}{2}\left(5 + \frac{81}{8}\right)\right) = \max\left(7, \frac{121}{16}\right) = \frac{121}{16} = 7.5625$$

in node “0”, the option is **not exercised**

Just for comparison: As can easily be checked by computation, the value of a conventional plain vanilla put option in this model is 8.5 for both the European and the American type.

2.10 The Complexity of Pricing Path-Dependent Options in an N-Step Binomial Model in General and, For Example, for Lookback Options

We have used the binomial model in the context of path-dependent option pricing for two purposes: first, to plausibilize the result of Theorem 2.12 and second, to approximate the fair value of path-dependent options (of both European- and American-style) for an underlying over a Wiener model.

There are only a few special cases (such as barrier options) where an explicit and exact computation of the expected value in Theorem 2.12 is possible.

In one of the following chapters, we will show how a simple Monte Carlo method can be used to approximate the expected value in Theorem 2.12 and thus the fair value of a European path-dependent option. The Monte Carlo method in this basic version cannot be used for pricing American path-dependent options.

One could argue, of course, that the American path-dependent options can be approximated—to arbitrary accuracy—by means of the N-step binomial model, as

exemplified above. That is correct in theory, but will often fail in reality due to the numerical complexity of the task.

As we have seen above, to fully price an American version of an option in an N-step binomial model, it is necessary to treat each path to each node separately. There are a total of 2 paths of length 1, 2^2 paths of length 2, 2^3 paths of length 3, etc. $\dots, 2^N$ paths of length N . So, in total, we are dealing with $2 + 2^2 + 2^3 + \dots + 2^N = 2^{N+1} - 2$ paths.

This means that even in a binomial model with just 30 steps (which is the minimum required to obtain a reasonably reliable result—especially for more complex products and for longer-dated options), we already have 2 billion paths to handle. This might just be feasible, but definitely takes far too long for an effective pricing tool. Thus, in the case of American path-dependent options, we often need to use other more subtle numerical methods to determine the expected value in Theorem 2.12.

In some cases, however, it is possible to price path-dependent (European and American) derivatives in the N-step binomial model even for large N . How that can be done, and why, and under what conditions, is what we are going to discuss using a specific example, namely, American lookback options.

The payoff of an American lookback option depends at any time t on the minimum or maximum value that the underlying asset has attained up to that point and possibly on the current underlying price S_t .

In the following, we are going to examine the example of an American lookback option with payoff $S_t - \min((S_u)_{u \in [0, t]})$ at each point in time $t \in [0, T]$.

We will now price this option in an N-step binomial model. For this purpose, we proceed according to the following algorithm (the procedure described below will be carried out and illustrated for a specific numerical example at a later point):

Algorithm

Step 1 (Determine Potential Minima)

In each node Z of the N-step binomial model, we note which minima may have occurred along the paths to that node. Determination of potential minima can be done for each node Z based simply on the values in the two preceding nodes X and Y that lead directly to Z in one step. How to calculate the values at the node Z is shown in the following graph (see Fig. 2.18).

We use the following notations:

z is the value of the underlying asset at node Z

v_1, v_2, \dots, v_x are the potential minima in X that are smaller than z

$v_{x+1}, v_{x+2}, \dots, v_\kappa$ are the potential minima in X that are greater than z

w_1, w_2, \dots, w_y are the potential minima in Y that are smaller than z

$w_{y+1}, w_{y+2}, \dots, w_\eta$ are the possible minima in Y that are greater than z

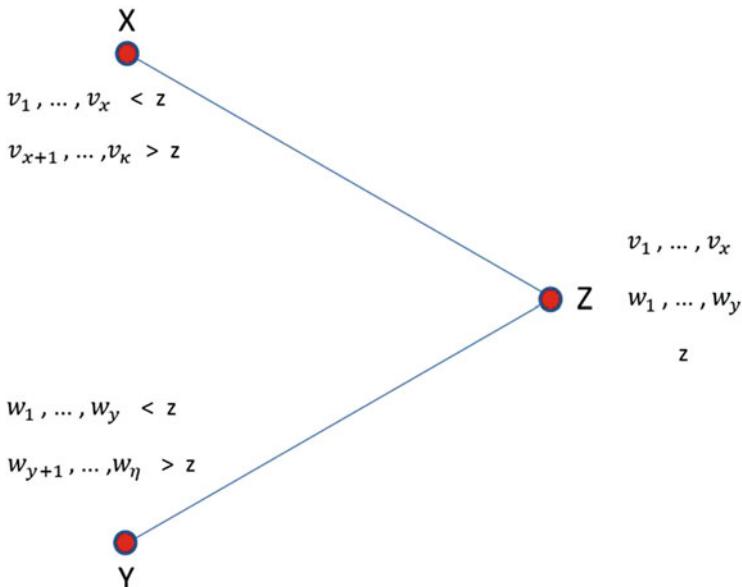


Fig. 2.18 Step 1 for efficiently pricing lookback options in the binomial model (determination of potential minima)

So obviously the only potential minima up to node Z are the values v_1, v_2, \dots, v_x , w_1, w_2, \dots, w_y and possibly the value z .

Important Note

The number of possible minima on paths to a node Z at time $n \cdot dt$ is of course limited by the number of different values that the underlying can possibly attain on the path from 0 to Z . This number is at most as large as there are different values for the underlying up to time $(n - 1) \cdot dt$ (plus 1 for the value of the underlying in Z itself).

So at the most, this number is equal to $1 + 2 + 3 + \dots + (n - 1) + 1 = \frac{n \cdot (n - 1)}{2} + 1 \leq \frac{n^2}{2}$.

Thinking this through, it is relatively easy to understand that the number of possible minima (for sufficiently large n) must actually be substantially smaller than $\frac{n^2}{4}$, and real applications show that in most cases, the number is indeed even much smaller than that.

Step 2 (Split up the Graph)

In the next step, we construct a new graph. In this graph, the previous nodes X , Y , Z , etc. are split into several nodes, that is, into exactly as many nodes as there are possible minima up to that node. We denote the new nodes by $(X, v_1), (X, v_2), \dots, (X, v_k)$ and $(Y, w_1), (Y, w_2), \dots, (Y, w_\eta)$ and $(Z, v_1), (Z, v_2), \dots, (Z, v_x), (Z, w_1), (Z, w_2), \dots, (Z, w_y)$ as well as (Z, z) , if z is a potential minimum.

In each such new node denoted by, say, (Q, τ) , the payoff of the lookback option is uniquely given by $q - \tau$ (current value q of the underlying minus the minimum τ on the respective path to Q).

The edges in the new graph have been chosen as follows (based again on the situation shown in Fig. 2.18)

$$(X, v_1) \rightarrow (Z, v_1)$$

$$(X, v_2) \rightarrow (Z, v_2)$$

...

$$(X, v_x) \rightarrow (Z, v_x)$$

$$(Y, w_1) \rightarrow (Z, w_1)$$

$$(Y, w_2) \rightarrow (Z, w_2)$$

...

$$(Y, w_y) \rightarrow (Z, w_y)$$

The nodes $(X, v_{x+1}), (X, v_{x+2}), \dots, (X, v_K)$ and $(Y, w_{y+1}), (Y, w_{y+2}), \dots, (Y, w_\eta)$ are all connected with (Z, z) .

The new situation for the section of the graph from Fig. 2.18 is shown in Fig. 2.19.

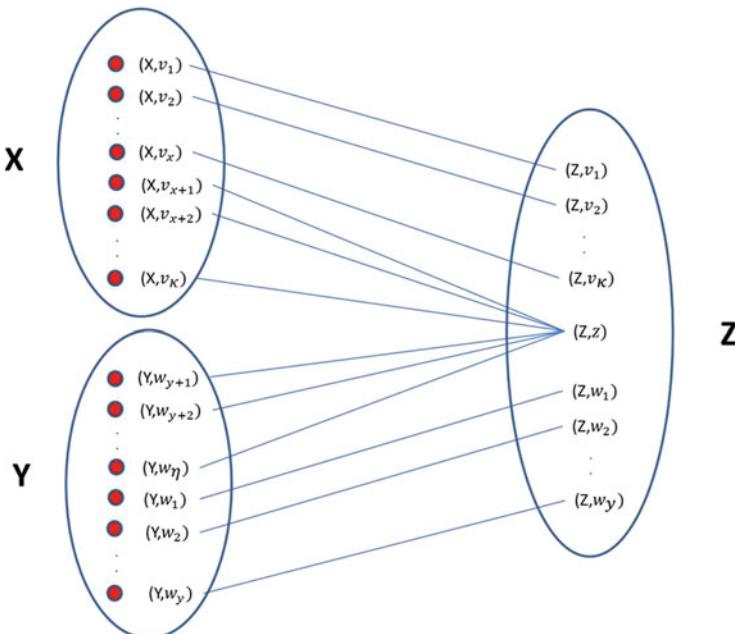


Fig. 2.19 Step 2 for efficiently pricing lookback options in the binomial model (splitting the graph)

Step 3 (Backwardation)

The value of the option at expiration $T = N \cdot dt$ is, of course, again equal to the payoff at the respective end point. We are now looking at the extended split graph. The end points are all pairs (U, m) consisting of a possible end value U of the underlying and a possible minimum m reached on the way from 0 to U .

The payoff and hence the value of the lookback option at the point (U, m) are given by $u - m$.

Thus, we know the value of the option in all possible situations at time $N \cdot dt$. In the same way as before, we now successively calculate (in the split-up graph) the values of the option in all possible situations at time $(N - 1) \cdot dt$, then at time $(N - 2) \cdot dt, \dots$ and finally at time 0 by backwardation.

The backwardation process is explained again below. The situation and notations are those used in Fig. 2.20.

In Fig. 2.20, we want to determine the fair value $f^{(a)}(X, v_i)$ of the lookback option at time $(n - 1) \cdot dt$ and at the node (X, v_i) of the extended graph.

For this purpose, we assume that the previous backwardation steps have already yielded the lookback option's fair values in all nodes for the time points $n \cdot dt$ and later.

We denote the nodes that can be reached from the node (X, v_i) in the next time step by (U, l_j) and by (Z, m_k) . The fair values of the lookback option at these nodes have already been calculated (based on the above assumption), and we denote them by $f^{(a)}(U, l_j)$ and by $f^{(a)}(Z, m_k)$.

The payoff if the option is immediately exercised at the node (X, v_i) is given by $x - v_i$.

The value of the lookback option at time $(n - 1) \cdot dt$ at the node (X, v_i) is then computed—based on the same arguments we applied earlier in deriving the backwardation method—using

$$f^{(a)}(X, v_i) = \max(x - v_i, e^{-rdt} \left(p' \cdot f^{(a)}(U, l_j) + (1 - p') \cdot f^{(a)}(Z, m_k) \right))$$

and the option is exercised at the node (X, v_i) if and only if

$$x - v_i > e^{-rdt} \left(p' \cdot f^{(a)}(U, l_j) + (1 - p') \cdot f^{(a)}(Z, m_k) \right)$$

We continue to compute the fair value in this way until time 0, i.e. until the fair value of the lookback option is determined at time 0.

End of Algorithm

Now, what about the complexity of the algorithm? We are not going to answer this question in detail but will give only a very rough upper estimate:

At each point in time $n \cdot dt$, an original node X essentially splits (see the “Important note” above) into maximally $\frac{n^2}{4}$ nodes in the extended graph. So, at

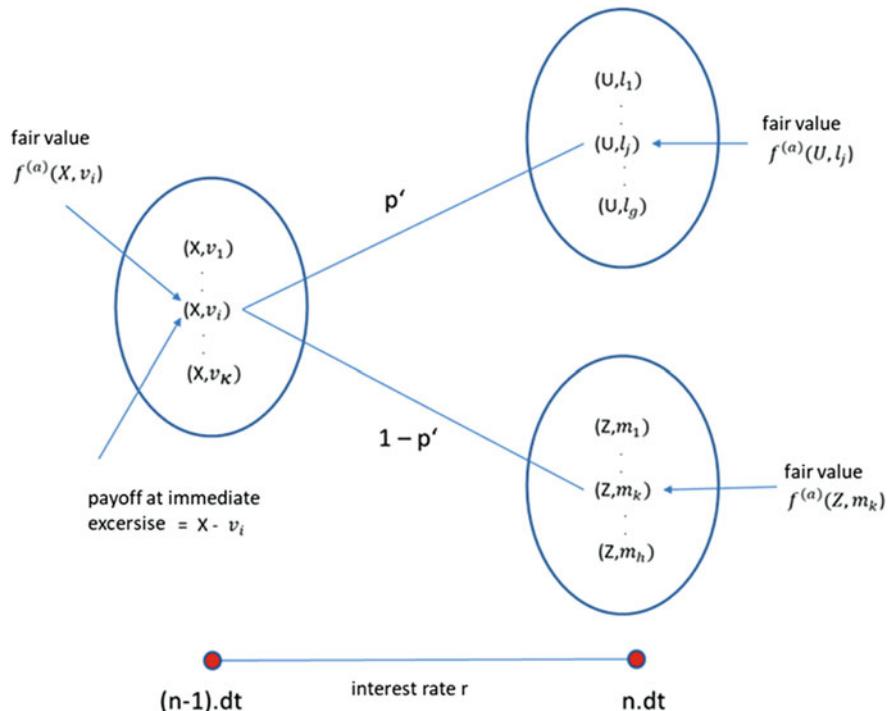


Fig. 2.20 Step 3 for efficiently pricing lookback options in the binomial model (backwardation)

any time $n \cdot dt$ in the extended graph, we are essentially (given sufficiently large n) dealing with $n \cdot \frac{n^2}{4} = \frac{n^3}{4}$ nodes at most.

This means that in total, we essentially need to perform valuations at $\frac{1^3}{4} + \frac{2^3}{4} + \frac{3^3}{4} + \dots + \frac{(N-1)^3}{4} < \frac{N^4}{16}$ nodes at most. Each valuation requires essentially **one** computational step.

Thus, in the case of a 30-step model, we have to perform valuations in a maximum of about 50,000 nodes to arrive at a complete valuation of the option. In the vast majority of cases, however, the number of nodes to be taken into account is substantially (!) lower. For comparison, recall that the traditional approach to pricing a path-dependent option would require us to work with more than 2 billion paths.

Let us analyse briefly which of the lookback option's properties makes it possible to work with much fewer nodes rather than having to consider every single possible path:

It is essential to note that not every single path leading to a node X generates a different instantaneous payoff, but that, in fact, "relatively few" different payoffs are possible in each individual node X ; and another essential aspect is that these

relatively few possible payoffs in a node X are also easily calculated (with little numerical effort).

In such cases, pricing American path-dependent options in a binomial model is numerically feasible and efficient.

For European path-dependent options—unless an explicit formula for the expected value in Theorem 2.12 and thus for the fair value can be derived anyway (as in the case of barrier options)—one would not work with a binomial model, but with other numerical methods such as the Monte Carlo method, which we will get to know in one of the next paragraphs.

For American path-dependent options of a more complex form (e.g. American arithmetic Asian options), one needs to develop and apply much more subtle pricing procedures. We will deal with this in the next volume of this book project.

To illustrate the above—somewhat more complex—valuation method for American lookback options, we are going to price such an option in a four-step binomial model in the following section.

2.11 Valuation of an American Lookback Option in a Four-Step Binomial Model (Numerical Example)

We are now going to value an American lookback option with **payoff** $S_t - \min((S_u)_{u \in [0, t]})$ at each point in time $t \in [0, T]$ within the four-step binomial model shown in Fig. 2.21.

First, we note at each node the potential minima that can be attained up to that node (see Fig. 2.22).

The next step was to split the graph (see Fig. 2.23). In each node of the extended graph in Fig. 2.23, the respective payoff has been entered in green.

Now we can do the backwardation, starting from the payoffs at expiration $T = 4 \cdot dt$ (with $f(x,y)$ denoting the option's respective fair value at the node $(X,\textcolor{red}{y})$)

$$\begin{aligned} f(54, 16) &= \max\left(38, \frac{1}{2}(f(81, 16) + f(27, 16))\right) = \max\left(38, \frac{1}{2}(65 + 11)\right) = \\ \max(38, 38) &= 38 \end{aligned}$$

exercise!

$$\begin{aligned} f(18, 16) &= \max\left(2, \frac{1}{2}(f(27, 16) + f(9, 9))\right) = \max\left(2, \frac{1}{2}(11 + 0)\right) = \\ \max(2, 5.5) &= 5.5 \end{aligned}$$

do not exercise!

$$\begin{aligned} f(18, 12) &= \max\left(6, \frac{1}{2}(f(27, 12) + f(9, 9))\right) = \max\left(6, \frac{1}{2}(15 + 0)\right) = \\ \max(6, 7.5) &= 7.5 \end{aligned}$$

do not exercise!

$$\begin{aligned} f(18, 8) &= \max\left(10, \frac{1}{2}(f(27, 8) + f(9, 8))\right) = \max\left(10, \frac{1}{2}(19 + 1)\right) = \\ \max(10, 10) &= 10 \end{aligned}$$

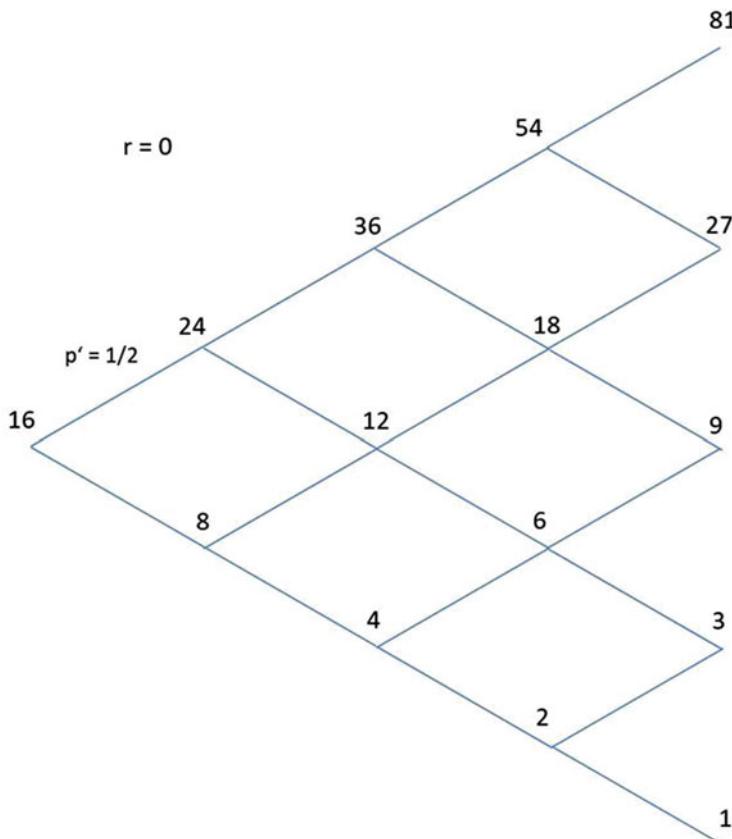


Fig. 2.21 Example: Lookback option in the four-step binomial model

exercise!

$$f(6, 6) = \max \left(0, \frac{1}{2}(f(9, 6) + f(3, 3)) \right) = \max \left(0, \frac{1}{2}(3 + 0) \right) = \max(0, 1.5) = 1.5$$

do not exercise!

$$f(6, 4) = \max \left(2, \frac{1}{2}(f(9, 4) + f(3, 3)) \right) = \max \left(2, \frac{1}{2}(5 + 0) \right) = \max(2, 2.5) = 2.5$$

do not exercise!

$$f(2, 2) = \max \left(0, \frac{1}{2}(f(3, 2) + f(1, 1)) \right) = \max \left(0, \frac{1}{2}(1 + 0) \right) = \max(0, 0.5) = 0.5$$

do not exercise!

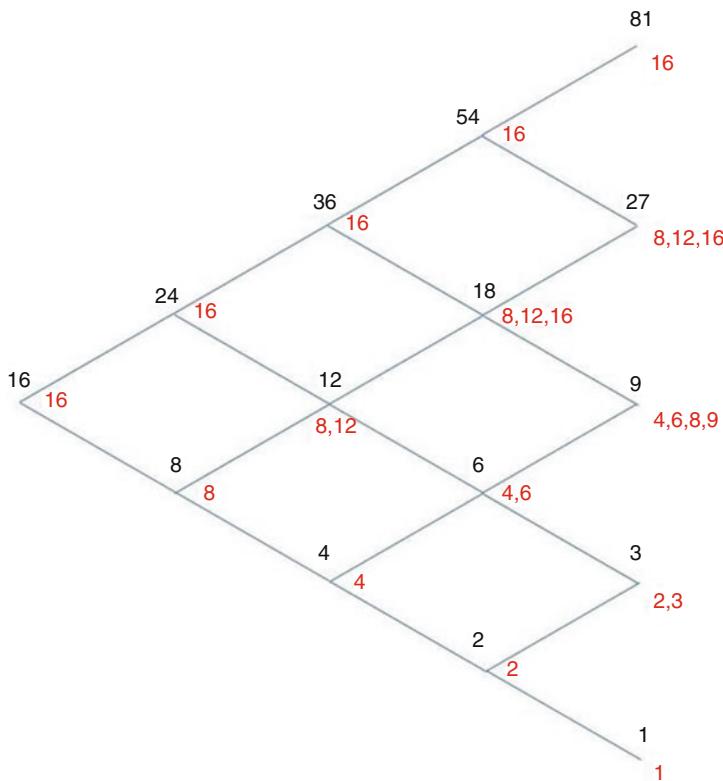


Fig. 2.22 Example: Lookback option in the four-step binomial model, step 1 (determine possible minima)

$$f(36, 16) = \max \left(20, \frac{1}{2}(f(54, 16) + f(18, 16)) \right) = \max \left(20, \frac{1}{2}(38 + 5.5) \right) = \max(20, 21.75) = 21.75$$

do not exercise!

$$f(12, 12) = \max \left(0, \frac{1}{2}(f(18, 12) + f(6, 6)) \right) = \max \left(0, \frac{1}{2}(7.5 + 1.5) \right) = \max(0, 4.5) = 4.5$$

do not exercise!

$$f(12, 8) = \max \left(4, \frac{1}{2}(f(18, 8) + f(6, 6)) \right) = \max \left(4, \frac{1}{2}(10 + 1.5) \right) = \max(4, 5.75) = 5.75$$

do not exercise!

$$f(4, 4) = \max \left(0, \frac{1}{2}(f(6, 4) + f(2, 2)) \right) = \max \left(0, \frac{1}{2}(2.5 + 0.5) \right) = \max(0, 1.5) = 1.5$$

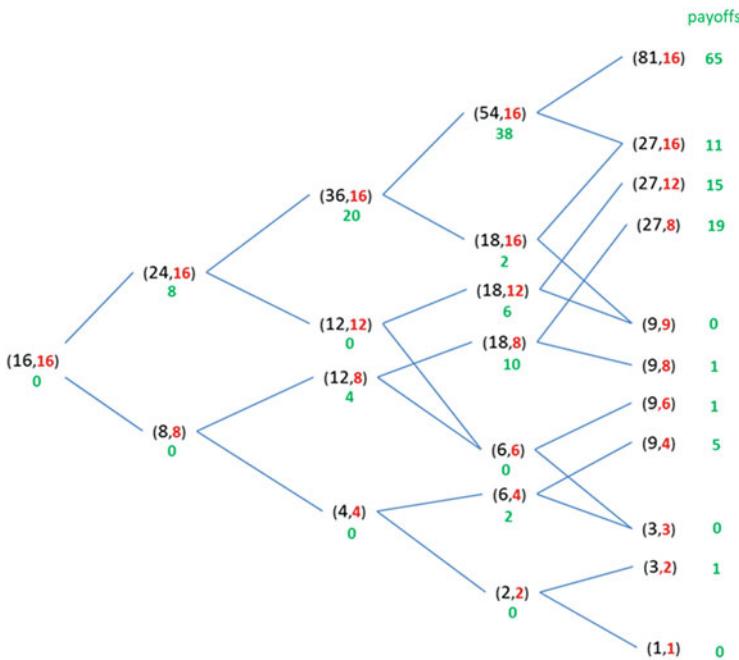


Fig. 2.23 Example: Lookback option in the four-step binomial model, step 2 (split the graph and determine the payoffs (green))

do not exercise!

$$f(24, 16) = \max \left(8, \frac{1}{2}(f(36, 16) + f(12, 12)) \right) = \max \left(8, \frac{1}{2}(21.75 + 4.5) \right) =$$

$$\max(8, 13.125) = 13.125$$

do not exercise!

$$f(8, 8) = \max \left(0, \frac{1}{2}(f(12, 8) + f(4, 4)) \right) = \max \left(0, \frac{1}{2}(5.75 + 1.5) \right) =$$

$$\max(0, 3.625) = 3.625$$

do not exercise!

$$f_0 = f(16, 16) = \max \left(0, \frac{1}{2}(f(24, 16) + f(8, 8)) \right) =$$

$$\max \left(0, \frac{1}{2}(13.125 + 3.625) \right) = \max(0, 8.375) = \mathbf{8.375}$$

do not exercise!

So, the value of the American lookback option at time 0 is 8.375.

At the points where we proposed an early exercise of the option, it would have been equally profitable in each case to keep the option. The value of the European version is therefore also 8.375.

2.12 Explicit Formulas for European Path-Dependent Options, For Example, Barrier Options

As mentioned earlier, it is sometimes possible to give an explicit formula for the fair price of a European path-dependent option on underlying assets over the Wiener model by finding an explicit formula for the expected value in Theorem 2.12, the general Black-Scholes formula for path-dependent options.

This is possible, for example, for certain barrier options (puts, calls, and simpler types of options), plus for the lookback options discussed above (in their basic form) and also, for example, for Asian options whose payoff is based on geometric averaging.

To derive these explicit formulas for barrier options and lookback options, we need a series of deeper insights from probability theory, and delving into these is beyond our scope. We will therefore give the basic results and main pricing formulas for barrier options without proof here and present just some numerical tests.

In the next chapter, however, we will derive and apply the exact pricing formula for “geometric Asian options on underlying assets over the Wiener model” with proof.

Accurate valuations of barrier options hinge on the following four results, which show how the fair price of a down-and-out barrier option or an up-and-out barrier option and of a down-and-in barrier option or an up-and-in barrier option relates to the fair price of options with the same parameters and a similar payoff function but without any barrier condition:

For this purpose, we are going to look at arbitrary European plain vanilla derivatives with payoff function Φ and their corresponding barrier versions.

The expiration date for all derivatives is denoted by T . We denote by t any time points in the interval $[0, T]$.

L always denotes the barrier of the respective barrier versions.

By Φ_L , we denote the following function: $\Phi_L(x) := \begin{cases} \Phi(x) & \text{if } x > L \\ 0 & \text{if } x \leq L \end{cases}$.

Furthermore, let $F(t, s, \Phi_L)$ be the fair price of the European plain vanilla option expiring at T and with payoff function Φ_L at time t and at an underlying price of s at time t .

Recall that we can determine this fair value $F(t, s, \Phi_L)$ using the Black-Scholes formula (to at least approximate arbitrary accuracy).

Our goal now is to find the fair prices $F_{DO}(t, s, \Phi, L)$, $F_{UO}(t, s, \Phi, L)$, $F_{DI}(t, s, \Phi, L)$, and $F_{UI}(t, s, \Phi, L)$ of the down-and-out, up-and-out, down-and-in, and up-and-in barrier options.

We have the following formulas:

$$F_{DO}(t, s, \Phi, L) = F(t, s, \Phi_L) - \left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot F\left(t, \frac{L^2}{s}, \Phi_L\right)$$

for $s > L$ (and 0 otherwise)

$$F_{UO}(t, s, \Phi, L) = F(t, s, \Phi - \Phi_L) - \left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot F\left(t, \frac{L^2}{s}, \Phi - \Phi_L\right)$$

for $s < L$ (and 0 otherwise)

$$F_{DI}(t, s, \Phi, L) = F(t, s, \Phi - \Phi_L) + \left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot F\left(t, \frac{L^2}{s}, \Phi_L\right)$$

$$F_{UI}(t, s, \Phi, L) = F(t, s, \Phi_L) + \left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot F\left(t, \frac{L^2}{s}, \Phi - \Phi_L\right)$$

As a simple example of one of these formulas, we are going to look at a down-and-out barrier call option with a barrier L that is smaller than the strike price K . The situation is shown in Fig. 2.24 (same graph as the one in Fig. 2.15).

In this case where $L < K$, we know that $\Phi_L(x) = \Phi(x) = \max(0, x - K)$ for all x (since $\Phi_L(x) = \Phi(x) = 0$ for all $x < L < K$). Therefore, based on the first of the above formulas:

$$F_{DO}(t, s, \Phi, L) = F(t, s, \Phi) - \left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot F\left(t, \frac{L^2}{s}, \Phi\right)$$

for $s > L$ (and 0 otherwise).

Thus: For the fair price $C_{DO}(t, s, K, L)$ of a down-and-out barrier call option with strike K and barrier L , we have

$$C_{DO}(t, s, K, L) = C(t, s, K) - \left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot C\left(t, \frac{L^2}{s}, K\right)$$

for $s > L$ (and 0 otherwise).

Here, $C(t, x, K)$ denotes the Black-Scholes price of a European plain vanilla call option with strike K at time t and at an underlying price of x .

Substituting the Black-Scholes formula for C would yield an explicit representation for the price of the down-and-out barrier call option.

If we let the barrier L go to 0, the factor $(\frac{L}{s})^{\frac{2r}{\sigma^2}-1}$ goes to 0, and we get $C_{DO}(t, s, K, 0) = C(t, s, K)$.

This is obviously correct, since the price can never fall below the barrier $L = 0$, and the barrier option would then become a completely ordinary (i.e. a plain vanilla) call option.

Immediately clear from the formula $C_{DO}(t, s, K, L) = C(t, s, K) - (\frac{L}{s})^{\frac{2r}{\sigma^2}-1} \cdot C(t, \frac{L^2}{s}, K)$ is also the trivial fact that the price of a down-and-out barrier call option

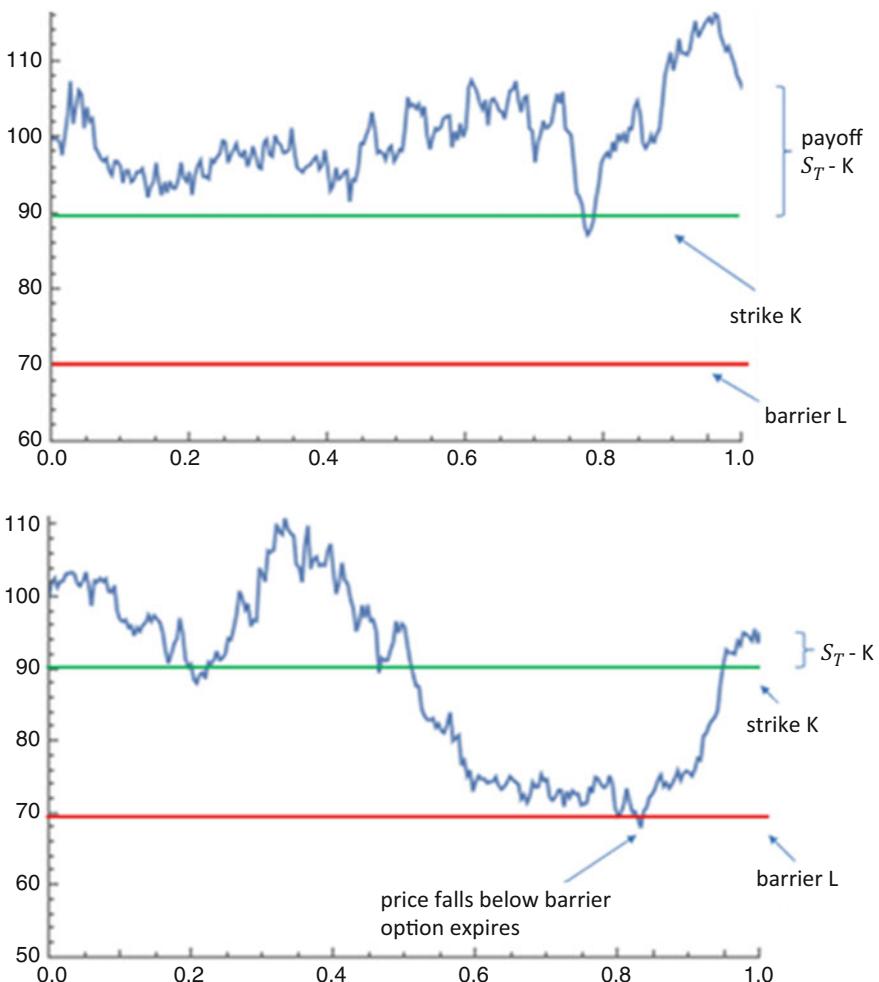


Fig. 2.24 Illustration of how a down-and-out barrier call option works

must always be smaller than the price of the corresponding plain vanilla call option:

$$C_{DO}(t, s, K, L) = C(t, s, K) - \left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot C\left(t, \frac{L^2}{s}, K\right) < C(t, s, K).$$

The difference $\left(\frac{L}{s}\right)^{\frac{2r}{\sigma^2}-1} \cdot C(t, \frac{L^2}{s}, K)$ between the price of the down-and-out barrier call option and the corresponding plain vanilla call option becomes larger the larger L is (because $C(t, x, K)$, as we know, is monotonically increasing in x). This too is a logical consequence (the higher the barrier L , the higher the risk that the barrier option becomes invalid and hence the lower its value).

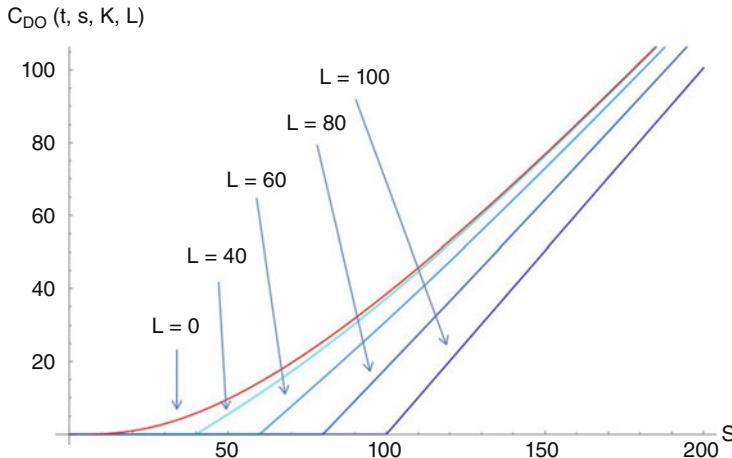


Fig. 2.25 Price movements of down-and-out barrier call options as a function of the underlying asset's price s for different values of the barrier L ($L <$ strike $K = 100$)

In Fig. 2.25, we use a concrete numerical example to illustrate how quickly the value of a down-and-out barrier call option decreases as the barrier L increases.

We choose the following parameters:

- Strike price $K = 100$
- Time to expiration $T = 1$
- Current point in time $t = 0$
- Risk-free interest rate $r = 0.01(1\%)$
- Volatility $\sigma = 1$

Here too, of course, we have software on our website that you can use for pricing barrier options.

2.13 Explicit Formulas for European Path-Dependent Options, For Example, Geometric Asian Options

So, as has been said, for a geometric Asian option, the payoff depends on the geometric mean of the underlying asset's prices at certain points in time $0 \leq t_1 < t_2 \dots < t_M \leq T$, i.e. on $\sqrt[M]{\prod_{i=1}^M S(t_i)}$.

For example, $S(t_i)$; $i = 1, 2, \dots, M$ can be the underlying asset's daily closing prices over the option's life.

A **geometric** Asian call option would then have the payoff

$$\Phi(S(t_1), S(t_2), \dots, S(t_M)) = \max \left(0, \sqrt[M]{\prod_{i=1}^M S(t_i)} - K \right)$$

and an **arithmetic** Asian put option would have the payoff

$$\Phi(S(t_1), S(t_2), \dots, S(t_M)) = \max \left(0, K - \frac{1}{M} \sum_{i=1}^M S(t_i) \right)$$

In Fig. 2.9, we illustrated how an **arithmetic** Asian call option works. In that figure, the red dots marked the prices, measured at equidistant points in time, that were used for computing the average. The magenta line represents the arithmetic average value, and the green line shows the strike price of the call option. Figure 2.26 depicts the same situation, but with an added brown line, which plots the now relevant geometric average value.

In this example, the geometric mean is lower than the arithmetic mean. In fact, this is always the case, not only in this particular example: *The geometric mean of a finite sequence a_1, a_2, \dots, a_M of positive values is always less than or equal to the arithmetic mean of these values. So we always have*

$$\sqrt[M]{\prod_{i=1}^M a_i} \leq \frac{1}{M} \sum_{i=1}^M a_i.$$

Equality between the arithmetic and geometric mean holds only if all values a_1, a_2, \dots, a_M are equal, i.e. if $a_1 = a_2 = \dots = a_M$ holds. (A very brief proof of this fact can be found at the end of this subsection.)

Yet that means the value of a geometric Asian call option is always smaller than that of an arithmetic Asian call option and the value of a geometric Asian put option is always greater than that of an arithmetic Asian put option.

According to Theorem 2.12, the Black-Scholes formula for European path-dependent derivatives, the fair value f_0^{arith} of an arithmetic Asian option or the fair value f_0^{geom} of a geometric Asian option is

$$f_0^{arith} = e^{-rT} \cdot E \left(\max \left(0, \frac{1}{M} \sum_{i=1}^M \tilde{S}(t_i) - K \right) \right)$$

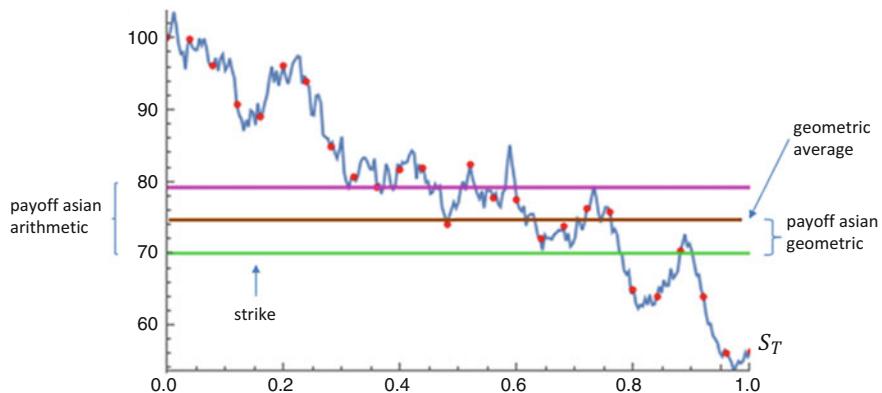


Fig. 2.26 Illustration of how a geometric and an arithmetic Asian call option works

and

$$f_0^{geom} = e^{-rT} \cdot E \left(\max \left(0, \sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} - K \right) \right),$$

where

$$\tilde{S}(t) = S(0) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)t + \sigma \sqrt{t} w} \text{ for all } t \in [0, T].$$

It is not possible to calculate the expected value $E \left(\max \left(0, \frac{1}{M} \sum_{i=1}^M \tilde{S}(t_i) - K \right) \right)$ explicitly in the case of an arithmetic Asian option. To determine this value (and thus the value of an arithmetic Asian option) approximately, we need to resort to numerical methods, such as the Monte Carlo method.

For geometric Asian options, however, it is possible to compute the expected value $E \left(\max \left(0, \sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} - K \right) \right)$ and hence its fair value explicitly, provided the successive measurement points $0 = t_0 < t_1 < t_2 \dots < t_M = T$ are equidistant. We denote that distance by Δt . Hence, $t_{i+1} - t_i = \Delta t$ for all $i = 0, 1, \dots, M - 1$.

We will perform this explicit calculation in the following.

The movement of the risk-neutral stock price \tilde{S} over time can, of course, be represented step by step, as follows:

$$\tilde{S}(t_i) = \tilde{S}(t_{i-1}) \cdot e^{\left(r - \frac{\sigma^2}{2}\right) \cdot \Delta t + \sigma \sqrt{\Delta t} w_{i-1}} \text{ for all } i = 1, 2, \dots, M.$$

Here, the w_i for $i = 0, 1, \dots, M-1$ are independent $\mathcal{N}(0, 1)$ -distributed random variables. If we then apply this representation to $\tilde{S}(t_{i-1})$ and then successively to $\tilde{S}(t_{i-2}), \tilde{S}(t_{i-3}), \dots, \tilde{S}(t_1)$, we obtain for $\tilde{S}(t_i)$

$$\tilde{S}(t_i) = S(0) \cdot e^{i \cdot \left(r - \frac{\sigma^2}{2}\right) \cdot \Delta t + \sigma \sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})}$$

and hence

$$\begin{aligned} \prod_{i=1}^M \tilde{S}(t_i) &= (S(0))^M \cdot \prod_{i=1}^M e^{i \cdot \left(r - \frac{\sigma^2}{2}\right) \cdot \Delta t + \sigma \sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})} = \\ &= (S(0))^M \cdot e^{\sum_{i=1}^M i \cdot \left(r - \frac{\sigma^2}{2}\right) \cdot \Delta t + \sum_{i=1}^M \sigma \sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})} = \\ &= (S(0))^M \cdot e^{\frac{M \cdot (M+1)}{2} \left(r - \frac{\sigma^2}{2}\right) \cdot \Delta t + \sigma \sqrt{\Delta t} \cdot \sum_{i=0}^{M-1} (M-i) \cdot w_i}. \end{aligned}$$

We denote the random variable $\sum_{i=0}^{M-1} (M-i) \cdot w_i$ by \tilde{w} and note the following:

\tilde{w} is the sum of M mutually independent normally distributed random variables $(M-i) \cdot w_i$. \tilde{w} is therefore also normally distributed.

The random variables $(M-i) \cdot w_i$ have expected value 0 and variance $(M-i)^2$.

Therefore, \tilde{w} has expected value 0 and variance $M^2 + (M-1)^2 + (M-2)^2 + \dots + 2^2 + 1^2 = \frac{M \cdot (M+1) \cdot (2M+1)}{6}$, that is, standard deviation $\sqrt{\frac{M \cdot (M+1) \cdot (2M+1)}{6}}$. Here, we have used a well-known summation formula for the sum of the squares of the first M natural numbers.

Thus (using among others $M \cdot \Delta t = T$),

$$\begin{aligned} \sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} &= S(0) \cdot e^{\frac{M+1}{2} \left(r - \frac{\sigma^2}{2}\right) \cdot \Delta t + \sigma \sqrt{\Delta t} \cdot \frac{1}{M} \cdot \tilde{w}} = \\ &= S(0) \cdot e^{\left(\frac{1}{2} + \frac{1}{2M}\right) \cdot \left(r - \frac{\sigma^2}{2}\right) \cdot T + \sigma \sqrt{T} \cdot \frac{1}{M^{1.5}} \cdot \tilde{w}}. \end{aligned}$$

(continued)

The random variable $\frac{1}{M^{1.5}} \tilde{w}$ still has expected value 0 and standard deviation

$$\frac{1}{M^{1.5}} \cdot \sqrt{\frac{M \cdot (M+1) \cdot (2M+1)}{6}} = \sqrt{\frac{(1+\frac{1}{M}) \cdot (2+\frac{1}{M})}{6}}.$$

We therefore write $\frac{1}{M^{1.5}} \tilde{w} = w \cdot \sqrt{\frac{(1+\frac{1}{M}) \cdot (2+\frac{1}{M})}{6}}$, and w is thus an $\mathcal{N}(0, 1)$ -distributed random variable.

We further set $\tilde{\sigma} := \sigma \cdot \sqrt{\frac{(1+\frac{1}{M}) \cdot (2+\frac{1}{M})}{6}}$ and thus obtain

$$\sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} = S(0) \cdot e^{\left(\frac{1}{2} + \frac{1}{2M}\right) \cdot \left(r - \frac{\sigma^2}{2}\right) \cdot T + \tilde{\sigma} \sqrt{T} w}.$$

Finally, we choose a value \tilde{r} , so that $\left(\frac{1}{2} + \frac{1}{2M}\right) \cdot \left(r - \frac{\sigma^2}{2}\right) = \tilde{r} - \frac{\tilde{\sigma}^2}{2}$; thus,

$$\begin{aligned} \tilde{r} &= \left(\frac{1}{2} + \frac{1}{2M}\right) \cdot \left(r - \frac{\sigma^2}{2}\right) + \frac{\tilde{\sigma}^2}{2} = \\ &= \left(\frac{1}{2} + \frac{1}{2M}\right) \cdot \left(r - \frac{\sigma^2}{2}\right) + \sigma^2 \frac{\left(1 + \frac{1}{M}\right) \cdot \left(2 + \frac{1}{M}\right)}{12} = \\ &= r \cdot \left(\frac{1}{2} + \frac{1}{2M}\right) - \sigma^2 \cdot \left(\frac{1}{12} - \frac{1}{12M^2}\right). \end{aligned}$$

Upon which we finally get the following representation:

$\sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} = S(0) \cdot e^{\left(\tilde{r} - \frac{\tilde{\sigma}^2}{2}\right) \cdot T + \tilde{\sigma} \sqrt{T} w}$, and to calculate the fair value of the geometric Asian option, we simply determine

$$\begin{aligned} e^{-rT} \cdot E \left(\max \left(0, \sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} - K \right) \right) &= \\ &= e^{-rT} \cdot E \left(\max \left(0, S(0) \cdot e^{\left(\tilde{r} - \frac{\tilde{\sigma}^2}{2}\right) T + \tilde{\sigma} \sqrt{T} \cdot w} - K \right) \right) = \\ &= e^{(\tilde{r}-r)T} \cdot \left(e^{-\tilde{r}T} \cdot E \left(\max \left(0, S(0) \cdot e^{\left(\tilde{r} - \frac{\tilde{\sigma}^2}{2}\right) T + \tilde{\sigma} \sqrt{T} \cdot w} - K \right) \right) \right). \end{aligned}$$

(continued)

This last expression $\left(e^{-\tilde{r}T} \cdot E \left(\max \left(0, S(0) \cdot e^{\left(\tilde{r} - \frac{\tilde{\sigma}^2}{2} \right) T + \tilde{\sigma} \sqrt{T} \cdot w} - K \right) \right) \right)$ is in fact nothing other than the fair value of a completely ordinary European plain vanilla call option with expiration T and strike K , however with modified parameters $\tilde{\sigma}$ and \tilde{r} instead of the parameters σ and r .

Using the formula for plain vanilla calls, we thus get

$$f_0^{geom} = S \cdot e^{(\tilde{r}-r)T} \mathcal{N}(d_1) - K \cdot e^{-rT} \cdot \mathcal{N}(d_2)$$

with

$$d_1 = \frac{\log \left(\frac{s}{K} \right) + \left(\tilde{r} + \frac{\tilde{\sigma}^2}{2} \right) T}{\tilde{\sigma} \sqrt{T}} \text{ and } d_2 = \frac{\log \left(\frac{s}{K} \right) + \left(\tilde{r} - \frac{\tilde{\sigma}^2}{2} \right) T}{\tilde{\sigma} \sqrt{T}},$$

and $\mathcal{N}(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2} dy$, where

$$\tilde{\sigma} := \sigma \cdot \sqrt{\frac{\left(1 + \frac{1}{M} \right) \cdot \left(2 + \frac{1}{M} \right)}{6}} \text{ and } \tilde{r} := r \cdot \left(\frac{1}{2} + \frac{1}{2M} \right) - \sigma^2 \cdot \left(\frac{1}{12} - \frac{1}{12M^2} \right).$$

So we now have an explicit formula for the fair value of a geometric Asian option.

In the special case that $M = 1$, we get $\tilde{\sigma} = \sigma$ and $\tilde{r} = r$ and hence again the original Black-Scholes formula for plain vanilla call options.

For M to infinity—that is, if we take the geometric mean over all assumed values—we get $\tilde{\sigma} = \sigma \cdot \sqrt{\frac{1}{3}}$ and $\tilde{r} = \frac{r}{2} - \frac{\sigma^2}{12}$.

In any case, $\tilde{\sigma} \leq \sigma$ and $\tilde{r} \leq r$ (and thus also always $e^{(\tilde{r}-r)T} \leq 1$).

Since the value of a European plain vanilla call option is always monotonically increasing in σ and in r , it follows that the value of a European geometric Asian option is always less than or equal to the value of the European plain vanilla option with the same parameters.

Let us illustrate this with a numerical example:

We choose

Strike price $K = 100$

Time to expiration $T = 1$

Risk-free interest rate $r = 0.01(1\%)$

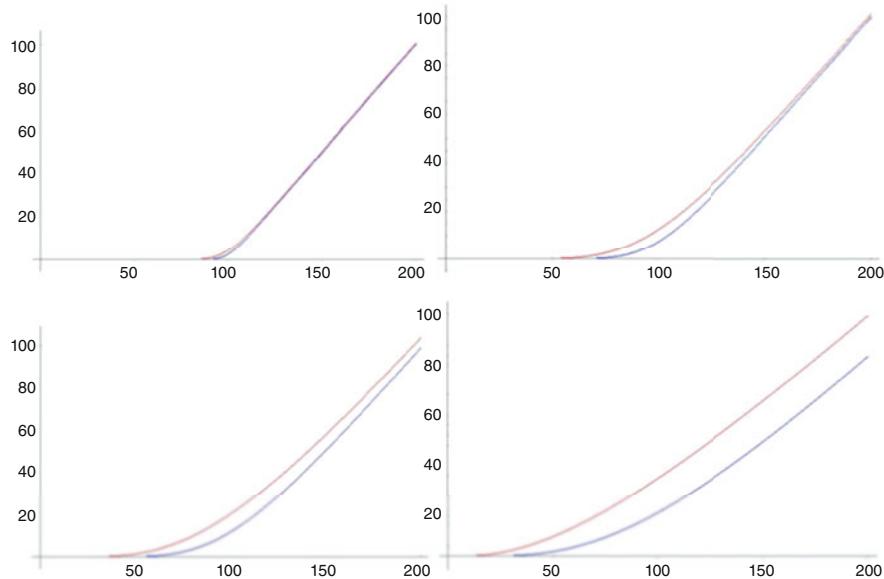


Fig. 2.27 Comparison of prices of plain vanilla call options (red) and geometric Asian call options (blue) as a function of the underlying asset's price for different volatility values (0.1, 0.3, 0.5, 1 from top left to bottom right)

Number of time points for computing the mean $M = 10$

For volatility, we choose four different values

$$\sigma = 0.1, 0.3, 0.5, 1$$

The results are illustrated in Fig. 2.27. The red curves are the price curves of the plain vanilla call option, and the blue ones represent the price curves of the geometric Asian option.

Having an explicit formula for pricing geometric Asian options will also be relevant and useful later when we use Monte Carlo methods to price arithmetic Asian options and try to accelerate these methods by means of so-called control variates.

Excursus

As emphasized above:

The geometric mean of a finite sequence a_1, a_2, \dots, a_M of positive values is always less than or equal to the arithmetic mean of these values. So we always have

(continued)

$$\sqrt[M]{\prod_{i=1}^M a_i} \leq \frac{1}{M} \sum_{i=1}^M a_i.$$

In the case of two values a_1 and a_2 (i.e. for $M = 2$), this inequality is

$$\sqrt{a_1 \cdot a_2} \leq \frac{a_1 + a_2}{2}$$

And this is equivalent to

$$\begin{aligned} a_1 \cdot a_2 &\leq \frac{(a_1 + a_2)^2}{4} \Leftrightarrow 4 \cdot a_1 \cdot a_2 \leq a_1^2 + 2 \cdot a_1 \cdot a_2 + a_2^2 \\ &\Leftrightarrow 0 \leq a_1^2 - 2 \cdot a_1 \cdot a_2 + a_2^2 \\ &\Leftrightarrow 0 \leq (a_1 - a_2)^2 \end{aligned}$$

and the last inequality is always true, since a square number is always greater than or equal to 0.

For arbitrary M , this inequality is usually proved by induction. This proof, however, is not entirely straightforward.

The shortest proof of this inequality is probably the one delivered by Hungarian mathematician György Polya, and we will present it briefly here:

The proof is based on the fact that the exponential function e^x is always $e^x \geq 1+x$. This is easily shown with a simple discussion of the curve $f(x) = e^x - 1 - x$. ($f(0) = 0$ and $f'(x)$ is less than 0 for x less than 0 and greater than 0 for x greater than 0.)

We set $x_i := \frac{a_i}{\frac{a_1+a_2+\dots+a_N}{N}} - 1$ and notice right away that $x_1 + x_2 + \dots + x_N = \frac{a_1+a_2+\dots+a_N}{\frac{a_1+a_2+\dots+a_N}{N}} - N = N - N = 0$.

It follows therefore that

$$\begin{aligned} 1 &= e^0 = e^{x_1+x_2+\dots+x_N} = \\ &= e^{x_1} \cdot e^{x_2} \cdot \dots \cdot e^{x_N} \geq (1 + x_1) \cdot (1 + x_2) \cdot \dots \cdot (1 + x_N) = \\ &= \frac{a_1}{\frac{a_1+a_2+\dots+a_N}{N}} \cdot \frac{a_2}{\frac{a_1+a_2+\dots+a_N}{N}} \cdot \dots \cdot \frac{a_N}{\frac{a_1+a_2+\dots+a_N}{N}} = \\ &= \frac{a_1 \cdot a_2 \cdot \dots \cdot a_N}{\left(\frac{a_1+a_2+\dots+a_N}{N}\right)^N}. \end{aligned}$$

(continued)

Thus,

$$1 \geq \frac{a_1 \cdot a_2 \cdot \dots \cdot a_N}{\left(\frac{a_1+a_2+\dots+a_N}{N}\right)^N};$$

hence,

$$\left(\frac{a_1 + a_2 + \dots + a_N}{N}\right)^N \geq a_1 \cdot a_2 \cdot \dots \cdot a_N$$

and so

$$\frac{a_1 + a_2 + \dots + a_N}{N} \geq \sqrt[N]{a_1 \cdot a_2 \cdot \dots \cdot a_N}$$

which is what we needed to show.

2.14 Brief Comment on Hedging Path-Dependent Derivatives

If path-dependent options (whether European or American) can be effectively priced by approximation in a binomial model, then approximate delta hedging as described earlier is also possible. We start again by approximating the Wiener model for the underlying asset with an N -step binomial model with (equidistant, if possible) time steps $n \cdot dt$ for $n = 0, 1, 2, \dots, N$, corresponding to the adjustment times for the hedging portfolio. The model is then split into the extended graph, as described in Sect. 2.10, and pricing is then performed in each node of this graph.

Throughout the life of the option, at each time point $n \cdot dt$ in that extended graph's respective node that we are currently at (given the path so far), the hedging portfolio is rebalanced such that we are perfectly hedged for the next step in the extended graph. (For the option as a derivative in the binomial model, this means perfect hedging.)

If path-dependent options can be priced explicitly (by means of the path-dependent version of the Black-Scholes equation, e.g. in the case of barrier options or in the case of the geometric Asian option), then we can differentiate the explicit valuation formula with respect to the underlying asset's price s to obtain the delta that we can then use for an (ideally perfect) hedging strategy.

In the case of more complex path-dependent options that need to be valued using, for example, Monte Carlo methods, the delta required to determine the hedging portfolio must then also be estimated by means of Monte Carlo simulation methods. How that is done will be addressed in a subsequent section.

2.15 Valuation of Derivatives Using Monte Carlo Methods, Basic Principle

Valuation of a derivative (be it plain vanilla or path-dependent) on an underlying asset that follows a Wiener model can be reduced, by using the Black-Scholes formula, to the calculation of an expected value.

The Monte Carlo method in turn is—when applied in its basic version—a very simple method for approximating expected values.

Most readers have likely already applied the Monte Carlo method in its simplest form at some point or other:

A random experiment may consist in rolling a certain dice. We are interested in the random variable Y , the “number rolled”, and wonder about the expected value of this random variable Y . One method that we could use to experimentally approximate this expected value is to generate a large number (N) of realizations y_1, y_2, \dots, y_N of that random variable Y and then use the average value of these realizations $\bar{Y} = \frac{y_1 + y_2 + y_3 + \dots + y_N}{N}$ as an approximation of the expected value.

In this example, we generate N realizations of the random variable Y simply by rolling the dice N times.

This example perfectly reflects the principle of the Monte Carlo method:

Approximate the expected value of a random variable Y by generating a large number N of realizations y_1, y_2, \dots, y_N of Y and computing the average over these realizations.

In real and general practice, Monte Carlo methods are usually applied as follows:

Suppose that we want to determine the expected value (the average value) $E(Y)$ of a random variable Y , where Y depends on various (mutually independent) random variables $\theta_1, \theta_2, \theta_3, \dots, \theta_s$ with known distribution and dependence structure, that is to say, $Y = Y(\theta_1, \theta_2, \theta_3, \dots, \theta_s)$. In this case, we would have a suitable random number generator on our computer produce N realizations $(\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_s^{(i)})$; $i = 1, 2, \dots, N$ of such vectors of random variables.

For each such vector, we compute the value of the random variable

$Y = Y(\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_s^{(i)})$ that results when such a vector (situation) occurs and determine the mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y(\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_s^{(i)})$$

as an approximation of $E(Y)$.

It is imperative in this context that these realizations are generated according to the known distribution and dependence structure of the random variables $\theta_1, \theta_2, \theta_3, \dots, \theta_s$.

Example 2.13 We are at time 0 and want to determine the fair value f_0 of a plain vanilla call option with expiration T and strike K on an underlying asset whose price

$S(t)$ in the time interval $[0, T]$ follows a Wiener model with trend μ and constant volatility σ . Let the risk-free interest rate in the time interval $[0, T]$ be r .

We already know that this value can be calculated explicitly using the Black-Scholes formula for call options. Nevertheless, to illustrate the Monte Carlo method, we are going to approximate the computation here using Monte Carlo (MC).

We know from the Black-Scholes theory that $f_0 = e^{-rT} \cdot E(\max(\tilde{S}(T) - K, 0))$. Here, $\tilde{S}(T) = S(0) \cdot e^{(r-\frac{\sigma^2}{2})T + \sigma \cdot \sqrt{T}w}$ with an $\mathcal{N}(0, 1)$ -distributed random variable w .

So, in the above terminology, we are looking at the expected value $E(Y)$ for the random variable $Y = Y(w) = \max(\tilde{S}(T) - K, 0) = \max\left(S(0) \cdot e^{(r-\frac{\sigma^2}{2})T + \sigma \cdot \sqrt{T}w} - K, 0\right)$.

Y thus depends only on one random variable $\theta_1 = w$.

We now run a random number generator for a total of N times, each time generating one independent $\mathcal{N}(0, 1)$ -distributed random number. We denote these values by $w^{(1)}, w^{(2)}, \dots, w^{(N)}$.

For each of these $w^{(i)}$, we compute the associated realization of Y , i.e. $Y(w^{(i)}) = \max\left(S(0) \cdot e^{(r-\frac{\sigma^2}{2})T + \sigma \cdot \sqrt{T}w^{(i)}} - K, 0\right)$.

The approximated value of f_0 is then given by $e^{-rT} \cdot \bar{Y}$ where $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y(w^{(i)})$.

Using our software (or Mathematica), we perform the approximation of f_0 below for the following numerical values:

Strike $K = 90$

Time to expiration $T = 1$

$r = 0.02$

$\sigma = 0.3$

Initial value $S(0)$ of the underlying asset = 100

The Black-Scholes formula for the call option yields the exact value $f_0 = 18.0691$.

We now run Monte Carlo simulation with 10,000 individual simulations.

To illustrate the convergence of the simulation, we continuously average the generated simulation values as we go. This means:

We compute $Y(w^{(i)})$ and concomitantly the emerging and successively improving approximations from 1, 2, 3, ..., N simulations; hence,

$$Y(w^{(1)}), \frac{Y(w^{(1)}) + Y(w^{(2)})}{2}, \frac{Y(w^{(1)}) + Y(w^{(2)}) + Y(w^{(3)})}{3}, \dots, \\ \frac{Y(w^{(1)}) + Y(w^{(2)}) + \dots + Y(w^{(N)})}{N}.$$

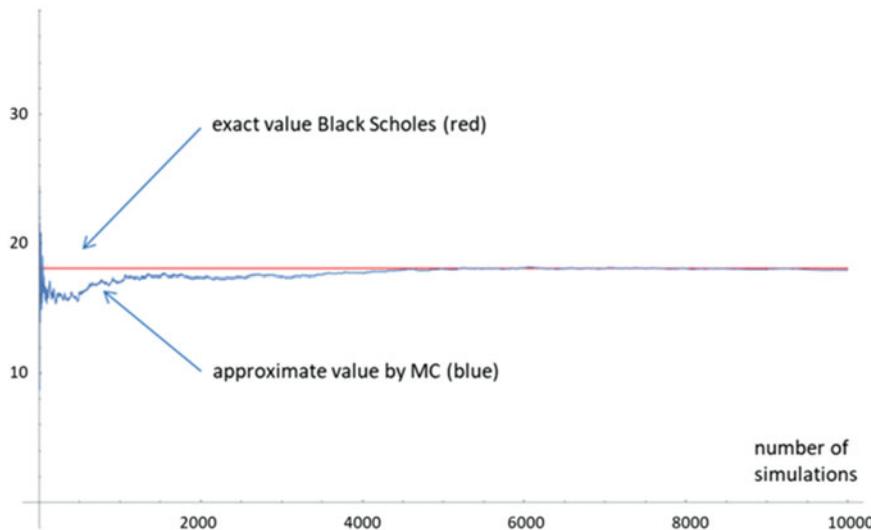


Fig. 2.28 Convergence for Monte Carlo simulation of the fair price of a plain vanilla call option, 10,000 simulations

We then plot these values, which tend to get increasingly better and closer to the true values, in a graph against the number of simulations used (see Fig. 2.28 for our specific example).

The red horizontal line represents the exact reference value—as determined using the Black-Scholes formula—at $f_0 = 18.0691$.

As an approximation of the option price, we could now use the last average value $\bar{Y} = \frac{Y(w^{(1)}) + Y(w^{(2)}) + \dots + Y(w^{(N)})}{N}$ ($= 17.9244$ in our example) or else the average over some of the last (e.g. the last 1000) approximations ($= 17.9864$ in our example).

Of course, each time we re-run the MC simulation, we will get (usually slightly) different results, since each time we run it, we are working with a new set of generated random numbers.

Comment

When calculating successive approximations $\bar{Y}(w^{(1)})$, $\frac{Y(w^{(1)})+Y(w^{(2)})}{2}$, $\frac{Y(w^{(1)})+Y(w^{(2)})+Y(w^{(3)})}{3}$, ..., $\frac{Y(w^{(1)})+Y(w^{(2)})+\dots+Y(w^{(N)})}{N}$, we do not recalculate one approximation after the other, of course (this would require about N^2 operations), but proceed recursively.

That is, if we denote $E(i) := \frac{Y(w^{(1)})+Y(w^{(2)})+\dots+Y(w^{(i)})}{i}$, then $E(i+1) = \frac{i \cdot E(i) + Y(w^{(i+1)})}{i+1}$.

(continued)

So, $E(i+1)$ can essentially be calculated from $E(i)$ and $Y(w^{(i+1)})$ in just one step. Calculating all approximations therefore essentially requires only N computational steps.

The option price obtained using MC simulation will, of course, differ from the exact price (we are going to look at just how big those differences can be **expected** to be in one of the next subsections). But, considering that the parameters needed for computing the option price are (more or less reliably) estimated values, such as (above all) the volatility parameter σ , we can easily accept such minor deviations as seen in our above example.

If we replace the volatility σ , which we had assumed to be 0.3 in our numerical example, by, say, 0.29 or 0.31 and calculate the fair values of the call option with those volatilities, we get fair values of 17.7304 and 18.4094, respectively.

We illustrate the situation in Fig. 2.29. There we ran MC simulation with 100,000 scenarios for $\sigma = 0.3$ and compared the simulation convergence with the exact values for $\sigma = 0.3$, 0.29, and 0.31. Thus, a slightly different estimate of the parameter σ has a significantly larger impact than the error of the Monte Carlo approximation.

When determining the fair price of a plain vanilla (i.e. a non-path-dependent) European derivative by MC simulation, it is not necessary to simulate the underlying

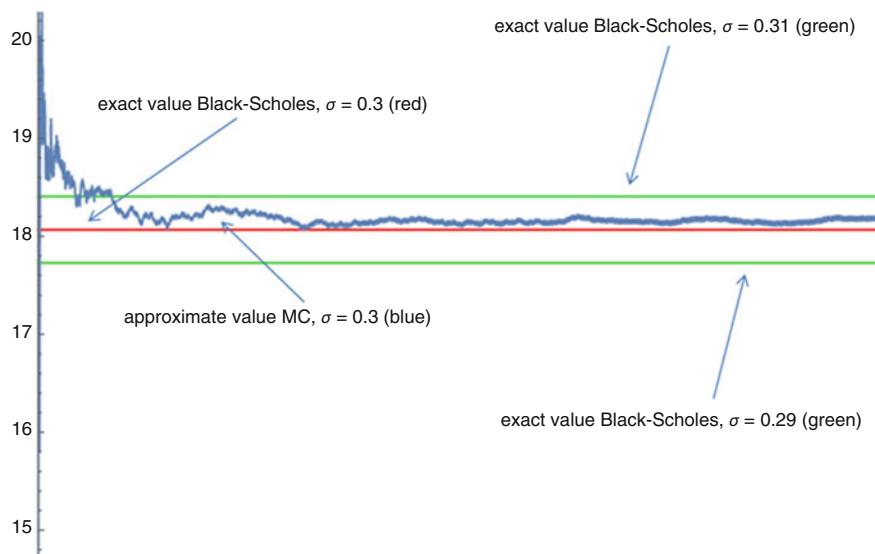


Fig. 2.29 Convergence for Monte Carlo simulation of the fair price of a plain vanilla call option, 100,000 simulations and comparison with results for modified volatility

asset's entire performance over time; it is sufficient to simulate only the last price of the underlying at time T .

This means: For each simulation scenario, we only need to generate **one** random number w . This situation changes significantly when it comes to pricing path-dependent options by MC simulation.

2.16 Valuation of European Path-Dependent Derivatives with Monte Carlo Methods

When pricing a European path-dependent option with payoff function Φ (for an underlying asset over a Wiener model) using MC simulation, we need to first of all distinguish between two different situations:

- (a) The payoff depends only on the values of the path at fixed predefined times $0 \leq t_1 < t_2 < \dots < t_s \leq T$ (e.g. Asian option)
- (b) The payoff depends on the asset's entire price path (e.g. barrier options or lookback options)

In case (a), we will again simulate values for the risk-neutral price of the underlying asset at time points t_1, t_2, \dots, t_s using the relationship that exists for the Wiener model, namely,

$$\tilde{S}(v) = \tilde{S}(u) \cdot e^{\left(r - \frac{\sigma^2}{2}\right) \cdot (v-u) + \sigma \sqrt{v-u} w}$$

for all u in $[0, T]$. In particular,

$$\begin{aligned} \tilde{S}(t_i) &= \tilde{S}(t_{i-1}) \cdot e^{\left(r - \frac{\sigma^2}{2}\right) \cdot (t_i - t_{i-1}) + \sigma \sqrt{t_i - t_{i-1}} w_i} = \dots = \\ S(0) \cdot e^{\left(r - \frac{\sigma^2}{2}\right) t_i + \sigma (\sqrt{t_i - t_{i-1}} w_i + \sqrt{t_{i-1} - t_{i-2}} w_{i-1} + \dots + \sqrt{t_2 - t_1} w_2 + \sqrt{t_1} w_1)} \end{aligned}$$

According to the Black-Scholes formula for path-dependent options, the option's fair price f_0 at time 0 is

$$f_0 = e^{-rT} \cdot E(\Phi(\tilde{S}(t_1), \dots, \tilde{S}(t_s))) = e^{-rT} \cdot E(\Psi(w_1, w_2, \dots, w_s)).$$

Here, $\Psi(w_1, w_2, \dots, w_s) :=$

$$\begin{aligned} &:= \Phi\left(S(0) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)t_1 + \sqrt{t_1} w_1}, S(0) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)t_2 + \sqrt{t_2 - t_1} w_2 + \sqrt{t_1} w_1}, \right. \\ &\quad \dots, S(0) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)t_s + \sigma (\sqrt{t_s - t_{s-1}} w_s + \sqrt{t_{s-1} - t_{s-2}} w_{s-1} + \dots + \sqrt{t_2 - t_1} w_2 + \sqrt{t_1} w_1)} \Big). \end{aligned}$$

Here, w_1, w_2, \dots, w_s are always independent $\mathcal{N}(0, 1)$ -distributed random variables.

So, for each simulation, we need to generate a vector $(w_1^{(i)}, w_2^{(i)}, \dots, w_s^{(i)})$ of s independent $\mathcal{N}(0, 1)$ -distributed random numbers for $i = 1, 2, \dots, N$, if our plan is to run N simulations again. Thus, the effort increases s -fold compared to the valuation of plain vanilla options.

We determine $\Psi(w_1^{(i)}, w_2^{(i)}, \dots, w_s^{(i)})$ for all $i = 1, 2, \dots, N$ and use

$\bar{\Psi} := \frac{\sum_{i=1}^N \Psi(w_1^{(i)}, w_2^{(i)}, \dots, w_s^{(i)})}{N}$ as an approximation for $E(\Phi(\tilde{S}(t_1), \tilde{S}(t_2), \dots, \tilde{S}(t_s)))$, so $e^{-rT} \cdot \bar{\Psi}$ as an approximation for f_0 .

In the second case (b), we cannot approximate the required value

$E(\Phi((\tilde{S}_t)_{t \in [0, T]})$) directly by MC simulation, because with MC simulation, we cannot generate an entire path on $[0, T]$, just a finite number of values on that path. So, our first step in case (b) would be to select a finite number of time points $0 \leq t_1 < t_2 < \dots < t_s \leq T$ at which the price is to be simulated. These simulated values can then be linearly connected to obtain a path on the entire time interval. We then proceed as in case (a). This “discretization” of the path, however, already leads to a first approximation error, as we no longer simulate the actually required value $E(\Phi((\tilde{S}_t)_{t \in [0, T]}))$, but a (slightly) different value. This (first) approximation error decreases, of course, the more time points t_1, t_2, \dots, t_s we choose. On the other hand, the complexity of the MC simulation increases as s grows (each individual simulation requires us to generate s random numbers).

We will illustrate the procedure for both case (a) and case (b) here with two numerical examples. In the case studies that we are going to look at later, we will then frequently deal with the valuation of path-dependent derivatives by means of MC simulation.

2.17 Monte Carlo Valuation of Asian Options

We are now going to price a European geometric Asian call option with the following parameters:

Strike $K = 90$

Time to expiration $T = 1$

$r = 0.02$

$\sigma = 0.3$

So the option's time to expiration is 1 year. We compute the average based on 12 values measured at equal distances (at the end of each trading month).

The underlying asset's initial price $S(0)$ is 100 points.

The exact formula for pricing geometric Asian options, which we derived in 2.13, yields a fair value of 13.061.

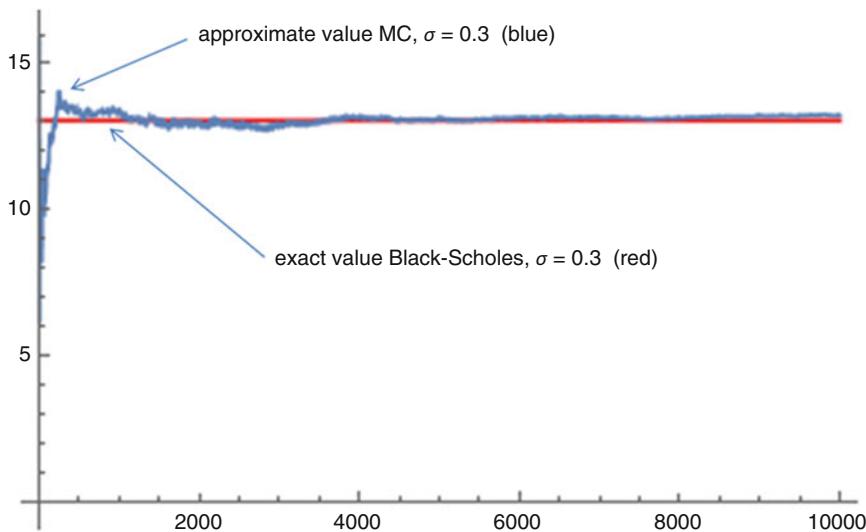


Fig. 2.30 Convergence for Monte Carlo simulation of the fair price of a European geometric Asian call option, 10,000 simulations

MC simulation with 10,000 scenarios produces the convergence graph displayed in the subsequent Fig. 2.30:

The approximate value as an average of the last 1000 approximations yields a value of 13.172 in this simulation.

The simulation was performed such that 10,000 paths were simulated in 12 steps. Figure 2.31 shows the first 30 paths generated for the MC simulation.

For comparison, we now also price the arithmetic Asian option with the same parameters by Monte Carlo simulation and compare that with the value of the geometric Asian option. The result is shown in Fig. 2.32.

2.18 Monte Carlo Valuation of Barrier Options

As another example of the application of Monte Carlo simulation in derivatives valuation, we are going to look at the case of a European down-and-out barrier call option. The advantage of this is that here again, we have a reference value that we can use to check how well the Monte Carlo method performs.

The basic parameters that we choose are:

Strike price $K = 90$

Time to expiration $T = 1$

$r = 0.02$

$\sigma = 0.3$

As barrier L , we choose $L = 60$.

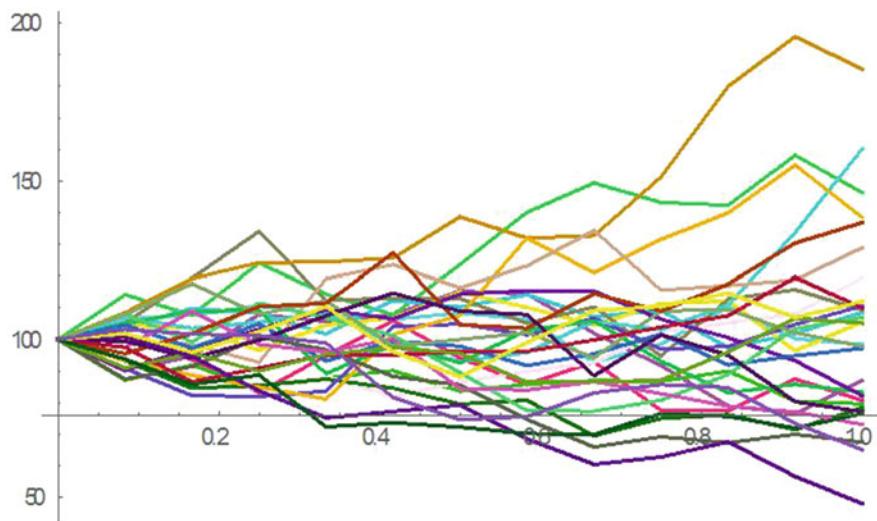


Fig. 2.31 The first 30 paths used to simulate the geometric Asian option

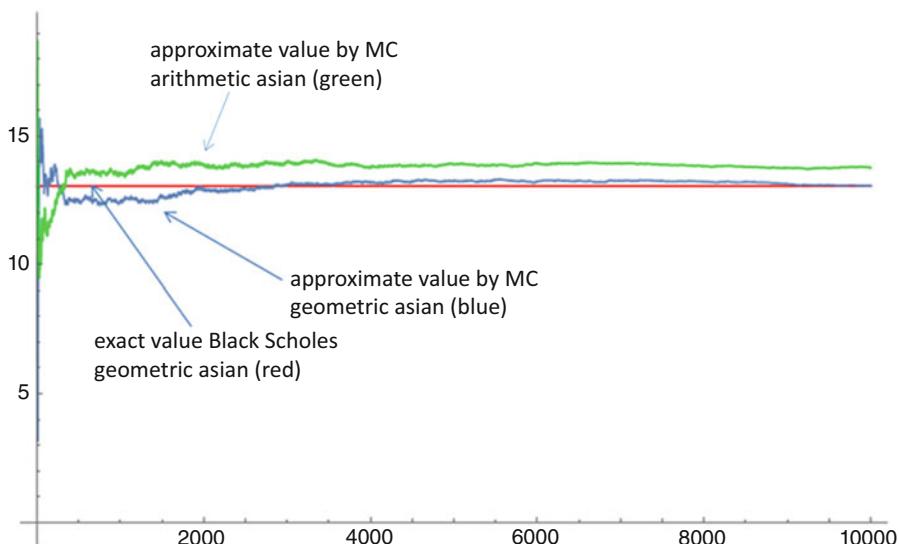


Fig. 2.32 Convergence for Monte Carlo simulation of the fair price of a European geometric Asian call option (blue) and a European arithmetic Asian call option (green), 10,000 simulations

First, we run simulations for the underlying asset's initial value of $S(0) = 100$. Using the formula for the fair price of down-and-out barrier calls from Sect. 2.12, we get the reference value 18.0602.

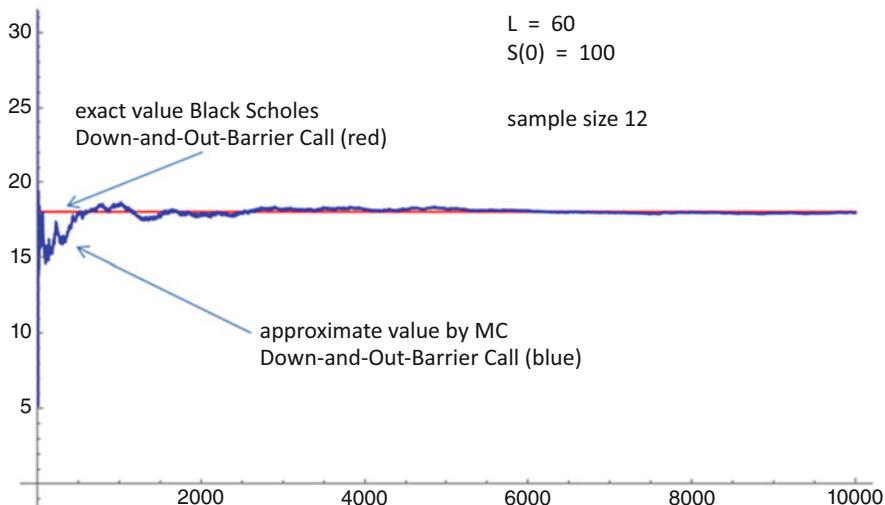


Fig. 2.33 Convergence for Monte Carlo simulation of the fair price of a down-and-out barrier call with barrier $L = 60$, initial value $S(0) = 100$, and 12 sample points

As described in Sect. 2.16 for case (b), we now need to perform a discretization first. A coarser discretization (fewer samples) means a poorer approximation of the actual problem, but on the other hand a less complex MC simulation. A finer discretization (more samples) will describe the actual problem more accurately, but will lead to a longer runtime for the MC simulation. We start with a very coarse discretization of $s = 12$ and a subdivision of the time interval $[0, T]$ into 12 time segments of equal length. We are again going to simulate 10,000 scenarios. The convergence behaviour is shown in Fig. 2.33.

The simulation yields an approximate value of 17.9715. We thus obtain a very good approximation and compelling convergence despite the fact that we chose only a very small number of sample points (one time point per month).

This situation worsens significantly when the initial value $S(0)$ is chosen closer to the value of the barrier. If we choose, say, $S(0) = 65$, the exact fair value is 1.24955. Simulation with the same 10,000 scenarios and 12 sample points now yields an approximate value of 1.549 and the convergence behaviour shown in Fig. 2.34. The simulation converges visibly, but toward a value that is obviously significantly different from the actual value.

So, to obtain a better result, we clearly need to substantially increase the number of sample points. Increasing the number of samples to 100 (control values at 100 equidistant time points on the interval $[0, T]$) yields the approximate value 1.427 for the exact value 1.24955 and the following convergence behaviour (Fig. 2.35):

Increasing the number of samples from 12 to 100 causes the runtime of the simulation to increase approximately eightfold. The outcome is a slight improvement over the simulation for 12 samples, but is still not satisfactory. It is only when

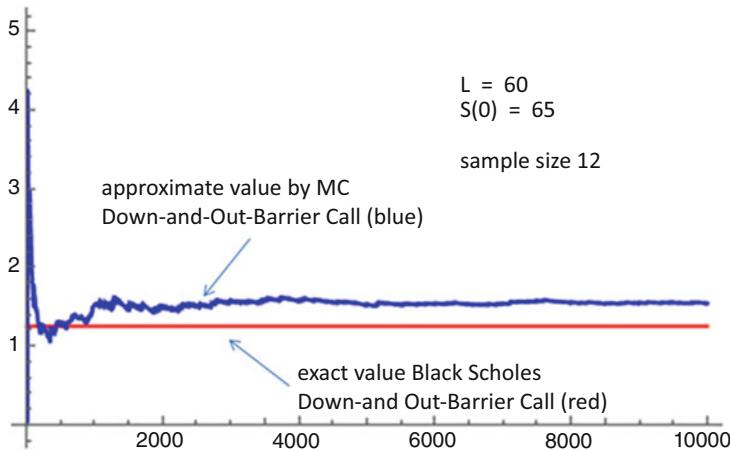


Fig. 2.34 Convergence for Monte Carlo simulation of the fair price of a down-and-out barrier call option with barrier $L = 60$, initial value $S(0) = 65$, and **12 sample points**

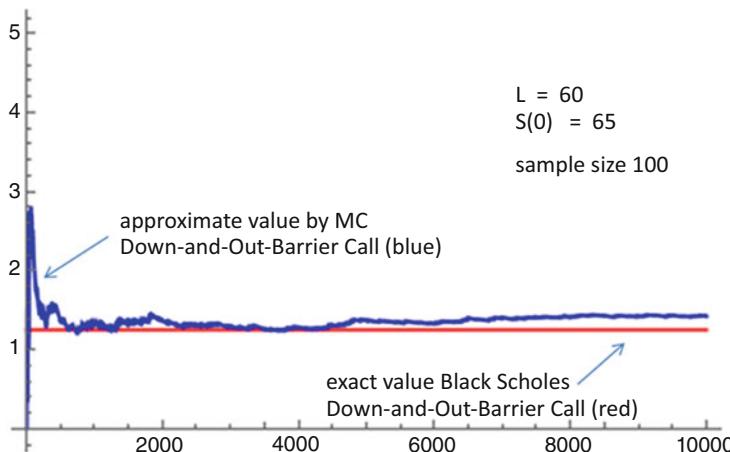


Fig. 2.35 Convergence for Monte Carlo simulation of the fair price of a down-and-out barrier call option with barrier $L = 60$, initial value $S(0) = 65$, and **100 sample points**

we increase the number of samples to 1000, at yet another approximately tenfold increase in the simulation runtime, and obtain the approximate value 1.2977 and the convergence behaviour shown in Fig. 2.36 that we get a reasonably acceptable result.

Here, the paths used for pricing the barrier option are simulated by means of 1000 individual steps each. The first 30 paths used for the MC simulation in the above example are shown in Fig. 2.37.

There are various approaches to improve the valuation of barrier options and accelerate the convergence rate when initial values are close to the barrier. However,

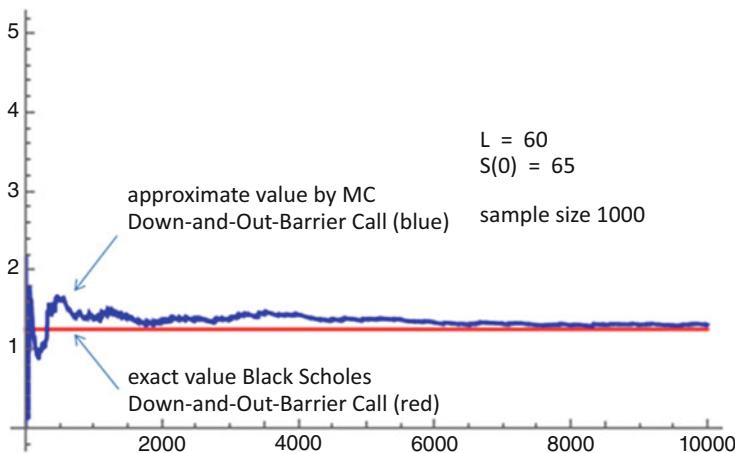


Fig. 2.36 Convergence for Monte Carlo simulation of the fair price of a down-and-out barrier call option with barrier $L = 60$, initial value $S(0) = 65$, and 1000 sample points

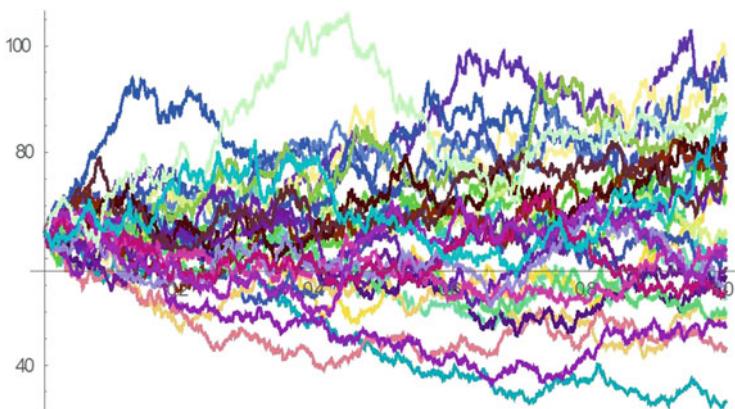


Fig. 2.37 30 simulation paths to determine the value of a down-and-out barrier call option with 1000 sample points

we will not address these variants here and now. Yet it should be clear from the example that we must always carefully weigh the reliability of MC simulation results—especially, of course, when we do not have exact values to compare them with.

2.19 Barrier Options in Turbo and Bonus Certificates

Barrier options are embedded (in part implicitly) in a wide variety of certificates (warrants) offered by investment firms or banks.

Product														
Product name: Classic TURBO warrants (put) on the DAX WKN/German securities identification number: CU68HN ISIN: DE000CU68HN4 Issued by: Commerzbank AG														
Objectives														
The product has a fixed life and is settled on the settlement date. If the price of the underlying never rises to or above the knock-out barrier (put) within the period from the issue date to the valuation date, the product is redeemed at a payout amount equal to the amount, multiplied by the multiplier, by which the reference price of the underlying is lower than the strike price on the valuation date. If the price of the underlying reaches or exceeds the knock-out barrier even just once, the product expires worthless and no payment is made.														
For calculation purposes, 1 index point corresponds to EUR 1.00.														
Static Data														
<table> <tbody> <tr> <td>Underlying: DAX Index</td> <td>Issue date: 16 October 2019</td> </tr> <tr> <td>Currency of the underlying: EUR</td> <td>Valuation date: 20 March 2020</td> </tr> <tr> <td>Currency of the product: EUR</td> <td>Settlement date: 27 March 2020</td> </tr> <tr> <td>Type: Put</td> <td>Multiplier: 0.01</td> </tr> <tr> <td>Strike price: 13,650 points</td> <td>Settlement type: Cash</td> </tr> <tr> <td>Knock-out barrier: 13,650 points</td> <td>Exercise type: European</td> </tr> <tr> <td>Reference price: Closing price of the index on the respective business day</td> <td></td> </tr> </tbody> </table>	Underlying: DAX Index	Issue date: 16 October 2019	Currency of the underlying: EUR	Valuation date: 20 March 2020	Currency of the product: EUR	Settlement date: 27 March 2020	Type: Put	Multiplier: 0.01	Strike price: 13,650 points	Settlement type: Cash	Knock-out barrier: 13,650 points	Exercise type: European	Reference price: Closing price of the index on the respective business day	
Underlying: DAX Index	Issue date: 16 October 2019													
Currency of the underlying: EUR	Valuation date: 20 March 2020													
Currency of the product: EUR	Settlement date: 27 March 2020													
Type: Put	Multiplier: 0.01													
Strike price: 13,650 points	Settlement type: Cash													
Knock-out barrier: 13,650 points	Exercise type: European													
Reference price: Closing price of the index on the respective business day														

Fig. 2.38 English translation of the key data for Classic TURBO warrant

An example of such a certificate is the “Classic TURBO warrant (put) on the DAX” offered by Commerzbank (now Société Générale) on 18 October 2019, for which we include an extract from its key investor documentation (KID) given in Fig. 2.38.

We see that this is an up-and-out barrier put option on the DAX with strike 13.650 and barrier 13.650 and expiration date 20 March 2020 (so, with about 5 months left until expiration). However, the turbo warrant has an exercise ratio of 0.01. This means: Only one hundredth of the payoff of an ordinary barrier put option will be paid out.

The quotes for this product on 18 October 2019 at 13:26:08 h were 10.14 // 10.15. At that time, the DAX stood at 12,674.38 points. We could therefore have bought this Classic TURBO warrant at that time at a price of 10.15 EUR.

For comparison, we are going to price the corresponding up-and-out barrier put option on the DAX with strike 13.650 and barrier 13.650 and expiration date 20 March 2020 using our Monte Carlo program. (As we know from Sect. 2.12, we would also have an explicit valuation formula at our disposal, but we are going to use just the Monte Carlo simulation here.)

The simulation (given a current implied volatility of the DAX of 14.5% and assuming a risk-free interest rate of 0%) yields a value of 1010.51 EUR. At an exercise ratio of 0.01, this would result in a price of 10.11, which is approximately the same as the price of the Classic TURBO warrant of 10.15. The price of the equivalent put option without barrier, on the other hand, would be 1125.53 (or 11.26 at an exercise ratio of 0.01).

Product	
Product name: Bonus Certificates Classic on the DAX index WKN/German securities identification number: CUOADL ISIN: DE000CUOADL2 Issued by: Commerzbank AG	
Objectives	
The product has a fixed life and is settled on the settlement date. There are several ways for the product to be redeemed on the settlement date, as follows: a) If the price of the underlying is never at or below the barrier during the observation period, the product will be redeemed in one of the two following ways: i) If the reference price of the underlying on the valuation date is at or below the bonus level, you will receive the bonus amount ii) Otherwise, you will receive a payout amount equal to the reference price of the underlying on the valuation date multiplied by the multiplier. b) If the price of the underlying is at or below the barrier at least once during the observation period, you will receive a cash amount on the settlement date equal to the reference price of the underlying on the valuation date multiplied by the multiplier.	
Static Data	
Underlying: DAX Index Currency of the underlying: EUR Currency of the product: EUR Barrier: 10,150 points Bonus level: 12,700 points Bonus amount: EUR 127	Reference price: Closing price of the index on the respective business day Issue Date: 30 January 2019 Observation period: 30 January 2019 to 20 December 2019 Valuation date: 20 December 2019 Settlement date: 31 December 2019 Multiplier: 0.01 Settlement type: Cash

Fig. 2.39 English translation of the key data for the Bonus Certificate Classic

Another example is the “Bonus Certificate Classic on the DAX” offered by Commerzbank (now Société Générale) on 17 October 2019. An excerpt from the product’s description is shown in Fig. 2.39.

The quotes for the product on 17 October 2019 at 10:16:02 h were 129.43 // 129.45. At that time, the DAX was at 12,685.56 points.

So, this Bonus Certificate Classic has the following properties:

If the DAX is above 12,700 points on 20 December 2019, we receive one hundredth (exercise ratio 100:1!) of the then current DAX price in EUR. If the DAX is below 12,700 points on 20 December 2019, we receive 127 EUR.

However, if the DAX falls below 10,150 points at any time during the certificate’s life, even if just once, our payout will be one hundredth of the DAX in EUR (the lower payout limit of 12,700 is then no longer valid).

On 17 October 2019 at 10:16, this certificate was available for purchase at 129.45 EUR.

The certificate’s profit function can thus be represented as shown in Fig. 2.40. (The red and blue graphs have the same shape for $x > 12,945$ and are shown slightly offset in Fig. 2.40, for better visibility.)

In principle, we get the same profit function if we combine an investment in one hundredth of the DAX with one hundredth of a down-and-out barrier put option with expiration 20 December 2019, strike 12,700, and barrier at 10,150. We say “in principle” because, depending on the actual price of the down-and-out barrier put option in question, the profit function may be slightly higher or lower compared to the profit function of the certificate.

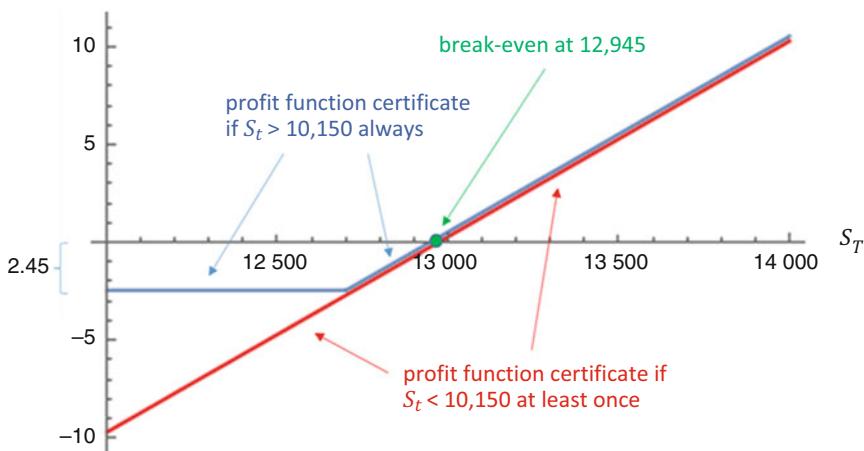


Fig. 2.40 Profit function for Bonus Certificate Classic on 17 October 2019

If we use our Monte Carlo approach to calculate the fair price of one hundredth of such a down-and-out barrier put option with the above parameters, then we get, when assuming the parameters,

$S_0 = 12,685$
 Barrier = 10,150
 Strike = 12,700
 $r = 0$
 Volatility = 13.1%
 Time left to expiration = $\frac{1}{12}$
 A price of about 2.64 EUR.

Therefore, if we create the Commerzbank Bonus Certificate Classic using an investment in the underlying instrument and this down-and-out barrier put option, then this combination (with the above data as on 17 October 2019) is priced at $126.85 + 2.64 = 129.49$ EUR, which is roughly the current price of the certificate.

2.20 Estimating Greeks (Especially Delta and Gamma) of Derivatives with Monte Carlo

If we want to hedge a derivative perfectly (at least in theory), we need the delta of that derivative at every moment of its life. This means that, if we do not have an explicit formula for the derivative's fair price, we have no choice but to estimate the delta too—using Monte Carlo, for example.

For hedging purposes, the delta needs to be estimated at every point in time t in $[0, T]$ at which the hedging portfolio is to be adjusted, for the parameters applicable

at each of these times. However, for the following, we can limit ourselves to $t = 0$ in each case.

So, we are at time 0 and the current price of the underlying asset is S_0 . By $F(s)$, we denote the fair value of the derivative (at time 0, with fixed parameters r and σ) when the underlying price is s . The delta then is $\Delta(S_0) = F'(S_0) = \frac{dF(s)}{ds} \Big|_{s=S_0}$. We now have two different ways to approximate this value.

First Approach: Using the Differential Quotient

We approximate $F'(S_0)$ for a fixed chosen small value h by $\frac{F(S_0+h)-F(S_0)}{h}$ and approximate this quantity by first calculating the fair price $F(S_0 + h)$ and the fair price $F(S_0)$ by Monte Carlo simulation, as described above. It is often recommended to approximate a kind of symmetrized differential quotient, namely, $\frac{F(S_0+h)-F(S_0-h)}{2h}$, instead of the usual differential quotient $\frac{F(S_0+h)-F(S_0)}{h}$. The results are indeed often better, but in the following, we are going to limit our discussion to the usual differential quotient (note that everything said below applies equally to the symmetrized quotient).

This very obvious approach has some pitfalls, however:

First (just like in the example of the barrier option in the previous section), we do not simulate the value directly but an approximate value. How close this approximation $\frac{F(S_0+h)-F(S_0)}{h}$ comes to the actual value of $F'(S_0)$ depends on how small our chosen h is. The smaller the h , the smaller the difference between the actual value and its approximation tends to be. Now, why then can we not just choose such an extremely small h that the difference between $\frac{F(S_0+h)-F(S_0)}{h}$ and $F'(S_0)$ becomes negligible?

Well, in the MC simulation of $F(S_0 + h)$, and of $F(S_0)$, we have to expect simulation errors of a certain magnitude (the larger the number of scenarios used for the simulation, the smaller the error tends to be). Suppose we have to expect an MC simulation error of a small positive magnitude of ε (e.g. $\varepsilon = 0.01$). If we now choose h less than or equal to ε (e.g. $h = \varepsilon = 0.01$), it could happen that, due to the MC simulation, the estimated value $F(S_0 + h)$ is too high by ε and that the estimated value $F(S_0)$ is too low by ε .

So instead of the value $\frac{F(S_0+h)-F(S_0)}{h}$, we get the value

$$\begin{aligned} \frac{F(S_0 + h) + \varepsilon - (F(S_0) - \varepsilon)}{h} &= \frac{F(S_0 + h) - F(S_0) + 2\varepsilon}{h} = \\ &= \frac{F(S_0 + h) - F(S_0)}{h} + \frac{2\varepsilon}{h} \geq \frac{F(S_0 + h) - F(S_0)}{h} + 2. \end{aligned}$$

We therefore need to be aware of a potentially significant deviation from the value $\frac{F(S_0+h)-F(S_0)}{h}$. We can prevent this only by choosing the parameter h such that it is significantly larger than ε , so as to be certain that the potential error term $\frac{2\varepsilon}{h}$ remains sufficiently small. However, this will adversely affect the approximation of $F'(S_0)$ by means of $\frac{F(S_0+h)-F(S_0)}{h}$.

Second Approach: Using the Derivative of the Payoff Function

We want to find $F'(S_0) = \frac{dF(s)}{ds} \Big|_{s=S_0}$.

Let us focus on plain vanilla first, i.e. on non-path-dependent options (using of course, as always so far, an underlying asset that follows a Wiener model). According to the Black-Scholes formula,

$$F(S_0) = e^{-rT} \cdot E(\Phi(\tilde{S}(t))) = e^{-rT} \cdot E\left(\Phi\left(S(0) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w}\right)\right)$$

and thus

$$F'(S_0) = \frac{dF(s)}{ds} \Big|_{s=S_0} = e^{-rT} \cdot \frac{d}{ds} E\left(\Phi\left(s \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w}\right)\right) \Big|_{s=S_0}.$$

If we now assume that we are allowed to swap the differential operator $\frac{d}{ds}$ for the expectation operator E (this is indeed possible in most real applications) and if we observe that differentiation of $\Phi\left(s \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w}\right)$ has to be done using the chain rule, we get

$$\begin{aligned} F'(S_0) &= e^{-rT} \cdot E\left(\frac{d}{ds}\left(\Phi\left(s \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w}\right)\Big|_{s=S_0}\right)\right) = \\ &= e^{-rT} \cdot E\left(\Phi'\left(s \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w}\right) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w} \Big|_{s=S_0}\right) \end{aligned}$$

and that is very helpful for what follows: If the payoff function Φ can be differentiated in a reasonable way, then this final expected value can again be simulated directly using Monte Carlo as usual. We don't have any approximation error between derivative and differential quotient, we do not have to choose a fixed value h in advance, and more importantly, this small value h does not appear in a denominator, which—as shown above—could cause difficulties.

Example 2.14 To better illustrate this second approach, let us first consider the case of a European plain vanilla call option with strike K . The payoff function here is given by $\Phi(x) = \max(0, x - K)$. The function Φ is differentiable everywhere except for $x = K$.

We have (see Fig. 2.41) $\Phi'(x) = \begin{cases} 0 & \text{for } x < K \\ 1 & \text{for } x > K \end{cases}$.

The fact that Φ is not differentiable in K can be disregarded, as we can define $\Phi'(K) := 1$, for example, and thus $\Phi'(x) = \begin{cases} 0 & \text{for } x < K \\ 1 & \text{for } x \geq K \end{cases}$.

We denote this function by $1_{x \geq K}(x)$. And so,

$F'(S_0) = e^{-rT} \cdot E \left(1_{x \geq K} \left(S_0 \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w} \right) \cdot e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}w} \right)$, which we can now simulate using Monte Carlo.

We run both approaches to simulating the delta of a European plain vanilla call option in the following and compare the approximations with the exact delta value, which is available to us as a reference value in this case. Remember: The delta of a call option at time 0 at a current underlying asset price of S_0 is given by $\mathcal{N}(d_1)$ with $d_1 = \frac{\log(\frac{S_0}{K}) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}$.

For the parameters that we have chosen, this results in a delta of 0.7149, which is our reference value for the following simulations.

We first simulate the delta using the first approach (approximation of the differential quotient). A first attempt with $h = 0.1$ and 10,000 scenarios yields the completely unsatisfactory situation shown in Fig. 2.42.

A slightly better yet still unsatisfactory outcome is obtained when choosing $h = 0.5$ and running 10,000 scenarios (Fig. 2.43).

A couple of further attempts demonstrate that a stable good convergence behaviour close to the actual delta value is not obtainable until we choose (e.g.) $h = 1$ and generate 100,000 scenarios. The simulation with the outcome shown in Fig. 2.44 yields an approximate value of 0.742.

When simulating the delta using the second approach, i.e. by approximating the expected value of the derivative of the payoff, we get very good results in a much shorter timespan, as is clearly shown in Fig. 2.45 (10,000 simulations, approximate value 0.719).

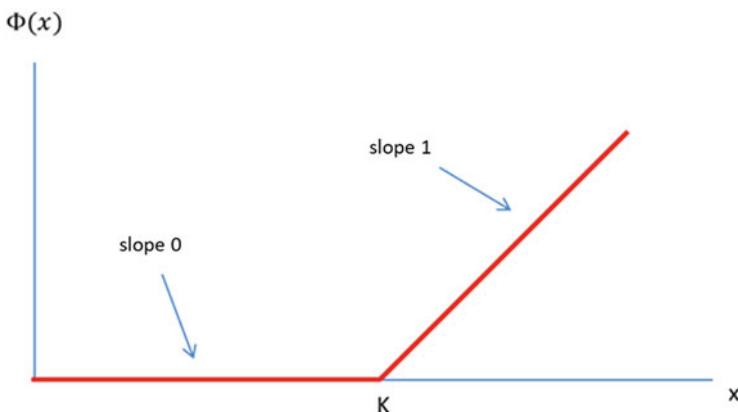


Fig. 2.41 Payoff function of a call option

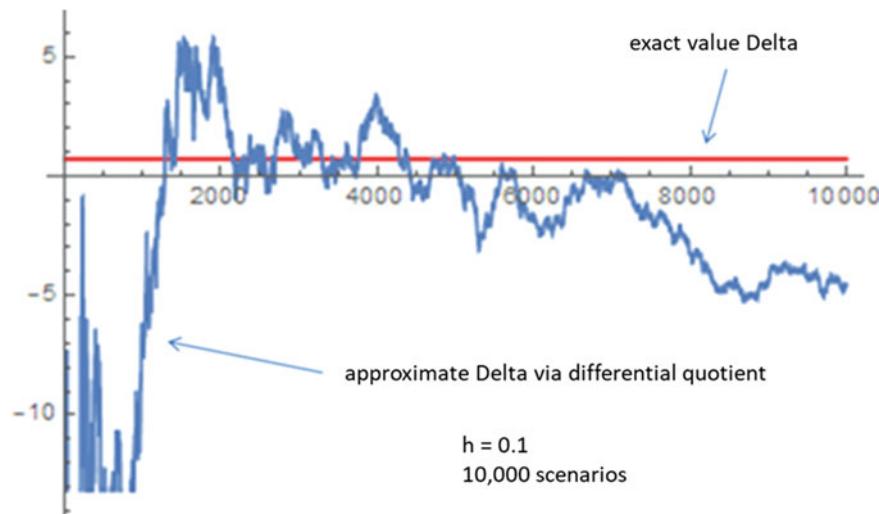


Fig. 2.42 Delta approximation for a plain vanilla call option using the differential quotient, $h = 0.1$, 10,000 scenarios

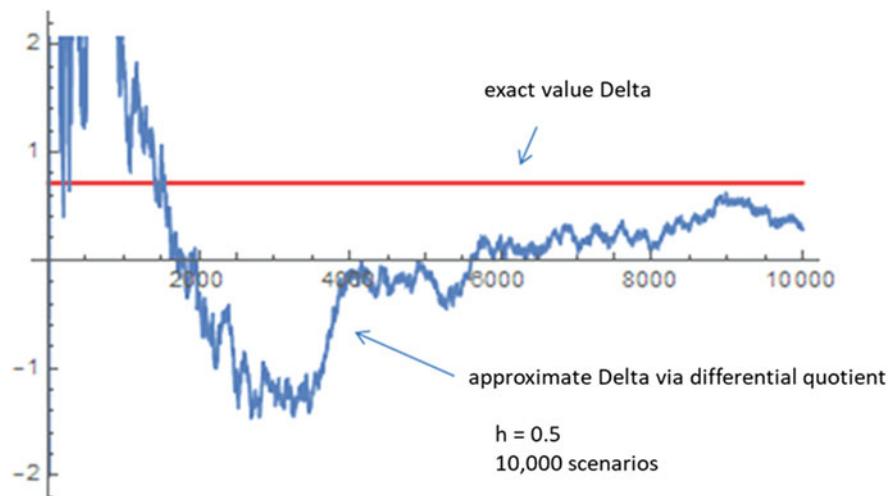


Fig. 2.43 Delta approximation for a plain vanilla call option using the differential quotient, $h = 0.5$, 10,000 scenarios

This approach can of course be applied similarly in estimating the other Greeks (approximation of the differential quotient or using the expected value of the derivative of the payoff with respect to the relevant parameter in each case).

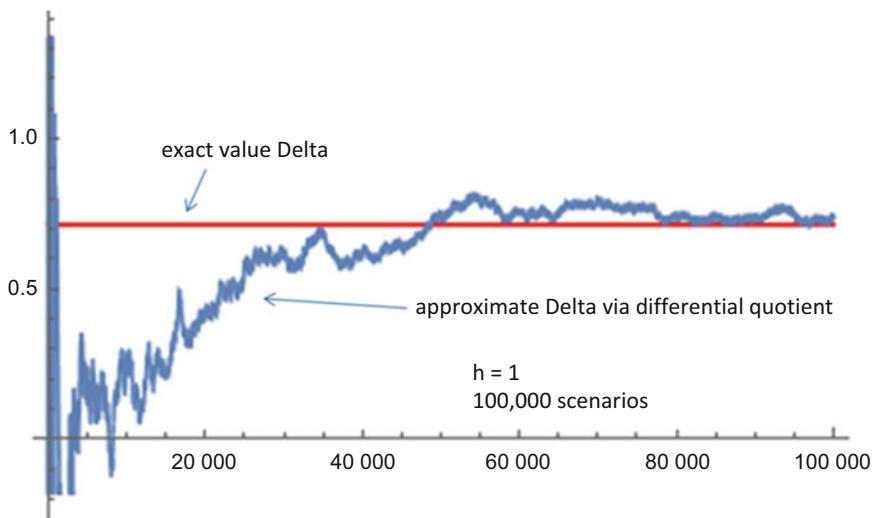


Fig. 2.44 Delta approximation for a plain vanilla call option using the differential quotient, $h = 1$, 100,000 scenarios

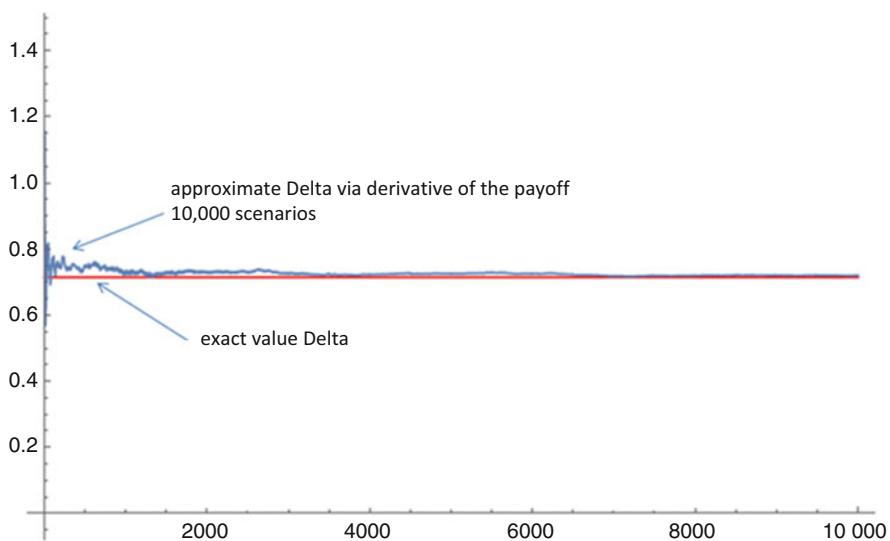


Fig. 2.45 Delta approximation for a plain vanilla call option using the derivative of the payoff, 10,000 scenarios

An exception to this is the gamma, which, as we recall, is defined as the second derivative of the option price with respect to the underlying asset's price S . Here too, we can of course start to experiment again by simulating an approximating differential quotient:

So again, we choose a parameter h and seek to approximate $F''(S_0) = \frac{d^2F(s)}{ds^2}\Big|_{s=S_0}$ by differential quotients, i.e. by

$$\begin{aligned} \frac{F'(s+h) - F'(s)}{h} &\sim \frac{\frac{F(s+2h)-F(s+h)}{h} - \frac{F(s+h)-F(s)}{h}}{h} = \\ &= \frac{F(s+2h) + F(s) - 2F(s+h)}{h^2} \end{aligned}$$

or by the symmetrized versions of the differential quotients, which may be better in many cases, i.e. by

$$\begin{aligned} \frac{F'(s+h) - F'(s-h)}{2h} &\sim \frac{\frac{F(s+2h)-F(s)}{2h} - \frac{F(s)-F(s-2h)}{2h}}{2h} = \\ &= \frac{F(s+2h) + F(s-2h) - 2F(s)}{4h^2}. \end{aligned}$$

Of course, to determine the gamma approximately, we can again try to differentiate the payoff function twice and go for direct approximation. However, this is not possible even for the simplest version, the plain vanilla call option: It is not only that F' is not differentiable in $x = K$ (when we determined the delta, we saw that this is not necessarily a problem); what compounds the problem is that the one-sided derivatives in this point are infinitely large, and this is a fact that renders this approach impracticable in the case of gamma.

Another conceivable variant would be to choose a parameter h , approximate $F'(s+h)$ and $F'(s)$ by means of the second approach, and then use these approximations (denoted by $D(s+h)$ and $D(s)$) to approximate the gamma using $\frac{D(s+h)-D(s)}{h}$.

The first approach using the differential quotient was carried out with the same parameters as above for the plain vanilla call option. The reference value, i.e. the exact gamma, is 0.01132. A typical simulation outcome (using the symmetrized approach for the differential quotients, choosing $h = 1$ and running 100,000 scenarios) is shown in Fig. 2.46.

The result is not usable in this form, which is partly due to the fact that the gamma is very small. The suggested approximate value for this simulation was 0.0567.

We are also going to try the second approach, i.e. determine delta at S_0 and $S_0 + h$ using MC, differentiating the payoff function, and using these values to approximate the gamma by means of the differential quotient. Choosing $h = 0.5$ and 100,000 scenarios typically yields a convergence behaviour as shown in Fig. 2.47.

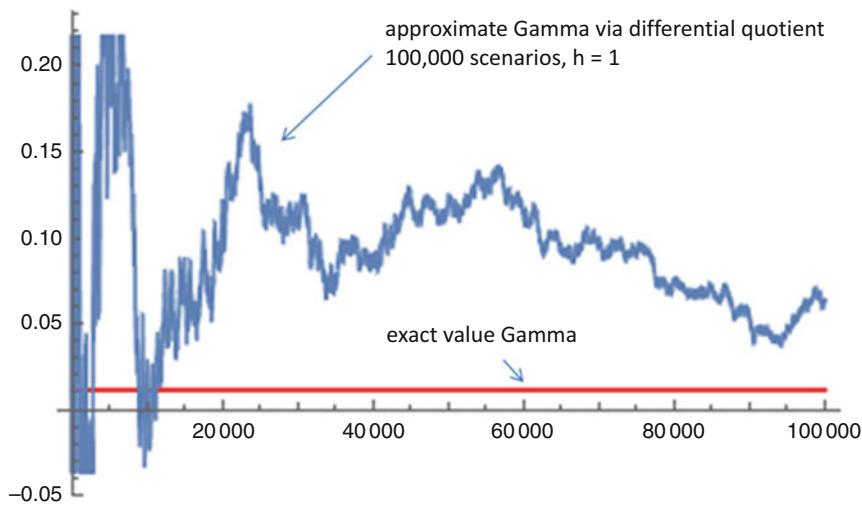


Fig. 2.46 Gamma approximation for a plain vanilla call option using the differential quotient, $h = 1$, 100,000 scenarios

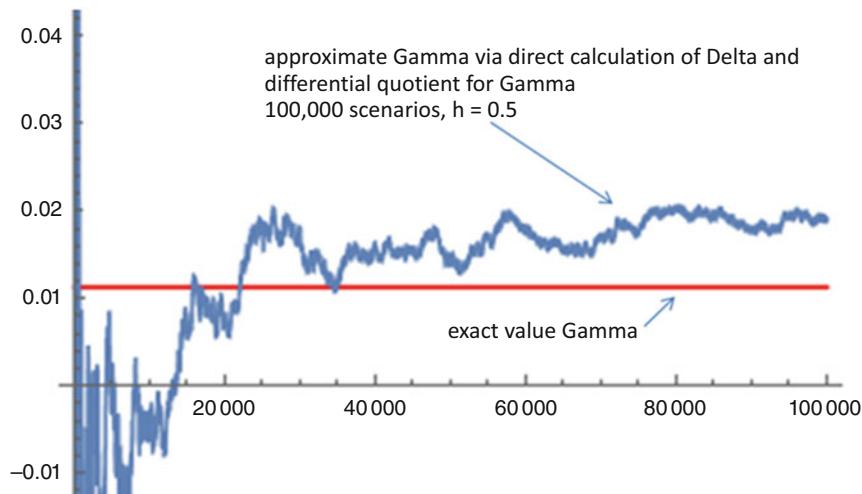


Fig. 2.47 Gamma approximation for a plain vanilla call option by direct MC computation of delta and the differential quotient for gamma, $h = 0.5$, 100,000 scenarios

The simulation shown in Fig. 2.47 suggests an approximate value of 0.0188. While both approximation and convergence behaviour are significantly better than for the simulation using only the differential quotient, the outcomes should still be viewed with caution.

2.21 Estimating Delta and Delta Hedging for Path-Dependent Derivatives (e.g. Geometric Asian Option)

In this chapter, we are going to show how to estimate the delta of a path-dependent option and how delta hedging is performed, using the geometric Asian option as an example. We know that, based on the exact formula for the fair price of a geometric Asian call option that we derived in Sect. 2.13, we could also calculate the delta explicitly. However, we want to exemplify the Monte Carlo method for path-dependent options and will use the geometric Asian option for that in order to have an exact reference value for the Monte Carlo approximation.

As parameters, we will again use our earlier values:

Strike price $K = 90$

Time to expiration $T = 1$

Risk-free interest rate $r = 0.02$

Volatility $\sigma = 0.3$

Initial value S_0 of the underlying asset = 100.

We choose $M = 12$ and use the prices at time points $\frac{T}{12}, 2 \cdot \frac{T}{12}, 3 \cdot \frac{T}{12}, \dots, 12 \cdot \frac{T}{12}$ to calculate the geometric mean.

To determine the reference value, i.e. the delta of the geometric Asian call, let us recall Sect. 2.13. There we showed that the fair value of the geometric call option is given by

$$e^{(\tilde{r}-r)T} \cdot \left(e^{-\tilde{r}T} \cdot E \left(\max \left(0, S(0) \cdot e^{(\tilde{r}-\tilde{\sigma}^2)T + \tilde{\sigma}\sqrt{T}w} - K \right) \right) \right).$$

This last expression $\left(e^{-\tilde{r}T} \cdot E \left(\max \left(0, S(0) \cdot e^{(\tilde{r}-\tilde{\sigma}^2)T + \tilde{\sigma}\sqrt{T}w} - K \right) \right) \right)$ is in fact nothing other than the fair value of a completely ordinary European plain vanilla call option with expiration T and strike K , however with modified parameters $\tilde{\sigma}$ and \tilde{r} instead of the parameters σ and r .

$$\tilde{\sigma} := \sigma \cdot \sqrt{\frac{\left(1 + \frac{1}{M}\right) \cdot \left(2 + \frac{1}{M}\right)}{6}} \text{ and } \tilde{r} := r \cdot \left(\frac{1}{2} + \frac{1}{2M}\right) - \sigma^2 \cdot \left(\frac{1}{12} - \frac{1}{12M^2}\right).$$

The delta of a call option with parameters $\tilde{\sigma}$ and \tilde{r} is—as we know from Volume I Section 4.37—given by $\mathcal{N}(\tilde{d}_1)$ with $\tilde{d}_1 = \frac{\log\left(\frac{S}{K}\right) + \left(\tilde{r} + \frac{\tilde{\sigma}^2}{2}\right)T}{\tilde{\sigma}\sqrt{T}}$.

Therefore, the delta of the geometric Asian call option is $e^{(\tilde{r}-r)T} \cdot \mathcal{N}(\tilde{d}_1)$.

Substituting the parameters of our example yields the reference value 0.7401 for the delta at time $t = 0$.

We now want to estimate the delta using Monte Carlo, and to that end, we choose the approach that consists in differentiating the payoff function. To differentiate the payoff function with respect to S_0 , we can proceed in the same way as in the case of the plain vanilla call option in the previous section. The payoff of the geometric Asian call is, as we know,

$$\max\left(0, \sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} - K\right).$$

Here,

$$\tilde{S}(t_i) = S(0) \cdot e^{i\left(r - \frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})}$$

and thus,

$$\sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} = S(0) \cdot \sqrt[M]{\prod_{i=1}^M e^{i\left(r - \frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})}}.$$

For later purposes, we note that the last root expression—which does not depend on S_0 !—can also be written as

$$\sqrt[M]{\prod_{i=1}^M e^{i\left(r - \frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})}} = \frac{1}{S_0} \cdot \sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)}.$$

Differentiating the payoff with respect to S_0 therefore means differentiating the following expression with respect to S_0 :

$$\max\left(0, S(0) \cdot \sqrt[M]{\prod_{i=1}^M e^{i\left(r - \frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})}} - K\right)$$

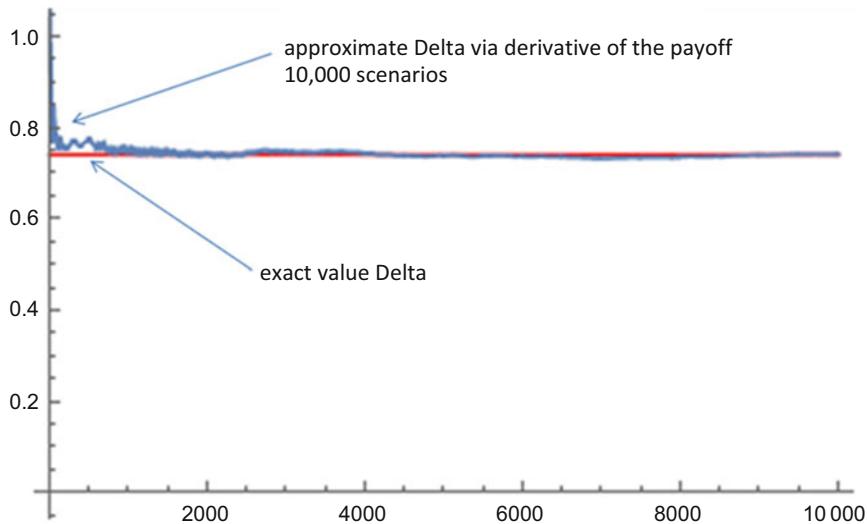


Fig. 2.48 Approximation of the delta of a geometric Asian call option by differentiating the payoff function

Using the chain rule, just like in the case of the plain vanilla call option, we obtain the following result for the derivative with respect to S_0 :

$$\begin{aligned}
 & 1_{x \geq K} \left(S(0) \cdot \sqrt[M]{\prod_{i=1}^M e^{i \left(r - \frac{\sigma^2}{2} \right) \Delta t + \sigma \sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})}} \right) \\
 & \cdot \sqrt[M]{\prod_{i=1}^M e^{i \left(r - \frac{\sigma^2}{2} \right) \Delta t + \sigma \sqrt{\Delta t} \cdot (w_0 + w_1 + \dots + w_{i-1})}} = \\
 & = 1_{x \geq K} \left(\sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)} \right) \cdot \frac{1}{S_0} \cdot \sqrt[M]{\prod_{i=1}^M \tilde{S}(t_i)}
 \end{aligned}$$

So this is the value that we need to simulate with Monte Carlo to approximate the delta of the geometric Asian call. For each scenario i , we will again have to generate an M -dimensional vector $(w_1^{(i)}, w_2^{(i)}, \dots, w_M^{(i)})$ of independent $\mathcal{N}(0, 1)$ -distributed random numbers in order to calculate a possible realization of the above value.

We run the simulation using 10,000 scenarios and again obtain a very compelling convergence behaviour, as shown in the example in Fig. 2.48. The approximate value obtained is 0.7407.

We are now going to take this simulation infrastructure we have created and use it for delta hedging of the geometric Asian call. (For the record, we reiterate here: Using MC would not be necessary in the case of the geometric Asian option, since we have exact formulas for its fair value and delta.)

To dynamically hedge the option over its entire lifetime, we will proceed as follows:

- We first choose the adjustment times for the hedging portfolio.
For example, we could use the time points $\frac{1}{12}, 2 \cdot \frac{1}{12}, 3 \cdot \frac{1}{12}, \dots, 12 \cdot \frac{1}{12}$ at which the prices for calculating the average are determined.
- We determine the fair price of the option at time 0 using MC. We already did this in Sect. 2.17. There, we determined an approximate value of 13.17 USD (rounded to the nearest cent). We assume that we sold the option at this price (+ expenses) and we can now use the amount of 13.17 dollars to execute the hedging strategy. 13.17 is exactly equal to the value of our hedging portfolio at time 0, and we therefore denote it by HP_0 .
- We use MC to calculate the delta Δ_0 of the option at time 0. We just did that above. The approximate value for the delta that we obtained was 0.7407. We therefore buy 0.7407 units of the underlying asset and pay 74.07 for that. To accomplish this, we need to take out a loan at interest rate r (for simplicity, we assume that r remains constant over the option's life) in the amount of $74.07 - 13.17 = 60.90$ USD initially for 1 month.
- After 1 month (time point $\frac{1}{12}$), the price of the underlying asset has changed to the value $S_{\frac{1}{12}}$. Our debt has increased due to the interest that accrued on our loan in that 1 month. Thus, our hedging portfolio has a new value, which we denote by $HP_{\frac{1}{12}}$.
- We can now re-estimate the volatility and proceed with this re-estimation, or we assume (for simplicity in explaining the principle) that the volatility remains relatively unchanged at $\sigma = 0.3$ until expiration.
- Based on the new price, the volatility used, and the now shorter time to expiration of $11 \cdot \frac{1}{12}$, we determine the new delta $\Delta_{\frac{1}{12}}$ at time $\frac{1}{12}$.
- However, to determine the delta (and also the fair price of the option, if desired for comparison or information purposes) at time $\frac{1}{12}$ as well as at the later times $L \cdot \frac{1}{12}$ for $L = 1, 2, 3, \dots, 11$, we need to adjust the procedure! Why? Because in this case, the future payoff depends not only on future (risk-neutral) prices but also on past prices that have already been realized at past sample points. How to proceed in this case is described in detail below.
- The hedging portfolio is rebalanced such that from now until $2 \cdot \frac{1}{12}$, we hold exactly $\Delta_{\frac{1}{12}}$ units of the underlying asset. Consequently, our loan/investment amount needs to be adjusted accordingly.
- At time $\Delta_{\frac{2}{12}}$, we determine the new value $HP_{\frac{2}{12}}$ of our hedging portfolio again.

We perform this procedure until time $T = 1$.

- At time $T = 12 \cdot \frac{1}{12} = 1$, the value of our hedging portfolio is HP_1 . We liquidate the portfolio and receive the amount of HP_1 . We (as the seller of the option) now have the obligation to pay out the payoff that is due. If our hedging was successful, then the proceeds HP_1 from the sale of the hedging portfolio should be large enough to cover that payoff.

On how to correctly determine the fair price of the option or the delta of the option using Monte Carlo simulation at a valuation and sampling time point t_L

By the time t_L , the price has already attained the values $S(t_1), S(t_2), \dots, S(t_L)$ that are used for calculating the mean. From the perspective of time point t_L , the future payoff of the option is given by

$$\max \left(0, \sqrt[M]{\prod_{i=1}^M S(t_i)} - K \right) = \max \left(0, V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M S(t_i)} - K \right)$$

Here, $V(L) := \sqrt[M]{\prod_{i=1}^L S(t_i)}$. $V(L)$ is a value that is fixed and known at time t_L .

If we want to value the option from the perspective of time point t_L , we proceed from the current price $S(t_L)$ and can expect a payoff of $\max \left(0, V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M S(t_i)} - K \right)$.

The value of the option from the perspective of t_L based on Black-Scholes is therefore given by

$$f_{t_L} = e^{-r(T-t_L)} \cdot E \left(\max \left(0, V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} - K \right) \right). \quad (2.1)$$

Here, $\tilde{S}(t_i) = S(t_L) \cdot e^{(i-L) \cdot (r - \frac{\sigma^2}{2}) \cdot \Delta t + \sigma \sqrt{\Delta t} \cdot (w_L + w_{L+1} + \dots + w_{i-1})}$.

(For simplicity, we are again assuming distances of equal length Δt between t_i and t_{i+1} . The general case is completely analogous.)

We denote the expression $e^{(i-L) \cdot (r - \frac{\sigma^2}{2}) \cdot \Delta t + \sigma \sqrt{\Delta t} \cdot (w_L + w_{L+1} + \dots + w_{i-1})}$ (for fixed L) in this case by $\kappa(i)$.

So, for pricing the option at time t_L , we need to use the formula (2.1) and approximate the expected value $E(\max(0, V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} - K))$ by Monte Carlo. To determine the delta of the option in t_L , we will now differentiate the payoff again, yet this time of course with

(continued)

respect to the current price $S(t_L)$. For this purpose, we write the payoff in the following form:

$$\begin{aligned} & \max \left(0, V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} - K \right) = \\ & = \max \left(0, V(L) \cdot S(t_L)^{\frac{M-L}{M}} \cdot \sqrt[M]{\prod_{i=L+1}^M \kappa(i)} - K \right) \end{aligned}$$

Differentiating the payoff with respect to $S(t_L)$ gives us

$$\begin{aligned} & 1_{x \geq \kappa} \left(V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} \right) \cdot V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \kappa(i)} \cdot \frac{M-L}{M} \cdot S(t_L)^{-\frac{L}{M}} = \\ & 1_{x \geq \kappa} \left(V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} \right) \cdot V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} \cdot \frac{M-L}{M} \cdot \frac{1}{S(t_L)}. \end{aligned}$$

Thus, to determine the delta of the option at time t_L , we need to use this formula and approximate the expected value $E \left(1_{x \geq \kappa} \left(V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} \right) \cdot V(L) \cdot \sqrt[M]{\prod_{i=L+1}^M \tilde{S}(t_i)} \cdot \frac{M-L}{M} \cdot \frac{1}{S(t_L)} \right)$ by Monte Carlo.

We now run the program on a specific numerical example of an underlying asset's price movements (see Fig. 2.49).

The prices on the 12 sampling dates in this example are

101.45, 100.78, 97.47, 94.34, 82.84, 94.83, 100.83, 119.35, 126.48, 132.45, 123.75, and 125.14.

The geometric mean of these prices is 107.21, and the resulting **payoff** at time $T = 1$ is therefore $107.21 - 90 = 17.21$.

The requisite deltas at the different points in time and the associated price values, which were all calculated in the same way as the Δ_0 by means of MC simulation according to the program described above, as well as the changes in the components and the total value of the associated hedging portfolio over time are summarized in the following Table 2.2. ("Unit value" refers to the current value of the underlying units held in the last period until now.)

We see that the value of the hedging portfolio remains very close to the fair price of the geometric Asian option at all times. Of particular interest is the fact that

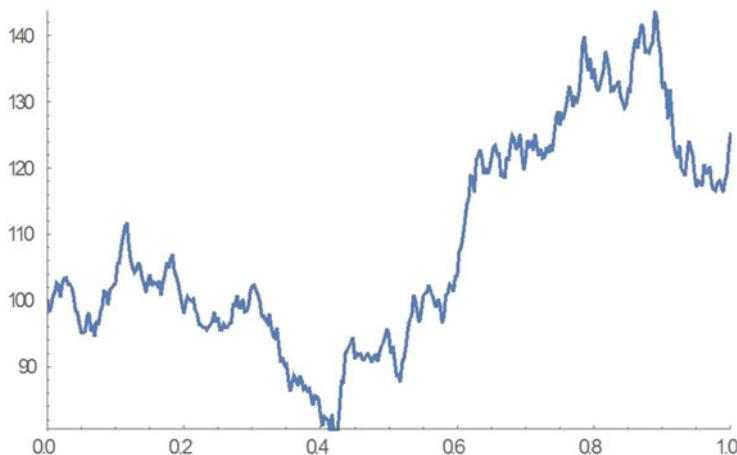


Fig. 2.49 Potential price movement of an underlying asset of a geometric Asian call option

Table 2.2 Deltas, price values, and the total value of the associated hedging portfolio

Time point	Price S	Delta	Option price	Unit value	Cash	Value HP
0	100	0.7407	13.17		13.17	13.17
$\frac{1}{12}$	101.45	0.7163	13.56	75.14	-61.00	14.14
$\frac{2}{12}$	100.78	0.6652	12.18	72.19	-58.62	13.56
$\frac{3}{12}$	97.47	0.5785	9.48	64.84	-53.56	11.27
$\frac{4}{12}$	94.34	0.4919	7.07	54.58	-45.19	9.39
$\frac{5}{12}$	82.84	0.241	1.98	40.75	-37.08	3.67
$\frac{6}{12}$	94.83	0.3903	5.64	22.85	-16.32	6.53
$\frac{7}{12}$	100.83	0.377	7.82	39.35	-30.53	8.82
$\frac{8}{12}$	119.35	0.2912	14.85	44.99	-29.24	15.76
$\frac{9}{12}$	126.48	0.2104	16.91	36.83	-19.03	17.80
$\frac{10}{12}$	132.45	0.1358	18.25	27.87	-8.83	19.04
$\frac{11}{12}$	123.75	0.072	17.08	16.81	1.06	17.86
$\frac{12}{12}$	125.14		17.21	9.01	8.97	17.98

“Option price” and “Value HP” are the most important values and should be compared, therefore in bold

the value of the hedging portfolio at expiration is slightly higher than the payable payoff. Figure 2.50 illustrates how the price curves of the hedging portfolio (blue) and the option (red) compare.

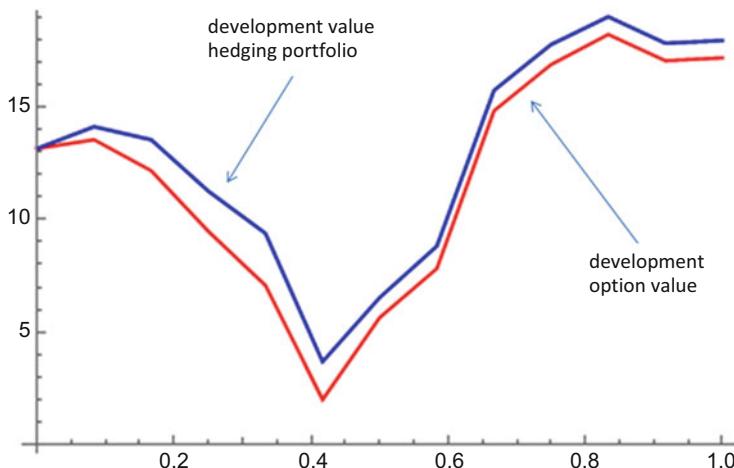


Fig. 2.50 Comparison of price curves for hedging portfolio (blue) and option (red)

2.22 Some Fundamental Remarks on Monte Carlo Methods and on the Convergence of Monte Carlo Methods

The outcome of a Monte Carlo method is always a “stochastic outcome”.

That is because it will always depend on the random numbers used in each simulation, and these random numbers will always differ from case to case.

In Sect. 2.15, we described the simplest version of a Monte Carlo method—determining the average expected value we would get when rolling a dice 10,000 times, for example, and calculating the mean over the 10,000 outcomes. We expect, of course, that each repetition of this experiment of 10,000 dice rolls will have a slightly different outcome, yet always very close to the actual expected average value of rolling a fair dice, that is, very close to 3.5. And in reality, this will practically always be the case. In principle however, we cannot rule out completely that at some point, the number 6 might be rolled in all of the 10,000 rolls, resulting in an estimated expected value of 6 for that particular simulation experiment.

The probability for that to happen is merely $\frac{1}{6^{10000}}$, which is a really small number, of course, but it is not equal to 0!

So, when using the Monte Carlo method—and assuming adequate parameters, i.e. that the number of scenarios used in a simulation is large enough and that (this being another important prerequisite) the random numbers are generated in an adequate way and meet the necessary quality criteria—then we can generally expect results close to the actual value.

But: We can never really be 100% sure of the outcome! The outcome is stochastic, i.e. dependent on random influences.

What then can be said about the quality of the convergence of Monte Carlo methods in general?

The mathematical law of large numbers in combination with the previously mentioned central limit theorem allows us to at least state probabilities that the outcome of a Monte Carlo simulation does not deviate from the actual correct value by more than a certain tolerance ε . (We again assume, of course, that the Monte Carlo method is performed correctly in every sense. We are going to tacitly assume this in all of the following.)

To illustrate this again with the dice experiment:

We can calculate the probability that, when performing 10,000 dice rolls (with a fair dice = correct execution of the MC experiment!), the average value rolled differs from the actual value 3.5 by (e.g.) less than $\frac{1}{20}$. So, what is the probability that the average value for 10,000 rolls is between 3.45 and 3.55?

Finding the answer is really quite simple:

- For an experiment of 10,000 rolls, we have a total of $6^{10,000}$ possible outcomes. An outcome in this context is a possible sequence of 10,000 integers between 1 and 6.

For example: 5, 2, 3, 2, 1, 6, 6, 5, 2, 4, . . . , 2, 6, 1, 3, 3, 4, 3, 3, 1, 5 (10,000 numbers)

or

1, 2, 3, 6, 4, 5, 1, 5, 5, 1, . . . , 2, 4, 3, 6, 6, 6, 1, 6, 4, 5 (10,000 numbers)

or

...

or

6, 6, 6, 6, 6, 6, 6, 6, 6, 6, . . . , 6, 6, 6, 6, 6, 6, 6, 6, 6 (10,000 numbers)

Each of these outcomes in 10,000 rolls is as likely as any other! This is a fact that seems hard to grasp for many. If asked to bet on a possible outcome, one might be willing to bet that one of the first two outcomes listed above will occur. No one, however, would be willing to bet on a 6 being rolled 10,000 times in a row. And yet, as hard as this may be to accept: Rolling a 6 10,000 times in a row is just as likely to occur as rolling the first or second sequence of numbers listed (or, rather, hinted at) above! The probability of any such a specific outcome is therefore $\frac{1}{6^{10,000}}$.

- We are now going to “count” the number A of possible outcomes of 10,000 rolls for which the sum of outcomes is between $10,000 \times 3.45 = 34,500$ and $10,000 \times 3.55 = 35,500$. (The average value of all of these outcomes is then between 3.45 and 3.55.)
- Now, the probability we are looking for is precisely $\frac{A}{6^{10,000}}$, i.e. the number of outcomes that yield the desired result divided by the number of all possible outcomes.
- So, the only thing that we need to do now is determine the variable A . In our example, we could calculate A exactly. It would be much less of an effort, however, if we determined A —not exactly, yet fairly accurately—by approximation. This is what we are going to do in the following, and we do so not just for convenience, but because we can then essentially apply this method

to the general question “What is the probability of Monte Carlo methods being how good?”.

- Now, if we denote by X_i the outcome of the i -th of 10,000 rolls, then $Y := \frac{X_1+X_2+\dots+X_{10,000}}{10,000}$ is the average sum of all the values rolled in the dice experiment. Each of the X_i represents a random variable (with randomly occurring values between 1 and 6). All X_i have the same probability distribution and are independent of one another. (The result of the i -th roll does not affect the outcome of the j -th roll in any way if $i \neq j$.)

The expected value μ of the random variable X_i is 3.5, and the variance σ^2 of the random variable X_i is $\frac{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}{6} = \frac{35}{12} = 2.91666\dots$, i.e. $\sigma = 1.70783\dots$

Thus, the expected value of Y is $\frac{\mu+\mu+\dots+\mu}{10,000} = \mu = 3.5$ and the variance of Y is $\frac{\sigma^2+\sigma^2+\dots+\sigma^2}{10,000^2} = \frac{\sigma^2}{10,000} = 0.00029166\dots$ and the standard deviation of Y is therefore $\frac{\sigma}{\sqrt{10,000}} = \frac{\sigma}{100} = 0.0170783$.

The central limit theorem states that in such a case, the distribution of the random variables Y , as the sum of independent identically distributed random variables, corresponds almost exactly to a normal distribution with expected value $\mu = 3.5$ and with standard deviation $\frac{\sigma}{100} = 0.0170783$ (or variance 0.000291666).

And with that we can calculate the **probability** $W_{[3.45,3.55]}$ for Y to be between **3.45** and **3.55** fairly accurately by

$$W_{[3.45,3.55]} \approx \frac{1}{\sqrt{2\pi} \cdot \frac{\sigma}{100}} \cdot \int_{3.45}^{3.55} e^{-\frac{(x-\mu)^2}{2(\frac{\sigma}{100})^2}} dx = \mathbf{0.996581}.$$

We now rephrase the above question in more general terms:

Let ε be a small positive number. For N rolls of the dice, what is the probability that the average value rolled deviates from the actual expected value of 3.5 by less than ε ?

To answer the question, we perform the exact same steps as above, except that we replace 10,000 by N and 3.45 by $3.5 - \varepsilon$ and 3.55 by $3.5 + \varepsilon$. The answer is now:

$$W_{[3.5-\varepsilon,3.5+\varepsilon]} \approx \frac{1}{\sqrt{2\pi} \cdot \frac{\sigma}{\sqrt{N}}} \cdot \int_{3.5-\varepsilon}^{3.5+\varepsilon} e^{-\frac{(x-3.5)^2}{2(\frac{\sigma}{\sqrt{N}})^2}} dx.$$

We are now going to fine-tune this last value somewhat:

First, we replace the integration variable x by a new integration variable $y := \frac{x-3.5}{\frac{\sigma}{\sqrt{N}}}$, then $\frac{dy}{dx} = \frac{\sqrt{N}}{\sigma}$, thus $dx = dy \cdot \frac{\sigma}{\sqrt{N}}$, and substituting this variable into the integral, we get

$$W_{[3.5-\varepsilon, 3.5+\varepsilon]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-\varepsilon \cdot \frac{\sqrt{N}}{\sigma}}^{\varepsilon \cdot \frac{\sqrt{N}}{\sigma}} e^{-\frac{y^2}{2}} dy.$$

Now we write the error tolerance ε in a slightly different form, namely, $\varepsilon = a \cdot \frac{\sigma}{\sqrt{N}}$, with an arbitrary constant a (e.g. $a = 1, 2, 3, \dots$), and we end up with

$$W_{[3.5-a \cdot \frac{\sigma}{\sqrt{N}}, 3.5+a \cdot \frac{\sigma}{\sqrt{N}}]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-a}^a e^{-\frac{y^2}{2}} dy.$$

For example, for $a = 1, 2, 3, 4$:

$$W_{[3.5 - \frac{1.707}{\sqrt{N}}, 3.5 + \frac{1.707}{\sqrt{N}}]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-1}^1 e^{-\frac{y^2}{2}} dy = 0.68 \dots$$

$$W_{[3.5 - \frac{3.414}{\sqrt{N}}, 3.5 + \frac{3.414}{\sqrt{N}}]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-2}^2 e^{-\frac{y^2}{2}} dy = 0.95 \dots$$

$$W_{[3.5 - \frac{5.121}{\sqrt{N}}, 3.5 + \frac{5.121}{\sqrt{N}}]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-3}^3 e^{-\frac{y^2}{2}} dy = 0.9973 \dots$$

$$W_{[3.5 - \frac{6.828}{\sqrt{N}}, 3.5 + \frac{6.828}{\sqrt{N}}]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-4}^4 e^{-\frac{y^2}{2}} dy = 0.999937 \dots$$

So, if we take, for example, the last estimate as a benchmark for extremely high probability, our answer would be: Rolling a dice N times results in an average value that has extremely low probability (about 0.000063 ...) of deviating from the expected value of 3.5 by more than $\frac{6.828}{\sqrt{N}}$.

Remember: The value of 6.828 is exactly $4 \cdot \sigma$, where σ is the standard deviation of a single roll, i.e. a fixed value independent of how many times N we roll the dice in total. Increasing the number of rolls only changes the distance $\frac{1}{\sqrt{N}}$ from the expected value in the last formula.

If we want to improve the (stochastic) quality of our simulation result by a factor of $\frac{1}{10}$, we have to increase the number N of samples by a factor of 100.

All right, so that was the situation for our simple dice experiment. But what about general and generally much more complex Monte Carlo simulations? Hence: "What is the probability of Monte Carlo methods being how good?"

The answer is: The situation is exactly the same even in general and complex simulation tasks. We conduct the same experiment for a total of N times, all independent of one another, and denote the outcome for each scenario by $X_i; i = 1, 2, 3, \dots, N$. Then we determine the mean $Y := \frac{X_1 + X_2 + \dots + X_N}{N}$ as an approximation of the value that we are attempting to find by means of our simulation. According to the central limit theorem, Y for large N approximately follows a normal distribution. If we denote the expected value of the X_i by μ and the standard deviation of the X_i by σ , then, given the exact same rationale as in the

dice example above (except that here, we use the general μ instead of the value 3.5),

$$W_{\left[\mu-a \cdot \frac{\sigma}{\sqrt{N}}, \mu+a \cdot \frac{\sigma}{\sqrt{N}}\right]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-a}^a e^{-\frac{y^2}{2}} dy$$

and in particular

$$W_{\left[\mu-\frac{\sigma}{\sqrt{N}}, \mu+\frac{\sigma}{\sqrt{N}}\right]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-1}^1 e^{-\frac{y^2}{2}} dy = 0.68\dots$$

$$W_{\left[\mu-\frac{2\sigma}{\sqrt{N}}, \mu+\frac{2\sigma}{\sqrt{N}}\right]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-2}^2 e^{-\frac{y^2}{2}} dy = 0.95\dots$$

$$W_{\left[\mu-\frac{3\sigma}{\sqrt{N}}, \mu+\frac{3\sigma}{\sqrt{N}}\right]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-3}^3 e^{-\frac{y^2}{2}} dy = 0.9973\dots$$

$$W_{\left[\mu-\frac{4\sigma}{\sqrt{N}}, \mu+\frac{4\sigma}{\sqrt{N}}\right]} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-4}^4 e^{-\frac{y^2}{2}} dy = 0.999937\dots$$

Now, if μ is indeed the variable we want to find, then we know that, when using the Monte Carlo method, there is a very high probability for us to obtain, by way of our simulation, an approximate value whose distance from the actual value in its order of magnitude is not greater than $\frac{1}{\sqrt{N}}$.

The complexity of the simulation problem addresses the accuracy of this approximation only in terms of the standard deviation σ of the problem. This σ is independent of the number N of simulations, however.

(We will discuss methods to accelerate the convergence behaviour of Monte Carlo techniques in a subsequent chapter. These methods essentially consist in modifying the original simulation problem such that the standard deviation σ of the problem is reduced, thus improving the (stochastic) error estimates, while the outcome of the problem remains unchanged.)

This fact is often casually phrased as follows:

When using a Monte Carlo method, expect an error of magnitude $\frac{1}{\sqrt{N}}$.

As we have already seen in some of our above examples, the prerequisite “ μ is indeed the variable we want to find ...” is not always satisfied. For example, when MC is used to price a barrier option or a lookback option, then, because of discretization, we no longer simulate the value of the actual option but rather the value of the slightly changed discretized version.

Similarly, when determining the delta of an option by simulating the differential quotient (for a certain fixed increment h), we no longer approximate the actual delta but rather the differential quotient. So, in addition to the simulation error as such, there is also the error caused by the difference between the value we wanted to find (e.g. the value of the original barrier option, delta) and the value that was actually simulated (e.g. the value of the discretized barrier option, differential quotient).

2.23 Some Remarks on Random Numbers

Monte Carlo methods are based on the use of random numbers. The random numbers used in simulations today are generated exclusively by means of implemented random number generators.

Having deterministic (!) algorithms to generate number sequences that exhibit all the attributes of “randomly generated” number sequences (whatever that means!) is an extremely complex task. While until the late 1990s, one would occasionally come across generators—even in professional software packages—which generated random sequences with substantial deficiencies for use in simulations, the random number generators used today are based on extremely reliable and professional algorithms and can largely be used unreservedly.

The development of such algorithms is based mainly on principles of number theory and was perfected by intensive research work in this field, especially in the period from 1970 to 2000. Fundamental and seminal works in this research field were, for example, the three volumes on *The Art of Computer Programming* by Donald Knuth [1] or *Random Number Generation and Quasi-Monte Carlo-Methods* by Harald Niederreiter [2].

As noted above: Users of Monte Carlo methods can generally rely on the random number generators available today and need not really concern themselves with questions as to how and on which bases these random numbers (or, rather, “pseudo-random numbers”, as they merely imitate randomness) are actually generated, which criteria a “good generator” should meet, and how the quality of such a generator is tested and verified.

Nevertheless, for readers interested in getting some insight into the topic, especially with regard to gaining a (rudimentary) understanding of the foundations used in generating random numbers for simulations, we are going to provide some information about this fascinating field at the intersection of pure number theory, probability theory, mathematical algorithmics, mathematical philosophy, and “hard” applied mathematics.

Some preliminary remarks:

Random numbers distributed according to different types of distributions are required in a variety of simulations. For example, in our applications so far, we have always needed random numbers that were normally distributed. In later applications, we are going to simulate beta-distributed and gamma-distributed random variables and will need the random numbers required for that.

The simplest situation is one where we have **uniformly distributed** random numbers in the interval $[0, 1]$. A sequence x_1, x_2, x_3, \dots of numbers in $[0, 1]$ is uniformly distributed if for all a and b with $0 \leq a < b < 1$, the probability that an element of the sequence lies in the interval $[a, b]$ is equal to the length $b - a$ of the interval. In other words, this means

$$\lim_{N \rightarrow \infty} \frac{1}{N} \# \{1 \leq n \leq N \mid x_n \in [a, b]\} = b - a.$$

(Here, $\#$ denotes the number of elements of the subsequent set.)

If we need a sequence distributed according to any other distribution, we can generate it relatively easily from a given uniformly distributed sequence. (In most cases, it is not necessary to take this roundabout route, since practically all established random number generators offer the possibility to generate random numbers for most generally used distributions. Only when an application requires an extremely special and unusual distribution will it become necessary for users to construct the requisite sequence with that particular distribution from a uniformly distributed sequence.)

The general procedure for generating random numbers distributed according to a given distribution from uniformly distributed random numbers is as follows: Let the given distribution be defined on a subinterval $[A, B]$ of real numbers or even on the totality of real numbers, i.e. on the interval $(-\infty, \infty)$. Let it have a density function f that is defined on $[A, B]$. For simplicity, let us assume that f is positive throughout on $[A, B]$. (Otherwise, we could still proceed in the same way as described below, with some obvious adjustments.)

The distribution function F of the distribution is then given for $x \in [A, B]$ by $F(x) = \int_A^x f(u)du$ and is a strictly monotonically increasing function in x with $F(A) = 0$ and $F(B) = 1$. Therefore, the inverse function $G(x) := F^{-1}(x)$ exists as a function of $[0, 1]$ with respect to $[A, B]$ and is also strictly monotonically increasing. In the following, we need this function G to be such that we can actually use it for further calculations. In many cases, it may not be possible to calculate G explicitly from F . In such cases, we need to resort to approximations for G . There are various numerical methods for determining good approximations for inverse functions.

If we already have a “good” **uniformly distributed** random sequence x_1, x_2, x_3, \dots of numbers in $[0, 1]$, then we define a new sequence of numbers y_1, y_2, y_3, \dots in $[A, B]$ in the following way: $y_i := G(x_i)$.

We will show now that this new sequence has the desired distribution in $[A, B]$. To show this, we choose an arbitrary value x in $[A, B]$ and ask: What is the probability that a number from the sequence y_1, y_2, y_3, \dots lies in the subinterval $[A, x]$ of $[A, B]$? We denote this probability by $W(y_i \in [A, x])$, which gives us

$$\begin{aligned} W(y_i \in [A, x]) &= W(G(x_i) \in [A, x]) = W(F^{-1}(x_i) \in [A, x]) = \\ &= W(x_i \in [F(A), F(x)]) = W(x_i \in [0, F(x)]) \end{aligned}$$

Since the x_i are uniformly distributed, the probability that x_i lies within $[0, F(x)]$ is precisely $F(x)$. So, to summarize: The probability that y_i lies in $[A, x]$ is $F(x)$. This again means that F is the distribution function of the sequence y_1, y_2, y_3, \dots and, hence, that the sequence y_1, y_2, y_3, \dots has precisely the distribution we need.

The bottom line: If we want to briefly (!) ask ourselves in the following as to when a sequence is a “good pseudo-random sequence” and how to construct such a sequence, we can limit our approach exclusively to sequences that are uniformly distributed in the interval $[0, 1]$.

As emphasized in the dice example, any outcome of a finite sequence $x_1, x_2, x_3, \dots, x_N$ of numbers in $[0, 1]$ is theoretically possible in purely random experiments. Here and in all of the following, we are going to assume a large number N of values. (Remember: 10,000 rolls of 6 in a row are theoretically possible and equally likely as any other given configuration of 10,000 numbers between 1 and 6!)

So in a random experiment, it is possible, in principle, that an x_i in $[0, 1]$ that lies in the interval $\left[0, \frac{1}{2}\right]$ occurs N times in a row. It is possible but extremely unlikely. However, a good random number generator must also provide the theoretical (however unlikely) possibility that a value in the interval $\left[0, \frac{1}{2}\right]$ occurs N times in a row.

Thus: A good random number generator must also provide for the possibility “of extreme events”.

Much more likely even in random experiments, however, are outcomes $x_1, x_2, x_3,$

\dots, x_N that fall fairly equally into all subintervals $[a, b]$, i.e. for which the difference between $\frac{1}{N} \cdot \#\{1 \leq n \leq N \mid x_n \in [a, b]\}$ and $b - a$ is very small.

Let us consider, for example, the largest error occurring here, i.e. the maximum of all differences

$$D_N = \max_{a,b} \left| \frac{1}{N} \cdot \#\{1 \leq n \leq N \mid x_n \in [a, b]\} - (b - a) \right|,$$

where we take the maximum over all intervals $[a, b]$ with $0 \leq a \leq b \leq 1$. (The more accurate term here would be to take the supremum over all such intervals, as it is not necessarily sure that the maximum is attained.) We denote this supremum by D_N and refer to it as the “discrepancy” of the set of points $x_1, x_2, x_3, \dots, x_N$. In random experiments, there is a very high probability for the occurrence of point sets $x_1, x_2, x_3, \dots, x_N$ with small discrepancy.

This means: Also a good random number generator will have a high probability of producing point sets with small discrepancy.

The discrepancy of a point set can never be smaller than $\frac{1}{N}$. To illustrate this, let us consider, for example, the interval consisting only of the point x_1 . That is, $a = x_1$ and $b = x_1$. In this case, the number of points in the interval is at least equal to 1, and the length of the interval $[a, b]$ is $b - a = 0$. And thus the discrepancy is at least equal to $\frac{1}{N}$.

Now let us look at a highly uniformly distributed sequence, say, the set of N points that corresponds (in any order) to $\frac{0}{N}, \frac{1}{N}, \frac{2}{N}, \frac{3}{N}, \dots, \frac{N-2}{N}, \frac{N-1}{N}$ (see Fig. 2.51).

This set of N equidistant points has, as can easily be seen, a discrepancy of exactly $\frac{1}{N}$ and hence the smallest discrepancy that can possibly occur.

A brief comment: The discrepancy of a set of N points $x_1, x_2, x_3, \dots, x_N$ in $[0, 1]$ that is concentrated entirely on the subinterval $\left[0, \frac{1}{2}\right]$ has discrepancy of at



Fig. 2.51 Set of equidistant points in $[0, 1]$



Fig. 2.52 100 random points in $[0, 1]$ with discrepancy of around $\frac{1}{\sqrt{N}}$

least $\frac{1}{2}$, and a set of points that is concentrated entirely on one single point (say, all x_i have value 0) has discrepancy 1, which is the largest possible discrepancy. Thus, the discrepancy of a set $x_1, x_2, x_3, \dots, x_N$ is always a value between $\frac{1}{N}$ and 1.

But does the set of points shown in Fig. 2.51 really correspond to our idea of a “random set of points”? Definitely not! The distribution of these points is too uniform to be considered typical of a “randomly” generated set of points. Using sophisticated mathematical methods from discrepancy theory, it can be shown that random sequences $x_1, x_2, x_3, \dots, x_N$ have a very high probability of being point sets with discrepancy in the approximate range of $\frac{1}{\sqrt{N}}$. This value obviously tends to 0 as N increases, yet at a much slower rate than the optimal order of convergence $\frac{1}{N}$.

So a good random number generator should have a high probability of producing sets of points $x_1, x_2, x_3, \dots, x_N$ with a discrepancy approximately in the range of $\frac{1}{\sqrt{N}}$.

Figure 2.52 shows a set of 100 points with a discrepancy of almost exactly $\frac{1}{\sqrt{N}} = \frac{1}{10}$.

So, is that the answer to our question: “A random number generator is a good generator if it has a high probability of producing point sets $x_1, x_2, x_3, \dots, x_N$ with a discrepancy approximately in the range of $\frac{1}{\sqrt{N}}$ ”?

No, that is not the answer! It is merely a necessary but far from sufficient prerequisite for a random number generator to be a good generator. Measuring and assessing the discrepancy of the point sets produced by a generator is only **one** test that a random number generator must pass in order to be acceptable. However, there are numerous other tests that must also be passed.

We will address just **one** more of these tests that a generator must pass:

One way to generate point sets $x_1, x_2, x_3, \dots, x_N$ with discrepancy approximately in the range of $\frac{1}{\sqrt{N}}$ is as follows: To generate a given number N of points, we choose a suitable irrational number α (suitable means: The continued fraction representation of α meets certain properties. What exactly that means will not be discussed in more detail here.) And then we do the following: We multiply α sequentially by the numbers $1, 2, 3, \dots, N - 1, N$ and reduce each multiplication result modulo one (i.e. we take only the value after the decimal point). This gives us N numbers between 0 and 1, namely, $\{1.\alpha\}, \{2.\alpha\}, \{3.\alpha\}, \dots, \{N.\alpha\}$. The curly brackets $\{\cdot\}$ denote the reduction modulo one.

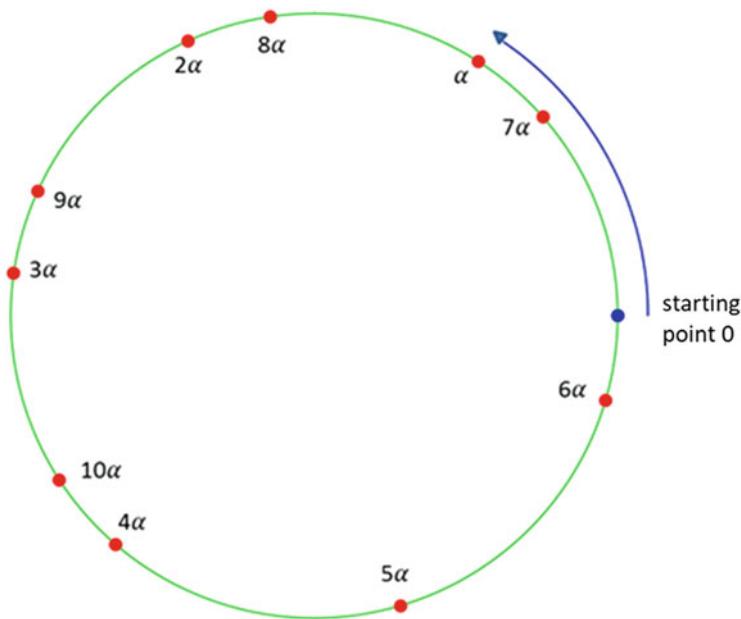


Fig. 2.53 Creating point sets by means of “step lengths” on a circle



Fig. 2.54 Set of points with discrepancy of magnitude $\frac{1}{\sqrt{N}}$

We can also visualize what is happening here as follows: We walk on a circle of circumference 1 with a step length of α . And every step that we place on that circle is a point in our set of points.

After completing the N steps, the circle is opened to an interval of $[0, 1]$, giving us the N points $x_1, x_2, x_3, \dots, x_N$ in $[0, 1]$. (See Fig. 2.53.)

In Fig. 2.54, we see 50 points that were generated in this way using a suitable α such that the discrepancy is approximately $\frac{1}{\sqrt{N}}$.

The discrepancy is of right order, and yet we immediately see: The outcome is not what “we think of as a random set of points”.

Why is that?

One of the most obvious reasons is the following:

If we look at the distances between successive points of this set and measure the length of these distances, we will notice (as we would for every set of points generated in this way) that there are only three different lengths! (See Fig. 2.55.)

This too completely contradicts our idea of a randomly generated set of points. We see that even if a generator (like the one above) produces point sets with suitable discrepancy, it may still be a completely unsuitable random number generator. To be acceptable, a good generator must pass a number of tests in addition to

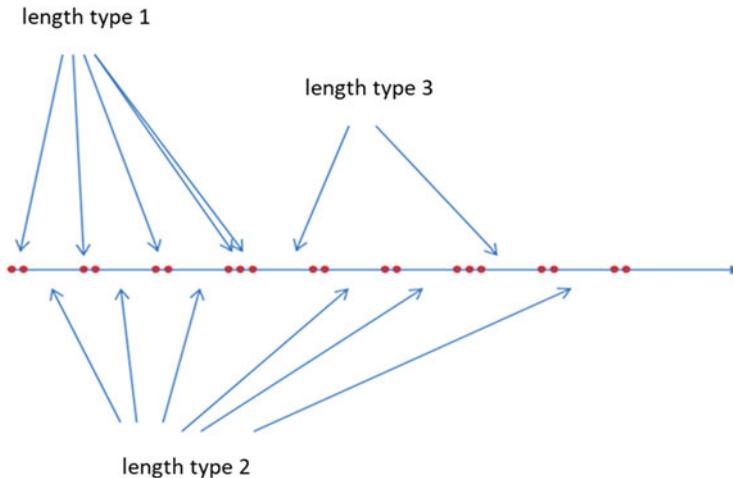


Fig. 2.55 Distances between consecutive points

the discrepancy test, including, as we just saw, a test as to whether distances between neighbouring points in the set have a high probability of being adequately distributed.

So we reiterate: The random number generators used in professional mathematical software today (as opposed to just 20 years ago) are based on highly qualitative and sophisticated algorithms. This means that the random number sequences provided by them can be unreservedly relied upon for use in simulations.

Our remarks here were specifically limited to uniformly distributed random numbers in the **unit interval**. Many applications, however, also require **d -dimensional random vectors** in the d -dimensional unit cube. The same goes for them too: The respective generator is only acceptable once it has passed a series of tests. However: In higher dimensions, these tests can sometimes become extremely complex. Just the complexity of calculating the discrepancy of a set of points increases enormously with dimension.

The process of designing and testing random number generators of the highest quality, implemented primarily in the period from around 1980 to 2010, has been of paramount importance for modern simulation techniques and is based on sophisticated mathematical methods and techniques.

2.24 A Remark on Quasi-Monte Carlo Methods

A key characteristic of Monte Carlo methods is that the error estimates regarding the quality of the simulation outcomes are only stochastic in nature. They are based on the fact that on average, the random number sequences used in those simulations lead to very good approximations (or have a high probability of doing so).

The quasi-Monte Carlo Method (QMC), on the other hand, uses fixed deterministic point sets for simulation instead of random ones. The development of highly specialized QMC methods essentially started in the early 1990s and is still ongoing to this day.

The design of suitable (and often high-dimensional) point sets for complex simulations is an extremely challenging task. The point sets used in such cases must always be **point sets with a discrepancy of essentially smallest possible order**.

QMC methods are not always applicable where MC methods are applicable. When they are applicable, however, they offer the great advantage of fixed deterministic error bounds. Furthermore, in many applications, they provide a much faster convergence of the approximations to the actual value than the pure Monte Carlo method.

We cannot deal with QMC methods in more detail in this book, so in the following, we will give just an idea of what form these deterministic error bounds can take. More information and references for further reading on the topic of QMC in financial mathematics can be found, for example, in [3].

Let us start with a hypothetical example of a simple two-dimensional application: Our task is to estimate the average depth of a lake (we assume for simplicity that it has a square surface). (The average depth can then be used to determine the exact volume of water in the lake). We could take depth measurements at various points of the surface and then calculate the average depth at those measurement points. This would then be an approximation of the actual average depth. Now, how can we choose suitable measuring points? Randomly, like in Fig. 2.56 (which would correspond to a Monte Carlo method), or in a highly regular distribution as in Fig. 2.57? Surprisingly, the choice of a “uniform grid” as shown in Fig. 2.57 would generally give rather poor outcomes (in any case, on average, no better than the pure MC method).

The answer is: It is best to choose (subtly constructed) quasi-random, low-discrepancy points that adequately combine randomness with regularity. An example of such a low-discrepancy two-dimensional point set used in QMC methods can be seen in Fig. 2.58. Now, what deterministic statement can we make regarding the error that we commit if we approximate the actual average depth T of the lake by using an approximate value T_N estimated by means of a QMC point set of N elements?

The answer is: There is a constant V (the so-called variation of the problem) that is determined only by the problem itself (not by the set of points and not by the number of points used), such that we have the following inequality:

$$|T - T_N| < V \cdot D_N.$$

Thus, the **convergence rate** of the approximation error when increasing the number of points for the simulation depends only on the discrepancy of the point set used.

The approximation error is **not only stochastically but definitely always** bounded by $V \cdot D_N$.

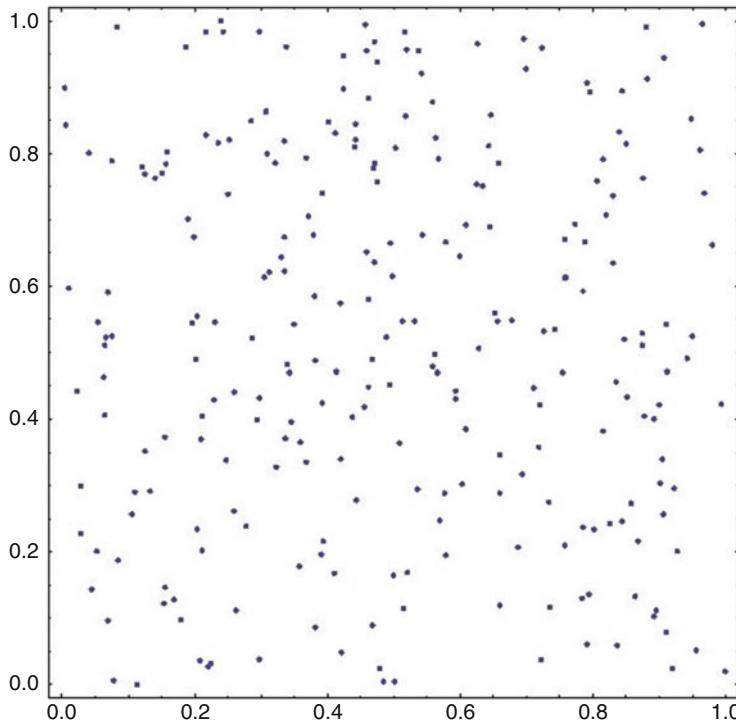


Fig. 2.56 Two-dimensional random points

With this in mind, let us compare the approaches proposed in Figs. 2.56, 2.57, and 2.58: The **Monte Carlo approach** (MC) provides **with high probability** an error in the range of no more than approximately $\sigma \cdot \frac{1}{\sqrt{N}}$ (σ is the standard deviation of the problem).

The discrepancy of a **two-dimensional uniform grid** is $\frac{1}{\sqrt{N}}$. Therefore, using a uniform grid will **certainly** lead to an error of approximately $V \cdot \frac{1}{\sqrt{N}}$ at most (V is the variation of the problem).

The discrepancy of a **low-discrepancy two-dimensional QMC point set**, like the one in Fig. 2.58, is of magnitude $\frac{\log N}{N}$. Therefore, using a **low-discrepancy two-dimensional QMC point set** will **certainly** lead to an error of approximately $V \cdot \frac{\log N}{N}$ at most.

Let us graphically compare the rates of convergence $\frac{1}{\sqrt{N}}$ and $\frac{\log N}{N}$ (see Fig. 2.59):

We see much faster (deterministic!) convergence for QMC with a low-discrepancy point set than for MC (stochastic) or compared with using a uniform grid.

It is not easy in most applications to give a good estimate for the variation V of the problem (nor, by the way, for the standard deviation σ of the problem), but

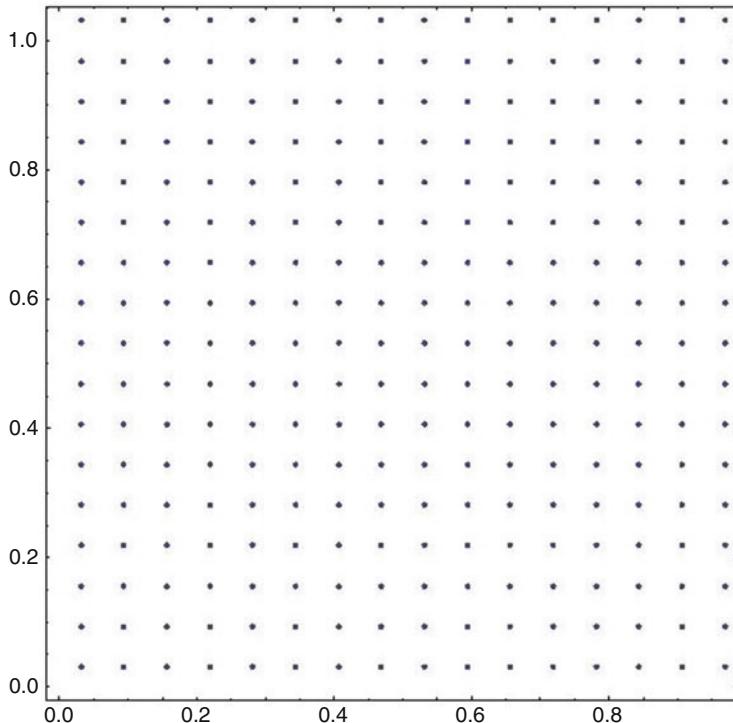


Fig. 2.57 A uniform grid

even if V were much larger than σ , the value $V \cdot \frac{\log N}{N}$ would become much smaller than $\sigma \cdot \frac{1}{\sqrt{N}}$ at a relatively fast rate. Figure 2.60 shows the same comparison as in Fig. 2.59, only now the value $10 \cdot \frac{\log N}{N}$ is compared with $\frac{1}{\sqrt{N}}$ (so, assuming a V of about 10 times the value of σ).

Here, we see a smaller error with the QMC approach than with the MC approach approximately from $N = 8000$ onward.

What about higher dimensions; what is the situation in such cases?

With the MC method (as we have seen), the stochastic error remains unchanged at $\sigma \cdot \frac{1}{\sqrt{N}}$ regardless of the dimension of the problem. Although higher-dimensional problems will generally be expected to have a higher standard deviation σ of the problem, the rate of the (stochastic) convergence, expressed by the term $\frac{1}{\sqrt{N}}$, is not affected by dimension.

The situation changes dramatically if one were to run the simulation using a higher-dimensional uniform grid.

The discrepancy of a uniform grid in dimension s with N points is about $\frac{1}{\sqrt[3]{N}}$.

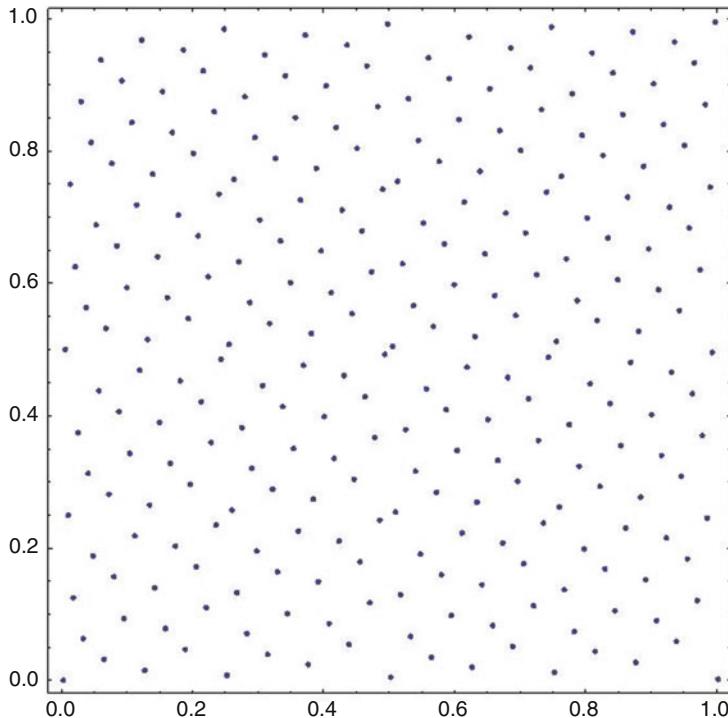


Fig. 2.58 Two-dimensional, low-discrepancy QMC point set

This magnitude $\frac{1}{\sqrt{sN}}$ is larger (i.e. worse) for dimension $s \geq 3$ than the MC convergence rate $\frac{1}{\sqrt{N}}$ and converges to zero extremely slowly for larger dimensions of s . The reason why this choice of points is poorly suited is illustrated in Fig. 2.62. Here, we see the right-side object from Fig. 2.61, i.e. a three-dimensional grid of 1000 points, from a slightly different perspective, and recognize that the points are all strung along relatively few lines. This results in large empty spaces between the points that no simulation point can fall into. These spaces remain critically under-represented in simulations (Fig. 2.62).

For example, if we have a 10-dimensional problem and want the simulation error (leaving the occurring constant V or σ aside) to be no larger than $\frac{1}{10}$ at most, we have to choose a uniform grid with at least 10^{10} (i.e. 10 billion) points.

In contrast, to arrive at a stochastic error of no more than $\frac{1}{10}$ by means of the MC method, 100 random points are already sufficient.

Now, what about low-discrepancy point sets in higher dimensions?

There are two main results to this.

The first is (in simplified terms):

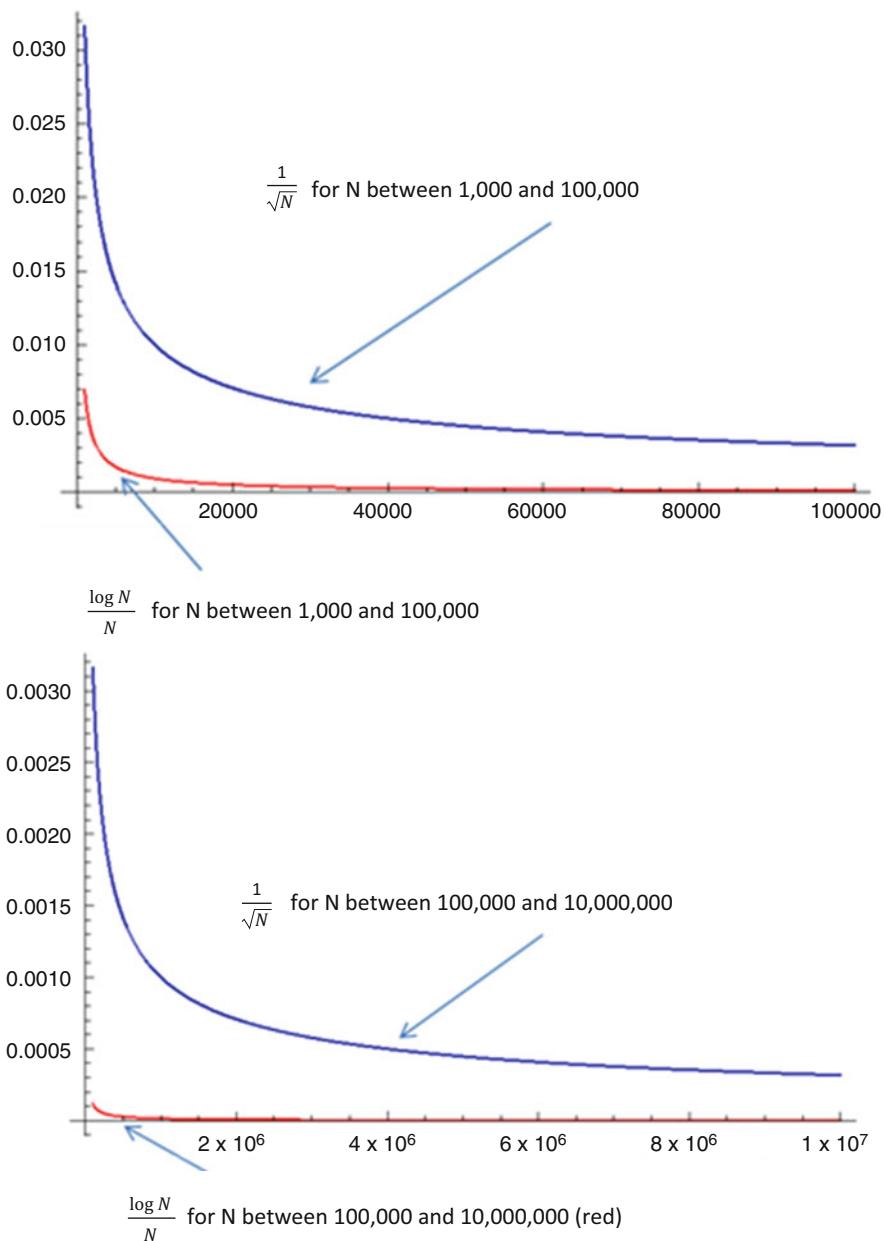


Fig. 2.59 Comparison of convergence rates for Monte Carlo and uniform grid (blue) and quasi-Monte Carlo with low-discrepancy point set (red)

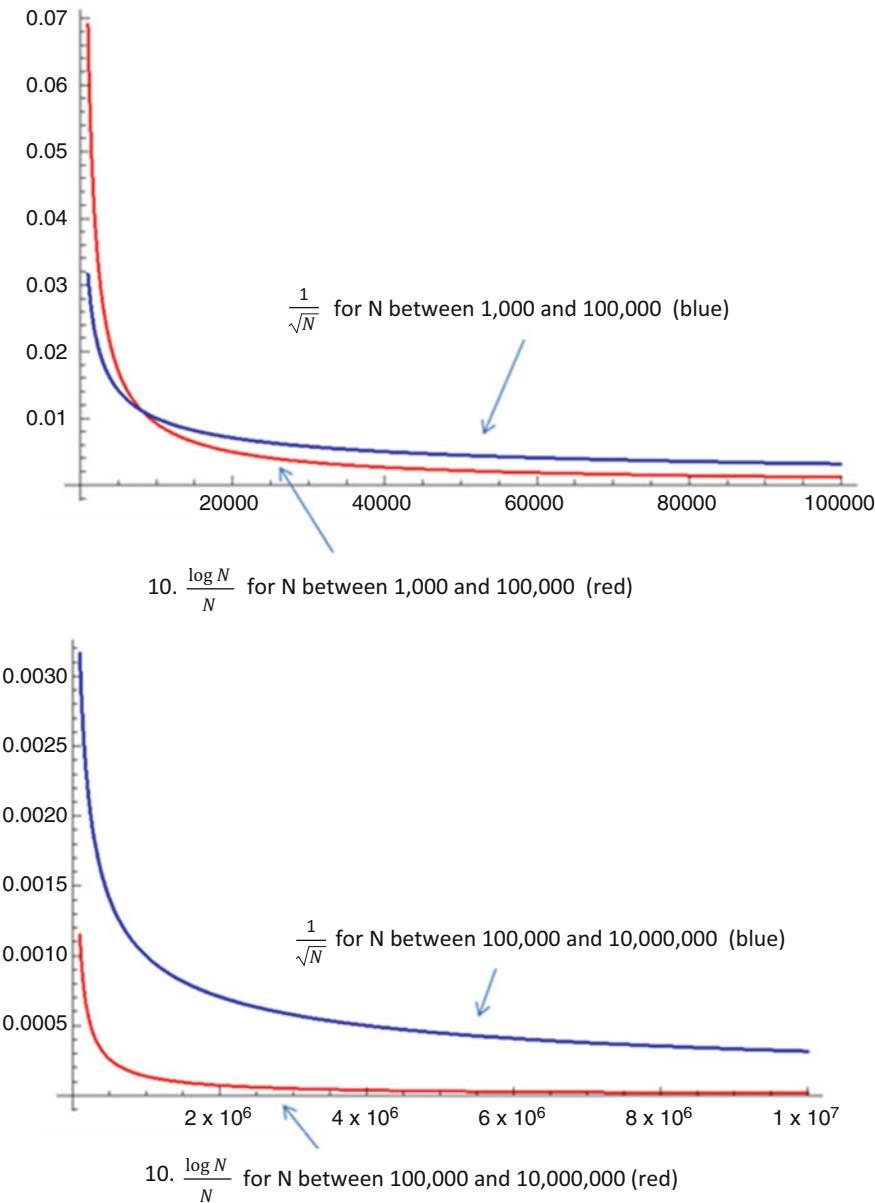


Fig. 2.60 Comparison of convergence rates for Monte Carlo and uniform grid (blue) and quasi-Monte Carlo with low-discrepancy point set (red) for $V \sim 10 \cdot \sigma$

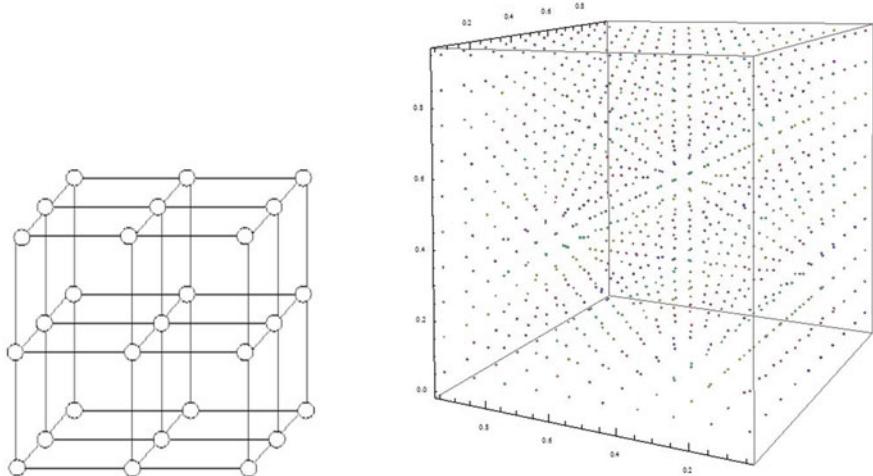


Fig. 2.61 Three-dimensional uniform grid consisting of 27 points (left) or of 1000 points (right)

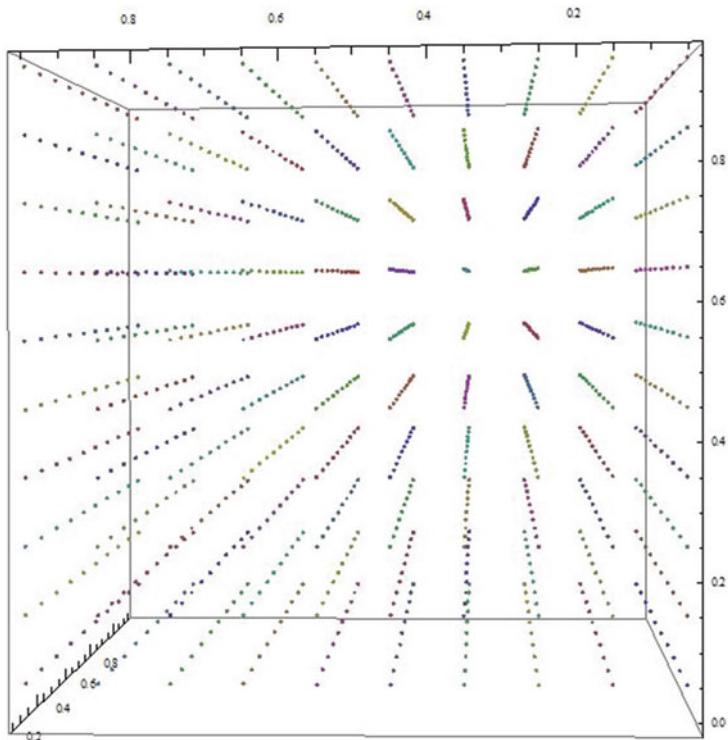


Fig. 2.62 Three-dimensional uniform grid consisting of 1000 points, different perspective

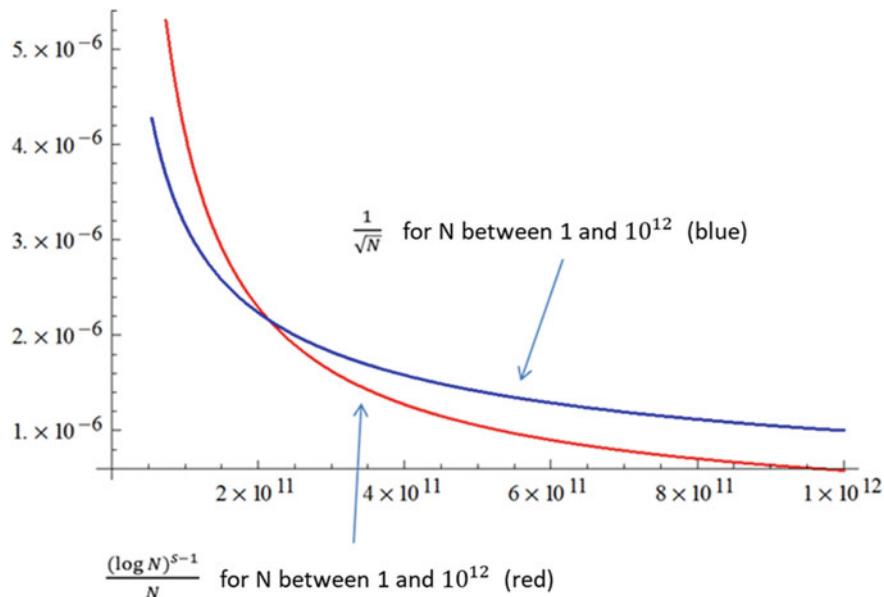


Fig. 2.63 Comparison of convergence rates for Monte Carlo (blue) and quasi-Monte Carlo with low-discrepancy point set (red) in dimension $s = 5$

In each dimension s , we can specify for each N a set consisting of N points, with a discrepancy of no more than $\frac{(\log N)^{s-1}}{N}$ at most.

The second result is:

In each dimension s , there is, for each N , a set consisting of N points with a discrepancy of no more than $3 \cdot \frac{\sqrt{s}}{\sqrt{N}}$ at most.

Let us have a closer look at the first of the two results and the discrepancy $\frac{(\log N)^{s-1}}{N}$ that occurs there:

For each dimension s , the expression $\frac{(\log N)^{s-1}}{N}$ converges to 0 faster than the MC error $\frac{1}{\sqrt{N}}$.

This has been graphically indicated in Fig. 2.63 for the dimension $s = 5$ as an example.

The graph clearly shows that as from a certain number of points N , the red line $\left(\frac{(\log N)^{s-1}}{N}\right)$ takes a course well below the blue line $\frac{1}{\sqrt{N}}$. However, the number of points N from which this is the case is so large that it can no longer be relevant for real applications. This means: For a realistic number of simulation points, the estimate $\frac{(\log N)^{s-1}}{N}$ is significantly larger than $\frac{1}{\sqrt{N}}$ and hence not particularly helpful.

Let us therefore move on to the second result:

The error bound that we obtain from this result is equivalent (apart from the moderate additional constant $3 \cdot \sqrt{s}$) to the stochastic error bound $\frac{1}{\sqrt{N}}$. Thus, using

such point sets, we can expect a simulation error that is **definitely always for any simulation** (not just stochastically with high probability as in the MC method) smaller than a constant times $\frac{1}{\sqrt{N}}!$

While this sounds perfect, there is a major catch: The second result is merely an **existence result**. Meaning: We know that such point sets exist for any s and any N , but so far we have not been able to construct them explicitly. And here we are on the frontline of research in the field of QMC methods. One of the most pressing unsolved research problems in this area is to develop specific construction methods that, for any given s and N , can specify point sets with N points in an s -dimensional unit cube with discrepancy smaller than $3 \cdot \frac{\sqrt{s}}{\sqrt{N}}$.

In fact, it is likely that many of the low-discrepancy point sets from the first result (i.e. those for which it is possible to give a discrepancy estimate of the form $\frac{(\log N)^{s-1}}{N}$) will also satisfy the discrepancy estimate $3 \cdot \frac{\sqrt{s}}{\sqrt{N}}$ (or at least of the form $c \cdot \frac{\sqrt{s}}{\sqrt{N}}$ with possibly a larger but fixed constant c). But so far it hasn't been possible to really prove this. The rationale behind this assumption is that the application of QMC methods using such low-discrepancy point sets most often yield simulation errors that are clearly well below the MC error of $\frac{1}{\sqrt{N}}$. However, so far, this is—as has been said—only an empirical observation and not a definitively proven mathematical fact.

2.25 An Example of Low-Discrepancy QMC Point Sets: The Hammersley Point Sets

We now want to give a vague (!) idea of possible construction methods for low-discrepancy point sets. Most of these methods and their analysis are based on techniques, some being quite sophisticated, from the field of number theory.

Number theory is one of the most fascinating areas of pure mathematics. It deals with highly fundamental questions regarding the properties of certain numbers or number ranges.

What makes the discipline of number theory particularly intriguing is the following: Many of the problems in number theory (some of which are still unresolved to this day) can be formulated in relatively simple terms so that they can be understood even by mathematical laypeople, and yet they are often extremely difficult to solve. Indeed, solving them requires deep mathematical techniques from a wide variety of other areas of mathematics and an exceedingly high degree of creativity. The greatest mathematicians in human history, such as Diophantus, Fermat, Euler, Gauss, Riemann, Hardy,

(continued)

Ramanujan, Erdős, Wiles, Bourgain, Tao, etc., all dealt with problems of number theory at some point.

Probably one of the best known unsolved problems in number theory is the **Goldbach conjecture**. It states that:

Any even number greater than or equal to 4 can be represented as the sum of two prime numbers!

It is indeed true that, for example, $4 = 2+2$, $6 = 3+3$, $8 = 3+5$, $10 = 5+5$, $12 = 5+7$, ..., $100 = 3+97$, ..., $10000 = 59+9941$, ...

But can this really hold for **all** even numbers greater than 4?

Another unsolved and prominent problem in number theory is the **twin primes problem**. Two consecutive odd numbers that are both prime numbers are called pairs of primes or twin primes, for example,

(3, 5), (5, 7), (11, 13), (17, 19), ..., (1,000,037; 1,000,039), etc.

The following is a question that remains unresolved to this day:

Are there infinitely many twin primes?

Another long debated problem that captivated generations of mathematicians and was finally solved in 1994 by British mathematician Andrew Wiles was Fermat's conjecture, which now, after its proof, is generally referred to as **Fermat's last theorem (or Fermat's great theorem)**. The (long suspected, but until then never proven) result of Fermat's theorem is:

There is no integer $n \geq 3$ such that the equation $x^n + y^n = z^n$ has positive integer solutions x, y, z .

(If the exponent is $n = 2$, then such solutions x, y, z do indeed exist. They are the so-called Pythagorean triples, such as $3^2 + 4^2 = 5^2$, or $5^2 + 12^2 = 13^2$, ... and infinitely many other examples.)

Another problem in number theory that is still unsolved and probably extremely difficult to solve is the following:

We know that, for example, the number $\sqrt{2}$ has an infinite and non-periodic decimal expansion. The expansion starts like this $\sqrt{2} = 1.4142135623730950488016887242096980785696718753769480731766797379\dots$ and is still not "understood" to this day. For example, it is still not known whether the digit 0 occurs infinitely often in this decimal representation or not. The same goes for the other nine digits. At least two of these digits must occur infinitely often, that much we know, but which ones they are is a question that nobody has been able to answer so far.

In the following, we will present what is probably the simplest construction of a low-discrepancy point set. We are going to focus on dimension 3 only, since that allows us to graphically illustrate the outcome. The concept of the construction works for any dimension, however.

- To construct N points x_1, x_2, \dots, x_N in dimension 3 with low discrepancy, we write each x_n in the form $x_n = (a_n, b_n, c_n)$.
- For a_n , we simply choose the value $a_n = \frac{n}{N}$.
- To represent b_n , we do the following:

We represent the number n in base 2, that is, in the binary system.

This gives us a representation of the form $n = \beta_0 + \beta_1 \cdot 2 + \beta_2 \cdot 2^2 + \dots + \beta_k \cdot 2^k$, and we then set

$$b_n := \frac{\beta_0}{2} + \frac{\beta_1}{2^2} + \frac{\beta_2}{2^3} + \dots + \frac{\beta_k}{2^{k+1}}$$

- To represent c_n , we do the following:

We represent the number n in base 3.

This gives us a representation of the form $n = \gamma_0 + \gamma_1 \cdot 3 + \gamma_2 \cdot 3^2 + \dots + \gamma_l \cdot 3^l$, and we then set

$$c_n := \frac{\gamma_0}{3} + \frac{\gamma_1}{3^2} + \frac{\gamma_2}{3^3} + \dots + \frac{\gamma_l}{3^{l+1}}$$

You can see the first 1000 points of this point sequence from different perspectives in Fig. 2.64. Hammersley point sets are probably the most easily generated low-discrepancy point sets. Yet they have some disadvantages.

For many—especially high-dimensional—applications, there are much better suited point sets, so-called digital point sets (such as Niederreiter point sets in particular). Such point sets are also provided in, for example, Mathematica (see, e.g. the “*Niederreiter*” given in the following Fig. 2.65 or “*Sobol*” commands).

So, in very many applications, QMC point sets definitely deliver better outcomes than a pure MC method. We will compare the methods in the chapter on multi-asset derivatives. But a word of caution is in order. Using a QMC method is not always possible. For example, if we want to simulate 30 paths of a stock price following a Wiener model with initial value $S_0 = 100$, trend $\mu = 0.1$, volatility $\sigma = 0.3$ over a period of $T = 1$ year, and 100 time steps, then, in earlier applications, we used 30 100-dimensional vectors of standard normally distributed random numbers and typically obtained a result such as the one shown in Fig. 2.66.

However, if we use 30 successive points of a 100-dimensional Hammersley point set for the simulation, then we typically get the following picture 2.67 (here, we did not use the first 30 points of a Hammersley point set but a much later segment; otherwise, the picture would differ even more from that of randomly generated paths).

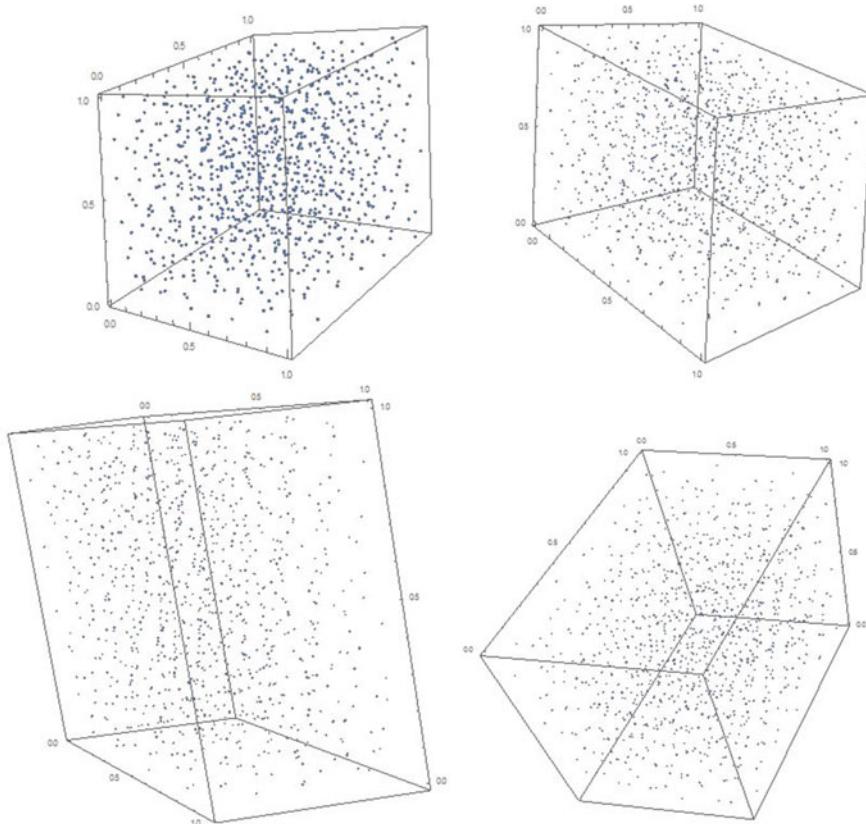


Fig. 2.64 The first 1000 points of a three-dimensional Hammersley point set from different perspectives

There are simply too many dependencies between successive points here. Simulating random paths in this way is therefore not a satisfactory method.

2.26 Variance Reduction Methods for Monte Carlo

Let us come back to the pure Monte Carlo method. We know that the simulation error in a Monte Carlo method is most likely in the range of $\sigma \cdot \frac{1}{\sqrt{N}}$. Various approaches have been developed with a view to accelerating the rate of convergence in a Monte Carlo procedure in certain cases.

We are not going to discuss these approaches in detail here, but we do want to give our readers at least a vague idea of how such methods work. So this is what we will do in this subsection. We will also present a numerical example for only one of these methods.

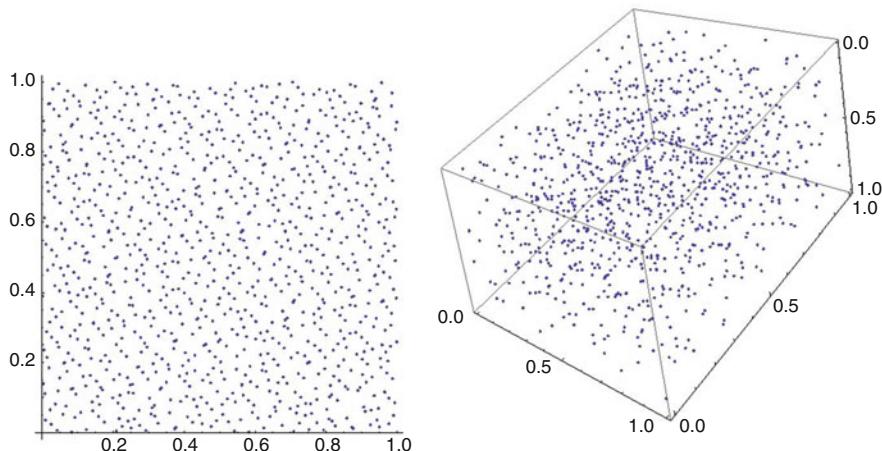


Fig. 2.65 1000 points of a two-dimensional and a three-dimensional Niederreiter point set

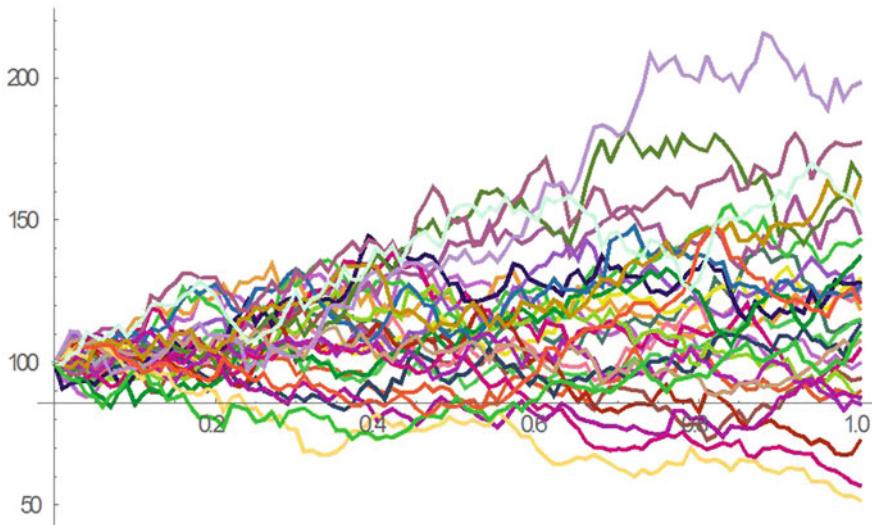


Fig. 2.66 30 simulations of a geometric Brownian motion using random numbers, 100 steps

Probably the two best known “acceleration methods” for Monte Carlo simulation are the “control variates” and the “importance sampling” techniques. In these two approaches, the simulation problem is modified so that, although the outcome of the problem remains unchanged, the standard deviation σ and thus the stochastic simulation error become smaller.

Other acceleration methods (antithetic variates, stratified sampling, etc.) have some impact on the random sequences used in the simulation, which, while essentially preserving the random nature of the sequences, can lead to an improved

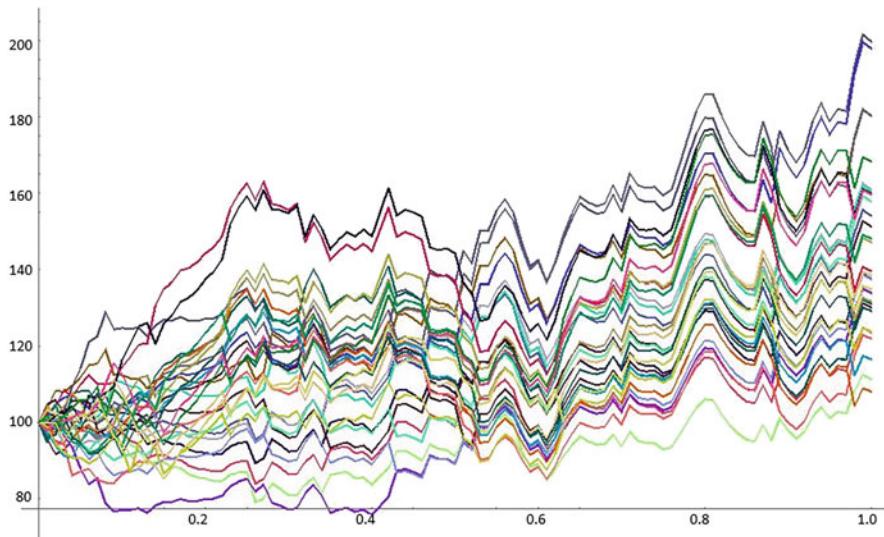


Fig. 2.67 30 simulations of a geometric Brownian motion using a segment from the Hammersley point set, 100 steps

convergence under certain circumstances. We are not going to discuss these methods any further here.

In **importance sampling**, the idea, put simply, is the following:

The objective when using a Monte Carlo method is essentially to find the expected value of a random variable X , where X depends on another random variable w that is distributed according to a distribution given by a density function f . So, what we are looking for is $\int X(w) \cdot f(w) dw$.

The Monte Carlo method tries to approximate this value by generating N different random numbers w_1, w_2, \dots, w_N distributed with respect to the density f , calculating each $X(w_i)$, and using $\frac{1}{N} \cdot \sum_{i=1}^N X(w_i)$ as an approximation for $\int X(w) \cdot f(w) dw$. The stochastic error is given by the magnitude $\sigma(X) \cdot \frac{1}{\sqrt{N}}$.

The idea now is to find another suitable density function g by analysing the problem (i.e. by analysing the random variable X) and rewrite the expected-value problem as follows:

$$\int X(w) \cdot f(w) dw = \int \left(X(w) \cdot \frac{f(w)}{g(w)} \right) \cdot g(w) dw$$

The expression $X(w) \cdot \frac{f(w)}{g(w)}$, which we denote by $Z(w)$, is yet another random variable. Thus, determining $\int X(w) \cdot f(w) dw$ is equivalent to determining $\int Z(w) \cdot g(w) dw$. We can approximate this value by $\frac{1}{N} \cdot \sum_{i=1}^N Z(u_i)$ where u_1, u_2, \dots, u_N

are now N different random numbers u_1, u_2, \dots, u_N distributed with respect to the density g . The stochastic error is now given by the magnitude $\sigma(Z) \cdot \frac{1}{\sqrt{N}}$.

So, if we manage to choose the density function g such that the standard deviation $\sigma(Z)$ of Z is significantly smaller than the standard deviation $\sigma(X)$ of X , then this process too can significantly improve the (stochastic) convergence of the Monte Carlo method.

For example: Let us assume a random variable X that takes only positive values.

Then for g , the ideal choice would be $g(w) = \frac{X(w) \cdot f(w)}{\int X(w) \cdot f(w) dw}$.

With this choice, g would always be positive and

$$\int g(w) dw = \int \frac{X(w) \cdot f(w)}{\int X(w) \cdot f(w) dw} = \frac{1}{\int X(w) \cdot f(w) dw} \cdot \int X(w) \cdot f(w) dw = 1$$

and thus g would indeed be a density function. The random variable Z would then have the form $Z(w) = X(w) \cdot \frac{f(w)}{g(w)} = \int X(w) \cdot f(w) dw$, which means it would be a constant and have standard deviation 0.

But: For us to be able to choose such a g , we would have to know $\int X(w) \cdot f(w) dw$, yet that is precisely the value we are actually looking for. So we can't assume that we will be able to find this g . However, the above observations do shed some light on how to choose such a g in principle. Our analyses should lead us to a density function g that is as proportional to $X(w) \cdot f(w)$ as possible.

The **control variates** method works as follows:

Again we want to estimate the expected value $\int X(w) \cdot f(w) dw$ of a random variable X . Using N random numbers w_1, w_2, \dots, w_N that are distributed with respect to the density f , we generate N realizations $X_i = X(w_i)$ for $i = 1, 2, 3, \dots, N$.

Now we use a second random variable Y to help us with the rest. Let this Y also depend on a random variable w distributed with respect to the density f . We assume that we know the expected value $E(Y) = \int Y(w) \cdot f(w) dw$ explicitly.

Using the same N random numbers w_1, w_2, \dots, w_N that we used to determine the X_i , we now also compute realizations $Y_i = Y(w_i)$ for $i = 1, 2, 3, \dots, N$ of Y .

In addition, we select a constant b , the exact form of which we will look at later.

Finally, we determine $Z_i := X_i - b \cdot (Y_i - E(Y))$ and use $\bar{Z} = \frac{Z_1 + Z_2 + \dots + Z_N}{N}$ as an approximation for $E(X)$.

In fact, $E(Z_i) = E(X_i - b \cdot (Y_i - E(Y))) = E(X_i) - b \cdot E(Y_i - E(Y)) = E(X)$, so we can indeed use \bar{Z} as an estimate for $E(X)$.

But what is the purpose of this approach?

Let us again compare the standard deviation (or variance) of $Z = X - b \cdot (Y - E(Y))$ (the variable we want to estimate) with the original standard deviation of X (using in the following the fact that $\sigma(Y - E(Y)) = \sigma(Y)$).

$\sigma^2(Z) = \sigma^2(X - b \cdot (Y - E(Y))) = \sigma^2(X) - 2 \cdot b \cdot \sigma(X) \cdot \sigma(Y) \cdot \rho_{XY} + b^2 \sigma^2(Y)$, where ρ_{XY} denotes the correlation between the random variables X and Y .

We now assume that b has been determined (at least approximately) such that the last expression $\sigma^2(X) - 2 \cdot b \cdot \sigma(X) \cdot \sigma(Y) \cdot \rho_{XY} + b^2 \cdot \sigma^2(Y)$ attains its minimum value.

We obtain this optimal value for b by differentiating the expression with respect to b and setting the derivative equal to 0, that is, by $2 \cdot b \cdot \sigma^2(Y) - 2 \cdot \sigma(X) \cdot \sigma(Y) \cdot \rho_{XY} = 0 \Leftrightarrow b = \frac{\sigma(X) \cdot \rho_{XY}}{\sigma(Y)}$.

We denote this optimal value for b by \tilde{b} . Plugging this optimal value $\tilde{b} = \frac{\sigma(X) \cdot \rho_{XY}}{\sigma(Y)}$ into the formula $\sigma^2(X) - 2 \cdot b \cdot \sigma(X) \cdot \sigma(Y) \cdot \rho_{XY} + b^2 \cdot \sigma^2(Y)$ for $\sigma^2(X - b \cdot (Y - E(Y)))$, we get $\sigma^2(X - b \cdot (Y - E(Y))) = \sigma^2(X) \cdot (1 - \rho_{XY}^2)$; thus,

$$\sigma(Z) = \sigma(X) \cdot \sqrt{1 - \rho_{XY}^2}.$$

What does that mean? How does that help us? If we understand how to find a random variable Y whose expected value we know and which has either a strongly positive or a strongly negative correlation with X , then we know that the random variable Z will have the same expected value as X , but a much smaller standard deviation than X . A Monte Carlo estimation of the expected value $E(X)$ using the random variable Z should therefore be a substantially faster way to obtain what we are looking for than a direct MC estimation using the random variable X .

However, this still leaves us with the question as to how to find the precise optimal value $\tilde{b} = \frac{\sigma(X) \cdot \rho_{XY}}{\sigma(Y)} = \frac{\text{cov}(X, Y)}{\sigma(Y)^2}$ ($\text{cov}(X, Y)$ denotes the covariance between X and Y). In fact, we will not be able to determine the exact value of \tilde{b} , but what we can do is determine an approximation \bar{b} for \tilde{b} after having calculated X_i and Y_i , as follows:

$$\bar{b} := \frac{\sum_{i=1}^N (x_i - \bar{X}) \cdot (y_i - E(Y))}{\sum_{i=1}^N (y_i - E(Y))^2} \quad \text{where } \bar{X} = \frac{x_1 + x_2 + \dots + x_N}{N}.$$

Note that by choosing a (more or less) suitable random variable Y as a control variate, we will also be able to estimate ρ_{XY} and thus the acceleration factor $\sqrt{1 - \rho_{XY}^2}$, as follows:

$$\overline{\rho_{XY}} := \frac{\sum_{i=1}^N (x_i - \bar{X}) \cdot (y_i - E(Y))}{\sqrt{\sum_{i=1}^N (x_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - E(Y))^2}}$$

In the following subsection, we will demonstrate the control variate method using a specific numerical example.

2.27 Using Monte Carlo with Control Variates to Value an Arithmetic Asian Option

We come back again to the example discussed in 2.17, where we priced a geometric Asian option and an arithmetic Asian option. The parameters were:

Strike $K = 90$

Time to expiration $T = 1$

$r = 0.02$

$\sigma = 0.3$

So the option's time to expiration is 1 year. We compute the average based on 12 values measured at equal intervals (at the end of each trading month).

The underlying asset's initial price $S(0)$ is 100 points.

In 2.17, we priced both options using the Monte Carlo method. But we also had an exact reference value for the geometric Asian version, namely, 13.061.

Now we want to do another valuation of the arithmetic Asian version, this time using the price of the geometric Asian option as a control variate.

We proceed as described in the previous chapter: We denote the number of sample values by N and

- compute N sample values X_i for the price of the arithmetic version
- estimate the approximation \bar{X} from the X_i
- compute N sample values Y_i for the price of the geometric version
- compute the exact reference value $E(Y)$
- estimate the values \bar{b} and ρ_{XY} from the X_i and the Y_i
- compute the N values $Z_i = X_i - b \cdot (Y_i - E(Y))$
- determine the approximate value \bar{Z} (resp. successive approximate values) from the Z_i
- graphically represent the outcomes.

The following figures illustrate the outcomes of 4 simulations that we performed, first with 100, then with 1000, with 10,000, and finally with 100,000 scenarios.

Simulation with $N = 100$ scenarios (displayed in Fig. 2.68):

Approximate value without control variate: 10.38

Approximate value with control variate: 13.43

Approximate value for b : 0.990736

Approximate value for ρ_{XY} : 0.98177

Simulation with $N = 1000$ scenarios (displayed in Fig. 2.69):

Approximate value without control variate: 13.6048

Approximate value with control variate: 13.5887

Approximate value for b : 1.03286

Approximate value for ρ_{XY} : 0.999485

Simulation with $N = 10,000$ scenarios (displayed in Fig. 2.70):

Approximate value without control variate: 13.3813

Approximate value with control variate: 13.5823

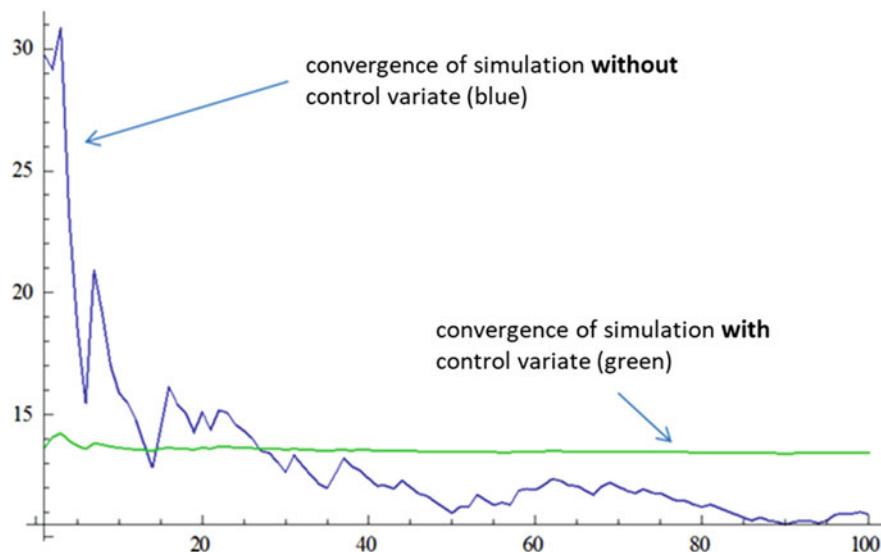


Fig. 2.68 Convergence for a simulation with 100 scenarios

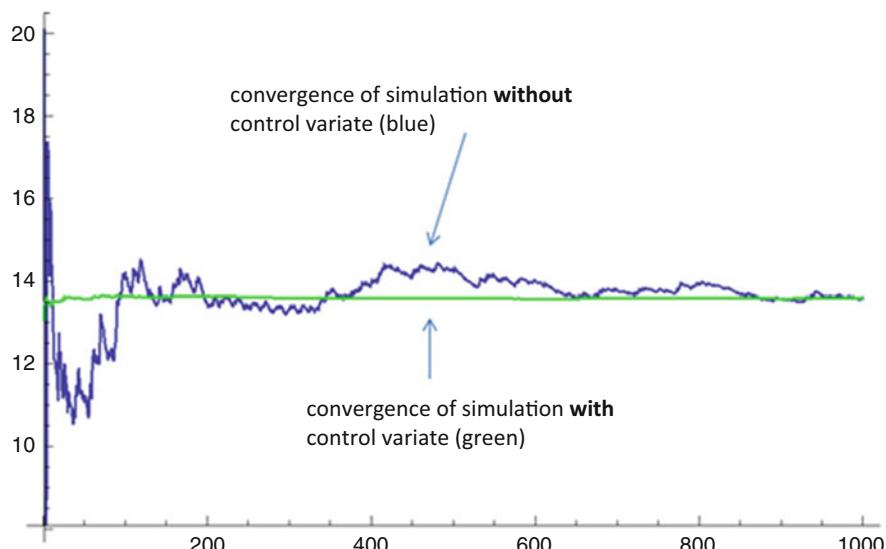


Fig. 2.69 Convergence for a simulation with 1000 scenarios

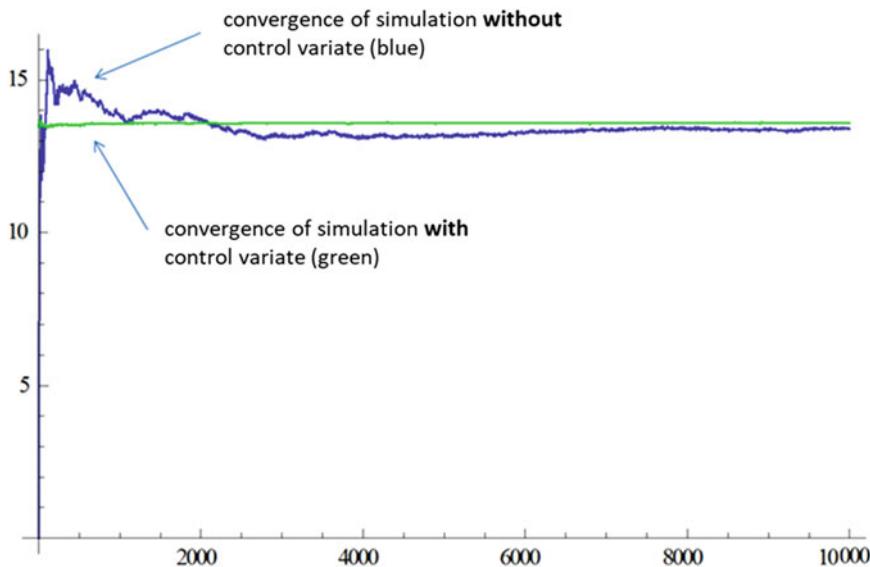


Fig. 2.70 Convergence for a simulation with 10,000 scenarios

Approximate value for b : 1.03411

Approximate value for ρ_{XY} : 0.999304

Simulation with $N = 100,000$ scenarios (displayed in Fig. 2.71):

Approximate value without control variate: 13.5887

Approximate value with control variate: 13.5746

Approximate value for b : 1.03331

Approximate value for ρ_{XY} : 0.999373

Especially in the first two images, we can see a much faster and extremely good convergence of the control variate method.

For comparison, an actual reference value for the arithmetic Asian option was also determined, with a much greater simulation and time effort, and the result was 13.577

2.28 Multi-asset Options

So far, we have dealt exclusively with derivatives whose payoff depended only on **one** underlying asset. Many derivatives on the market are, however, based on several assets. These assets can be of the same type (e.g. a basket of different stocks or currencies), or they can be of different types (e.g. a stock and a currency).

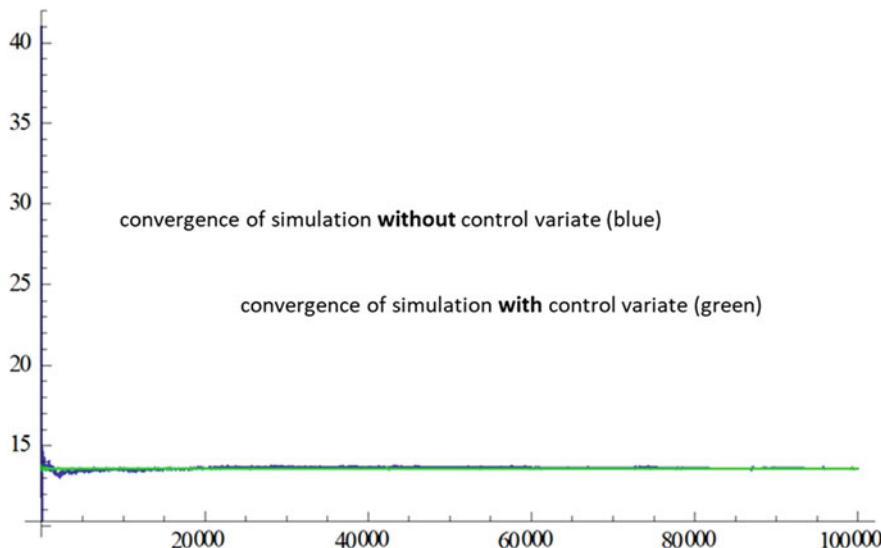


Fig. 2.71 Convergence for a simulation with 100,000 scenarios

In the following, we are going to refer to these options as “multi-asset-options” (other names used include “basket options”, “rainbow options”, and more).

So, if a European derivative is based on various assets, which (or, rather, the prices of which) we are going to denote by S_1, S_2, \dots, S_d , then the payoff at a time T is a function Φ of the prices $S_1(T), S_2(T), \dots, S_d(T)$, that is to say, the payoff is $\Phi(S_1(T), S_2(T), \dots, S_d(T))$.

For all of these options, there are also path-dependent versions, of course.

The payoffs of traded multi-asset options are sometimes of a very simple form, e.g.

$$\begin{aligned} \Phi(S_1(T), S_2(T), \dots, S_d(T)) &= \\ &= \max \left(0, \frac{\frac{S_1(T)}{S_1(0)} + \frac{S_2(T)}{S_2(0)} + \frac{S_3(T)}{S_3(0)} + \dots + \frac{S_d(T)}{S_d(0)}}{d} - K \right) \end{aligned}$$

(call option on the normalized mean stock price)

or

$$\Phi(S_1(T), S_2(T)) = \max(0, b \cdot S_1(T) - S_2(T))$$

(exchange option). The exchange option gives its holder the right to exchange one unit of the stock S_2 for b units of the stock S_1 at time T . (The holder will obviously execute this exchange only if at time T , the b units of the stock S_1 have a greater

value than one unit of the stock S_2 , and the exchange would result in a profit of $b \cdot S_1(T) - S_2(T)$.)

Or

$$\Phi(S_1(T), X(T)) = \max(0, S_1(T) - K) \text{ paid in currency } X.$$

This could be, for example, a call option traded in Europe on a US stock S_1 , with $X(T)$ denoting the price of one USD in EUR at time T .

However, sometimes multi-asset options are of a very complex and creative form. A frequently encountered version is, for example, the following:

$$\begin{aligned} \Phi(S_1(T), S_2(T), \dots, S_d(T)) &= \text{average value at time } T \\ &\quad \text{of the } k \text{ smallest values at time } T \text{ from} \\ &\quad \frac{S_i(T)}{S_i(0)}, i = 1, 2, \dots, d. \end{aligned}$$

(where k is a fixed given value smaller than d).

In the following, we will again assume that all underlying assets can be modelled by a geometric Brownian motion.

So, one approach to this would be to set

$$S_i(T) = S_i(0) \cdot e^{\mu_i T + \sigma_i \sqrt{T} w^{(i)}}$$

for every $i = 1, 2, \dots, d$, where μ_i are the trends and σ_i are the volatilities of the respective underlying asset and $w^{(1)}, w^{(2)}, w^{(3)}, \dots, w^{(d)}$ are independent standard normally distributed random variables.

The problem now, however, is that the products S_1, S_2, \dots, S_d will generally not be independent of each other and the $w^{(1)}, w^{(2)}, w^{(3)}, \dots, w^{(d)}$ should therefore also not be chosen as if they were independent of each other.

As a measure of the dependence, we will again use the correlations ρ_{ij} between the continuous returns of the underlying S_i and the underlying S_j , i.e.—at the level of the model—the correlation between the exponents $\mu_i T + \sigma_i \sqrt{T} w^{(i)}$ and $\mu_j T + \sigma_j \sqrt{T} w^{(j)}$, and this is in fact equal to the correlation between the standard normally distributed random variables $w^{(i)}$ and $w^{(j)}$.

Now, this begs the question: How can the correlation between different financial products be suitably integrated into the Wiener model?

2.29 Modelling Correlated Financial Products in the Wiener Model, Cholesky Decomposition

We refine the question posed at the end of the last subsection:

We are dealing with d financial assets S_1, S_2, \dots, S_d . For each pairing of these assets, S_i and S_j , let ρ_{ij} be the (estimated) correlation between the returns of these two assets.

Each of these financial assets can **in and of itself** be modelled by a Wiener model of the form $S_i(T) = S_i(0) \cdot e^{\mu_i T + \sigma_i \sqrt{T} W^{(i)}}$ with a standard normally distributed random variable $W^{(i)}$. But if we want to simulate the **parallel movements** of all d assets on **the same time interval**, then—to get a correct result—we need to generate the $W^{(1)}, W^{(2)}, W^{(3)}, \dots, W^{(d)}$ such that they reflect precisely this correlation. (By the way, here, we purposely chose the notation with capital letters $W^{(i)}$ for the dependent random variables!)

How can this be accomplished? To answer this, let us rephrase the question somewhat:

Is there any other representation for the $S_i(T)$ that provides the same model as the conventional Wiener model for each S_i but where the correlation is automatically integrated, so that there is no need to additionally impose it as a condition that the random variables have to satisfy?

The answer, of course, is “Yes”. And this is how it could be done:

We will again assume d independent standard normally distributed random variables $w^{(1)}, w^{(2)}, w^{(3)}, \dots, w^{(d)}$.

We construct each $W^{(i)}$ as a linear combination of the $w^{(1)}, w^{(2)}, w^{(3)}, \dots, w^{(d)}$, so, for example,

$$W^{(i)} = a_{i,1} \cdot w^{(1)} + a_{i,2} \cdot w^{(2)} + a_{i,3} \cdot w^{(3)} + \dots + a_{i,d} \cdot w^{(d)} \text{ for } i = 1, 2, 3, \dots, d.$$

As a sum (or linear combination) of independent normally distributed random variables, $W^{(i)}$ is again normally distributed and has expected value 0.

The variance of $W^{(i)}$ is $a_{i,1}^2 + a_{i,2}^2 + \dots + a_{i,d}^2$. So, the first condition that we place on $a_{i,j}$ is:

For each i the equation $a_{i,1}^2 + a_{i,2}^2 + \dots + a_{i,d}^2 = 1$ must hold.

Each of the $W^{(i)}$ is then standard normally distributed.

This representation of course automatically entails a certain dependence between the $W^{(i)}$. (Except, of course, in the case where, e.g. $a_{i,j}$ is 0 whenever $i \neq j$ and is 1 whenever $i = j$. In which case we are again looking at independent assets.)

Now, what does the dependence between a $W^{(i)}$ and a $W^{(j)}$ look like, specifically? More precisely: What is the correlation $\text{cor}(W^{(i)}, W^{(j)})$ between W_i and W_j ?

Well, let us evaluate that:

After all, we have

$$\begin{aligned} \text{cor}(W^{(i)}, W^{(j)}) &= \frac{E((W^{(i)} - E(W^{(i)})) \cdot (W^{(j)} - E(W^{(j)})))}{\sigma(W^{(i)}) \cdot \sigma(W^{(j)})} = \\ &= E(W^{(i)} \cdot W^{(j)}) = \\ &= E((a_{i,1} \cdot w^{(1)} + a_{i,2} \cdot w^{(2)} + \dots + a_{i,d} \cdot w^{(d)})) \cdot \\ &\quad (a_{j,1} \cdot w^{(1)} + a_{j,2} \cdot w^{(2)} + \dots + a_{j,d} \cdot w^{(d)}). \end{aligned}$$

$$\begin{aligned}
& \cdot \left(a_{j,1} \cdot w^{(1)} + a_{j,2} \cdot w^{(2)} + \dots + a_{j,d} \cdot w^{(d)} \right) = \\
& = E \left(\sum_{k=1}^d \sum_{l=1}^d a_{i,k} \cdot a_{j,l} \cdot w^{(k)} \cdot w^{(l)} \right) = \\
& = \sum_{k=1}^d \sum_{l=1}^d a_{i,k} \cdot a_{j,l} \cdot E \left(w^{(k)} \cdot w^{(l)} \right) = \sum_{k=1}^d a_{i,k} \cdot a_{j,k}.
\end{aligned}$$

Yet we wanted that $\text{cor}(W^{(i)}, W^{(j)}) = \rho_{ij}$, i.e. equal to the (estimated) correlation between the returns of the various assets.

For the case that $i = j$, we get, because $\rho_{ii} = 1$, the condition that we already placed above:

$$a_{i,1}^2 + a_{i,2}^2 + \dots + a_{i,d}^2 = 1.$$

For all other i and j , we have the condition

$$\sum_{k=1}^d a_{i,k} \cdot a_{j,k} = \rho_{ij}.$$

So, if we manage to choose the coefficients $a_{i,j}$ such that all of these conditions are satisfied, then we have found an adequate representation for the $W^{(i)}$ and thus a suitable way to model the correlated assets.

Provided that the correlations ρ_{ij} are suitable, we can actually specify coefficients $a_{i,j}$ such that all conditions are satisfied. The technique by which the coefficients can be found is called “Cholesky decomposition”.

The condition that the correlations must satisfy for a solution to be possible is: The correlation matrix $M = (\rho_{ij})_{i,j=1,2,\dots,d}$ is positive definite!

Cholesky decomposition yields a $d \times d$ matrix C for which $C \cdot C^T = M$.

Here, C^T denotes the transposed matrix of C , that is, the matrix C mirrored at the main diagonal.

The entries of such a transposed matrix C^T are precisely coefficients $a_{i,j}$ that satisfy the above equations. Determining such a Cholesky matrix C is best done with a mathematics software (such as Mathematica) or with the software on our website.

To illustrate how this technique works, we are going to do the following:

In a first example, we determine—manually, so to speak—the general representation of two correlated financial assets, and in a second example, we demonstrate the Cholesky decomposition and the representation of three correlated financial assets.

Example 2.15 (Two-Dimensional Cholesky Decomposition) The two-dimensional correlation matrix M of two financial assets S_1 and S_2 has the form $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where ρ is between -1 and 1 and denotes the correlation between the returns of S_1 and S_2 .

If we denote the Cholesky matrix C by $C = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$, then $C^T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and we have to determine a, b, c , and d such that $\begin{pmatrix} a & c \\ b & d \end{pmatrix} \cdot \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

This gives us the equations

$$a^2 + c^2 = 1$$

$$b^2 + d^2 = 1$$

$$a \cdot b + c \cdot d = \rho$$

At first glance, this system of equations seems to tell us that we are dealing with three equations in four unknowns. This could motivate us to choose an arbitrary (and very simple) value for one of the unknowns in the hope that we will be able to solve the remaining system of three equations in three unknowns.

We start the experiment by choosing $a = 1$, which immediately results in $c = 0$.

From the last equation then follows that $b = \rho$, which in turn, due to the second equation, leads to $d = \pm\sqrt{1 - \rho^2}$.

Since one solution is sufficient for our purposes, we choose $d = +\sqrt{1 - \rho^2}$, and our matrix C thus has the form $C = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{pmatrix}$; hence, $C^T = \begin{pmatrix} 1 & \rho \\ 0 & \sqrt{1 - \rho^2} \end{pmatrix}$.

So we choose $a_{1,1} = 1, a_{1,2} = 0, a_{2,1} = \rho, a_{2,2} = \sqrt{1 - \rho^2}$ and therefore finally $W^{(1)} = w^{(1)}$ and $W^{(2)} = \rho \cdot w^{(1)} + \sqrt{1 - \rho^2} \cdot w^{(2)}$.

The representation of the two financial assets S_1 and S_2 is therefore

$$S_1(T) = S_1(0) \cdot e^{\mu_1 T + \sigma_1 \sqrt{T} w^{(1)}}$$

$$S_2(T) = S_2(0) \cdot e^{\mu_2 T + \sigma_2 \sqrt{T} (\rho \cdot w^{(1)} + \sqrt{1 - \rho^2} w^{(2)})}$$

Example 2.16 The task is to find a representation for three financial assets S_1, S_2 , and S_3 of the form $S_i(T) = S_i(0) \cdot e^{\mu_i T + \sigma_i \sqrt{T} W^{(i)}}$, where the dependencies between the returns of these three assets correspond to the following correlation matrix M :

$$M = \begin{pmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 1 & 0.4 \\ 0.5 & 0.4 & 1 \end{pmatrix}$$

Cholesky decomposition with, e.g. Mathematica yields

$$C^T = \begin{pmatrix} 1 & 0.3 & 0.5 \\ 0 & 0.95393 \dots & 0.26207 \dots \\ 0 & 0 & 0.82542 \dots \end{pmatrix} \text{ and thus}$$

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0.3 & 0.95393 \dots & 0 \\ 0.5 & 0.26207 \dots & 0.82542 \dots \end{pmatrix}.$$

The representation of the financial assets $S_1, S_2, \text{ and } S_3$ is therefore

$$S_1(T) = S_1(0) \cdot e^{\mu_1 T + \sigma_1 \sqrt{T} w^{(1)}}$$

$$S_2(T) = S_2(0) \cdot e^{\mu_2 T + \sigma_2 \sqrt{T} \cdot (0.3 \cdot w^{(1)} + 0.95393 \cdot w^{(2)})}$$

$$S_3(T) = S_3(0) \cdot e^{\mu_3 T + \sigma_3 \sqrt{T} \cdot (0.5 \cdot w^{(1)} + 0.26207 \cdot w^{(2)} + 0.82542 \cdot w^{(3)})}$$

By the way, it is always possible to find a matrix C that is a lower triangular matrix. That is, the only random variables to occur in product S_i are $w^{(1)}, w^{(2)}, w^{(3)}, \dots, w^{(i)}$.

2.30 Valuation of Multi-asset Options

A multi-asset option can be priced by means of the following obvious extension of the Black-Scholes formula.

Theorem 2.17 *Let S_1, S_2, \dots, S_d be financial products with the following price representation:*

$$S_i(T) = S_i(0) \cdot e^{\mu_i T + \sigma_i \sqrt{T} \cdot (a_{i,1} \cdot w^{(1)} + a_{i,2} \cdot w^{(2)} + \dots + a_{i,d} \cdot w^{(d)})}$$

for $i = 1, 2, \dots, d$, where $a_{i,1}^2 + a_{i,2}^2 + \dots + a_{i,d}^2 = 1$ for all $i = 1, 2, \dots, d$.
(No payments or costs are incurred from S_i .)

Let D be a European multi-asset derivative on the underlying assets S_1, S_2, \dots, S_d with expiration T and payoff function Φ , i.e. a payoff at time T of the form $\Phi(S_1(T), S_2(T), \dots, S_d(T))$.

Then the fair price $F(t)$ of the derivative D at time t in $[0, T]$ is

$$F(t) = e^{-r(T-t)} \cdot E(\Phi(\tilde{S}_1(T), \tilde{S}_2(T), \dots, \tilde{S}_d(T)))$$

where \tilde{S}_i has dynamics

$$\tilde{S}_i(T) = S_i(t) \cdot e^{\left(r - \frac{\sigma_i^2}{2}\right)(T-t) + \sigma_i \sqrt{T-t} (a_{i,1} \cdot w^{(1)} + a_{i,2} \cdot w^{(2)} + \dots + a_{i,d} \cdot w^{(d)})}$$

with standard normally distributed independent random variables $w^{(1)}, w^{(2)}, \dots, w^{(d)}$. “ E ” in this equation denotes the expected value, and r is the risk-free interest rate $f_{0,T}$.

The result applies analogously to path-dependent multi-asset derivatives with the corresponding adjustments (essentially substituting $S_i(T)$ by $(S_i(t)_{t \in [0,T]})$).

Moreover, the result also applies analogously if some of the S_i incur payments or costs in the form of a continuous return q_i (essentially substituting r by $r - q_i$ in the representation of \tilde{S}_i).

Determining the expected value $E(\Phi(\tilde{S}_1(T), \tilde{S}_2(T), \dots, \tilde{S}_d(T)))$ is now done for the d random variables and will only be possible explicitly in very special cases.

In **integral form**, the **expected value** can be represented as a d -fold integral as follows:

$$\begin{aligned} E(\Phi(\tilde{S}_1(T), \tilde{S}_2(T), \dots, \tilde{S}_d(T))) &= \\ &= \frac{1}{\sqrt{2\pi}^d} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Phi(\tilde{S}_1(T), \tilde{S}_2(T), \dots, \tilde{S}_d(T)) \cdot \\ &\quad \cdot e^{-\frac{(w^{(1)})^2 + (w^{(2)})^2 + \dots + (w^{(d)})^2}{2}} dw^{(1)} dw^{(2)} \dots dw^{(d)}. \end{aligned}$$

Here, it is then necessary to substitute the expressions $\tilde{S}_i(T)$ in each case by $S_i(t) \cdot e^{\left(r - \frac{\sigma_i^2}{2}\right)(T-t) + \sigma_i \sqrt{T-t}(a_{i,1} \cdot w^{(1)} + a_{i,2} \cdot w^{(2)} + \dots + a_{i,d} \cdot w^{(d)})}$.

For Monte Carlo simulation of the expected value

$$E(\Phi(\tilde{S}_1(T), \tilde{S}_2(T), \dots, \tilde{S}_d(T)))$$

using N scenarios, we need to generate (in the standard, i.e. **non-path-dependent** case) N -independent vectors $(w_j^{(1)}, w_j^{(2)}, w_j^{(3)}, \dots, w_j^{(d)})$ for $j = 1, 2, \dots, N$ of independent standard normally distributed random numbers. (It is sufficient to simply generate $N \cdot d$ independent standard normally distributed random numbers and group them into N vectors with d entries each.)

For each such scenario $(w_j^{(1)}, w_j^{(2)}, w_j^{(3)}, \dots, w_j^{(d)})$, we compute for each $i = 1, 2, \dots, d$ a path for the prices of the (risk-neutral) underlying assets using

$$\tilde{S}_i(T) = S_i(t) \cdot e^{\left(r - \frac{\sigma_i^2}{2}\right)(T-t) + \sigma_i \sqrt{T-t}(a_{i,1} \cdot w_j^{(1)} + a_{i,2} \cdot w_j^{(2)} + \dots + a_{i,d} \cdot w_j^{(d)})}$$

and, using these price values, determine a scenario value for a payoff

$$X_j = \Phi(\tilde{S}_1(T), \tilde{S}_2(T), \dots, \tilde{S}_d(T))$$

The approximate value for the expected value is then again

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N}.$$

For Monte Carlo simulation of the expected value

$$E \left(\Phi \left((\tilde{S}_1(t))_{t \in [0, T]}, \dots, (\tilde{S}_d(t))_{t \in [0, T]} \right) \right)$$

using N scenarios, the first step in the case of a **path-dependent** derivative N is to generate a discretization of the time interval $[t, T]$ into M time intervals (of ideally equal length). The next step is then that for each scenario j ($j = 1, 2, \dots, N$), independent d -dimensional vectors are to be generated, consisting of entries which are all M -dimensional vectors.

$$\begin{aligned} & \left(\left(w_j^{(1)}(1), w_j^{(1)}(2), \dots, w_j^{(1)}(M) \right), \right. \\ & \quad \left(w_j^{(2)}(1), w_j^{(2)}(2), \dots, w_j^{(2)}(M) \right), \\ & \quad \left. \left(w_j^{(d)}(1), w_j^{(d)}(2), \dots, w_j^{(d)}(M) \right) \right) \end{aligned}$$

Each such M -dimensional vector $(w_j^{(k)}(1), w_j^{(k)}(2), \dots, w_j^{(k)}(M))$ consists of the random numbers used to generate a path in the j -th scenario for the k -th product.

(It is sufficient to simply generate $N \cdot d \cdot M$ independent standard normally distributed random numbers and group them into N vectors with d entries of M elements each.)

For each such scenario

$$\begin{aligned} & \left(\left(w_j^{(1)}(1), w_j^{(1)}(2), \dots, w_j^{(1)}(M) \right), \right. \\ & \quad \left(w_j^{(2)}(1), w_j^{(2)}(2), \dots, w_j^{(2)}(M) \right), \dots, \\ & \quad \left. \left(w_j^{(d)}(1), w_j^{(d)}(2), \dots, w_j^{(d)}(M) \right) \right) \end{aligned}$$

we compute for each $i = 1, 2, \dots, d$ a value for the prices of the (risk-neutral) underlying assets at time T and use these paths to determine a scenario value for a payoff

$$X_j = \Phi \left((\tilde{S}_1(t))_{t \in [0, T]}, (\tilde{S}_2(t))_{t \in [0, T]}, \dots, (\tilde{S}_d(t))_{t \in [0, T]} \right)$$

The approximate value for the expected value is then again

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N}.$$

The complexity of the simulation problem can already increase dramatically here. MC valuation of a path-dependent multi-asset derivative on $d = 30$ underlying assets and with a discretization into $M = 100$ time steps already requires 3000

random numbers per scenario. Thus, to run 10,000 scenarios, we would already have to generate 30 million random numbers.

For **multi-asset derivatives**, the use of **quasi-Monte Carlo methods** (QMC) is well worth considering. Especially a non-path-dependent multi-asset derivative lends itself to working with N -element d -dimensional QMC point sets.

In the next subsection, we will perform the valuation of a multi-asset option both in MC and for comparison in QMC.

Before performing a multi-asset option valuation as described above, it is a good idea in each case to first ask yourself a very pertinent question, as this could—if applicable—substantially reduce the complexity of the problem:

The question to be asked is:

Can the payoff of the multi-asset option $\Phi(S_1(T), S_2(T), \dots, S_d(T))$ be represented as a function Ψ of another function Y of $S_1(T), S_2(T), \dots, S_d(T)$, where Y can be modelled by a one-dimensional Wiener model?

To clarify what we mean by this, look at this example:

A popular version of a multi-asset option is a call option on the average normalized value of a basket of d stocks at a future point in time T , that is, an option with the payoff

$$\begin{aligned}\Phi(S_1(T), S_2(T), \dots, S_d(T)) &= \max\left(0, \frac{\frac{S_1(T)}{S_1(0)} + \frac{S_2(T)}{S_2(0)} + \frac{S_3(T)}{S_3(0)} + \dots + \frac{S_d(T)}{S_d(0)}}{d} - K\right) \\ &:= \Psi(Y(S_1(T), S_2(T), \dots, S_d(T)))\end{aligned}$$

with $\Psi(x) := \max(0, x - K)$ and $Y(S_1(T), S_2(T), \dots, S_d(T)) =$

$$\frac{\frac{S_1(T)}{S_1(0)} + \frac{S_2(T)}{S_2(0)} + \frac{S_3(T)}{S_3(0)} + \frac{S_d(T)}{S_d(0)}}{d}.$$

However, the value $\frac{\frac{S_1(T)}{S_1(0)} + \frac{S_2(T)}{S_2(0)} + \frac{S_3(T)}{S_3(0)} + \frac{S_d(T)}{S_d(0)}}{d}$ can in fact be viewed as an index calculated from the stocks S_1, S_2, \dots, S_d . It is therefore possible to represent the price movement of $Y(S_1(T), S_2(T), \dots, S_d(T)) = \frac{\frac{S_1(T)}{S_1(0)} + \frac{S_2(T)}{S_2(0)} + \frac{S_3(T)}{S_3(0)} + \frac{S_d(T)}{S_d(0)}}{d}$ by a one-dimensional Wiener model. This means that we are then only dealing with a one-dimensional problem.

The analogous procedure is also applicable, for example, to an Asian option on the mean of a basket of stocks.

2.31 Example of Pricing a Multi-asset Option with MC and with QMC

Let us consider the following example of a multi-asset option:

The option is valid from time 0 to time $T = 1$ year and is entered on a basket of 10 stocks S_1, S_2, \dots, S_{10} . The price of the stocks is normalized such that at time 0, all of them have the normalized price 100. In the following, we always refer exclusively to these normalized stock prices.

Table 2.3 Volatility and trend for the different stocks

	S_1	S_2	S_3	S_4	S_5
σ	0.1	0.15	0.2	0.25	0.3
$r - \frac{\sigma^2}{2}$	0.015	0.00875	0	-0.01125	-0.025
	S_6	S_7	S_8	S_9	S_{10}
σ	0.35	0.4	0.45	0.5	0.55
$r - \frac{\sigma^2}{2}$	-0.04125	-0.06	-0.08125	-1.105	-0.13125

Let the risk-free interest rate for the time range $[0, T]$ be constant with $r = 0.02$.

For each of the ten stock prices, we assume a Wiener model. Since the trend of these models is irrelevant to the valuation process, we do not estimate and specify the trend.

We denote the volatility of S_i by σ_i , the numerical values of which are given below. To value the option, we only need the risk-neutral version of the Wiener models for the S_i . We therefore use $S_i(t)$ to denote the risk-neutral price movements from the outset and won't need the extra notation $\tilde{S}_i(t)$.

The models for the price movements $S_i(t)$ therefore each have trend $r - \frac{\sigma_i^2}{2}$.

We assume that the stocks are numbered consecutively with increasing volatility and choose the corresponding parameters as given in the following Table 2.3:

Furthermore, we assume the following correlation matrix M for the returns of each stock:

$$M = \begin{pmatrix} 1. & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ 0.9 & 1. & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 \\ 0.8 & 0.9 & 1. & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 \\ 0.7 & 0.8 & 0.9 & 1. & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 \\ 0.6 & 0.7 & 0.8 & 0.9 & 1. & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 \\ 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1. & 0.9 & 0.8 & 0.7 & 0.6 \\ 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1. & 0.9 & 0.8 & 0.7 \\ 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1. & 0.9 & 0.8 \\ 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1. & 0.9 \\ 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1. \end{pmatrix}$$

Cholesky decomposition using Mathematica gives us

$$C = \begin{pmatrix} 1. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\ 0.9 & 0.43589 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\ 0.8 & 0.412948 & 0.435286 & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\ 0.7 & 0.390007 & 0.411103 & 0.434613 & 0. & 0. & 0. & 0. & 0. & 0. \\ 0.6 & 0.367065 & 0.386921 & 0.409048 & 0.433861 & 0. & 0. & 0. & 0. & 0. \\ 0.5 & 0.344124 & 0.362738 & 0.383482 & 0.406745 & 0.433013 & 0. & 0. & 0. & 0. \\ 0.4 & 0.321182 & 0.338556 & 0.357917 & 0.379628 & 0.404145 & 0.432049 & 0. & 0. & 0. \\ 0.3 & 0.29824 & 0.314373 & 0.332351 & 0.352512 & 0.375278 & 0.401189 & 0.430946 & 0. & 0. \\ 0.2 & 0.275299 & 0.290191 & 0.306786 & 0.325396 & 0.34641 & 0.370328 & 0.397796 & 0.429669 & 0. \\ 0.1 & 0.252357 & 0.266008 & 0.28122 & 0.298279 & 0.317543 & 0.339467 & 0.364646 & 0.393863 & 0.428174 \end{pmatrix}$$

and thus the coefficients $a_{i,j}$ in the representations

$$S_i(T) = S_i(0) \cdot e^{\left(r - \frac{\sigma_i^2}{2}\right)T + \sigma_i \sqrt{T}(a_{i,1} \cdot w^{(1)} + a_{i,2} \cdot w^{(2)} + \dots + a_{i,10} \cdot w^{(10)})}$$

of each of the stock prices are given for $i = 1, 2, \dots, 10$. $(a_{i,1}, a_{i,2}, \dots, a_{i,10})$ is precisely the i -th row in this matrix C^T .

Let the payoff of the option that we want to value be given by

$$\Phi(S_1(T), S_2(T), \dots, S_{10}(T)) := \max(0, \max(S_1(T), S_2(T), \dots, S_{10}(T)) - 100)$$

What is relevant for the payout amount at time T is the maximum value of the 10 stocks at time T (i.e. in 1 year).

In Fig. 2.72, we see a possible scenario for the price movements of the ten stocks over the course of a year, including the maximum and the resulting payoff. So, what needs to be determined now is the expected value $E(\max(0, \max(S_1(T), S_2(T), \dots, S_{10}(T)) - 100))$. We first determined the expected value with 10,000 scenarios using the pure Monte Carlo method (blue curve in the following graphs), then using QMC with Hammersley point sets (red curve in the following graphs), and finally using QMC with Niederreiter point sets (green curve in the following graphs).

In the MC method, we generated 10,000 simulations for 10 standard normally distributed random numbers $(w_j^{(1)}, w_j^{(2)}, w_j^{(3)}, \dots, w_j^{(10)})$ for $j = 1, 2, \dots, 10,000$, which were then used to generate the respective scenarios for possible values of $S_1(T), S_2(T), \dots, S_{10}(T)$.

In the QMC method with a Hammersley point set, we generated a 10-dimensional Hammersley point set consisting of 10,000 points $(w_j^{(1)}, w_j^{(2)}, w_j^{(3)}, \dots, w_j^{(10)})$; $j = 1, 2, \dots, 10,000$. Using the inversion method, this uniformly distributed (!) set of points was transferred to a standard normally distributed set of points, which was then used to generate the respective scenarios for possible values of $S_1(T), S_2(T), \dots, S_{10}(T)$.

In the QMC method with a Niederreiter point set, we generated a 10-dimensional Niederreiter point set consisting of 10,000 points $(w_j^{(1)}, w_j^{(2)}, w_j^{(3)}, \dots, w_j^{(10)})$; $j = 1, 2, \dots, 10,000$. Using the inversion method, this uniformly distributed (!) set of points was transferred to a standard normally distributed set of points, which was then used to generate the respective scenarios for possible values of $S_1(T), S_2(T), \dots, S_{10}(T)$.

Both Hammersley and Niederreiter point sets have been implemented in Mathematica (and of course in the software on our website).

Running the Mathematica implementation of the Niederreiter point sets multiple times yields slightly different outcomes, as each time, Mathematica will generate different segments of a Niederreiter sequence.

We have illustrated the outcomes for two different simulation runs in Figs. 2.73 and 2.74.

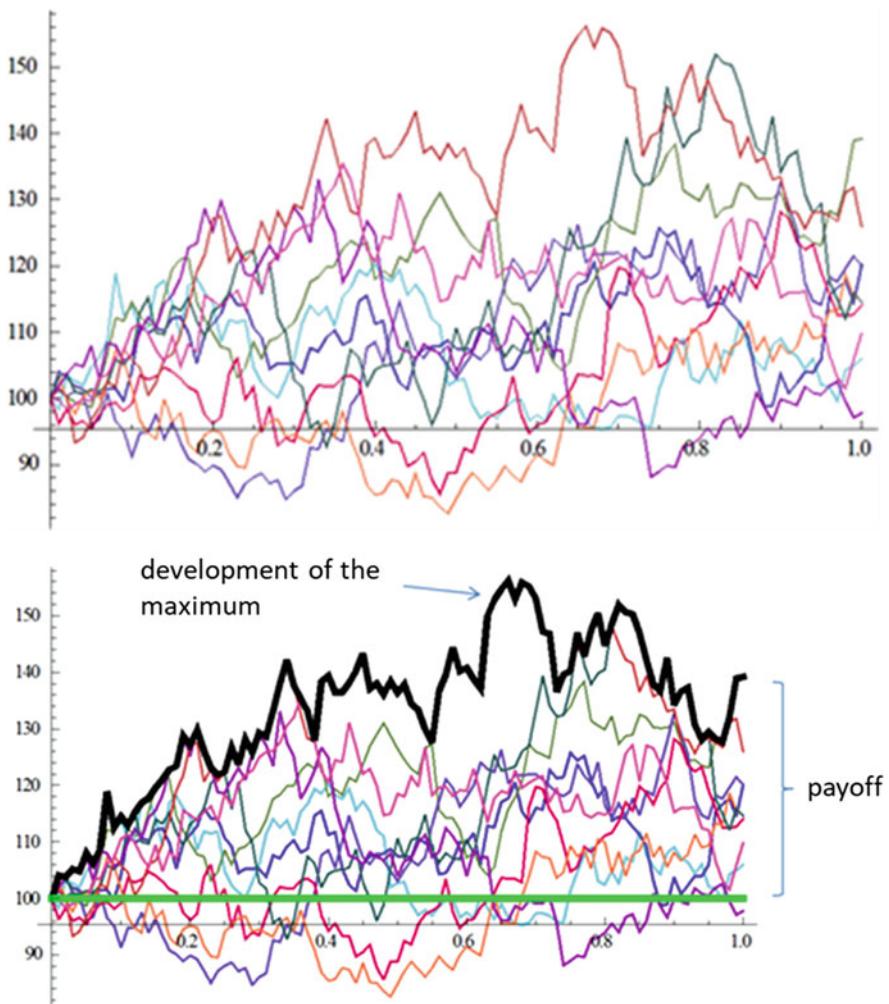


Fig. 2.72 A possible scenario and the resulting payoff, example of a multi-asset option

Both simulation examples illustrated in Figs. 2.73 and 2.74 exhibit a much faster stability in the outcomes for the two QMC techniques compared to the MC method. The approximate value obtained in both simulation examples is of the form 38.6 ... for both the Hammersley and Niederreiter point sets.

Using the MC method, the first simulation yields an approximate value of 38.89..., and the second simulation yields an approximate value of 38.13....

Much more time-consuming simulation to obtain a reference value shows that the actual value of the derivative is indeed likely to be 38.6....

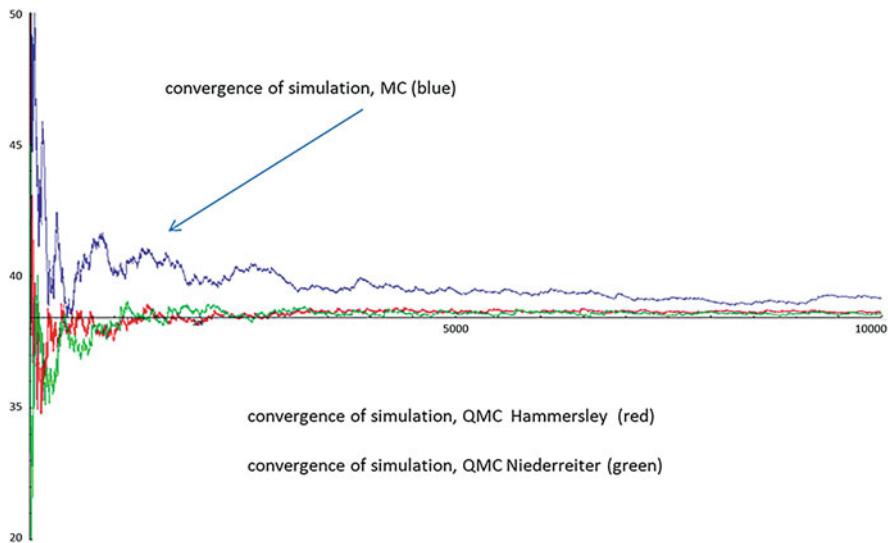


Fig. 2.73 Convergence behaviour, valuation of a multi-asset option with MC and with QMC, simulation 1

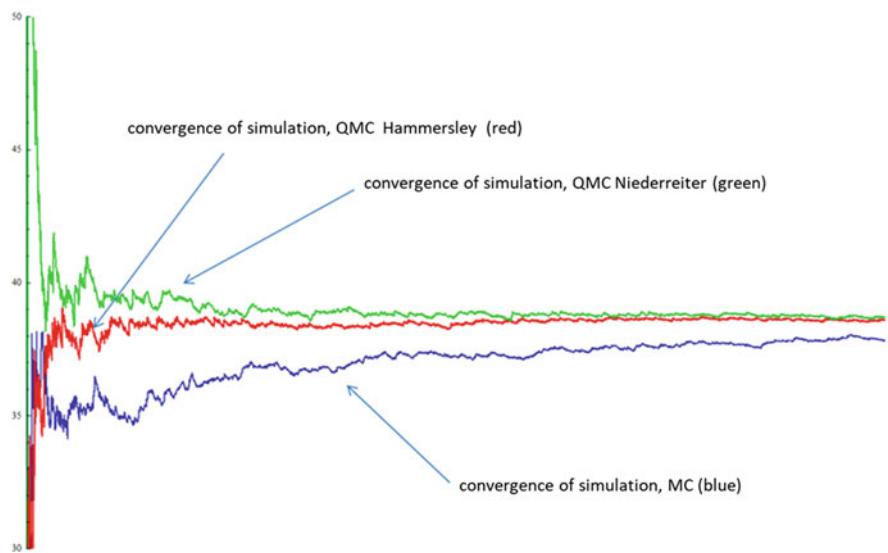


Fig. 2.74 Convergence behaviour, valuation of a multi-asset option with MC and with QMC, simulation 2

It is worth noting, however, that in our programs, the time required to run the QMC simulations was approximately five times greater than the time required to run the MC simulations.

References

1. Donald Knuth. *The Art of Computer Programming*. Addison-Wesley Professional, 2011.
2. Harald Niederreiter. *Random number generation and quasi-Monte Carlo methods*. Vol. 63. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. pp. vi+241. ISBN: 0-89871-295-5. DOI: 10.1137/1.9781611970081. <https://doi.org/10.1137/1.9781611970081>.
3. Gerhard Larcher and Gunther Leobacher. Quasi-Monte Carlo and Monte Carlo methods and their applications in finance. *Surveys on Mathematics for Industry*, 11:95–130, 2005.



Fundamentals: Stochastic Analysis and Applications, Interest Rate Dynamics, and Basic Principles of Pricing Interest Rate Derivatives

3

Keywords

Heuristic methods from stochastic analysis · Stochastic processes · Interest rate models · Ornstein-Uhlenbeck model · Zero-coupon bonds and interest rates · Interest rate swaps · Valuation of interest rate derivatives · Floors · Caps · Vasicek model · Hull-White model · Complete markets · The Black-Scholes market is complete · Valuation in incomplete markets

Before we can proceed, in Volume III Chapter 3, with some case studies to illustrate the basic techniques we have discussed so far, we first need to look at some fundamental methods from other fields of quantitative finance. In the following sections, we will touch upon these fields only superficially, focusing only on aspects that are indispensable for gaining a basic understanding of quantitative finance and financial engineering.

With this in mind, we will deal briefly with three central topics in quantitative finance in the next three sections, namely:

Modelling of Interest Rate Dynamics and Valuation of Interest Rate Derivatives

As this is the first time we deal with this topic, we will also need to provide some basic information about differential representation of stochastic processes. (The aim here is really to impart only an intuitive understanding of these mathematical expressions.)

Value-at-risk and Risk Management On the subject of risk management, we will also, for the time being, provide only some basic techniques used in value-at-risk calculation and credit risk management.

Portfolio Selection Again, for now, we will only present an introduction to the basic concepts of Markowitz's classic portfolio selection theory and of stochastic control theory (which will bring us to the so-called Merton ratio).

3.1 Modelling of Interest Rate Dynamics

To model stock prices, stock index prices, foreign exchange rates, and commodity prices, we have so far been using the Wiener model, that is, a geometric Brownian motion. The question we want to address now is how to adequately model interest rates, such as the 6-month Libor, a 3-month Euribor, a 1-year EUR swap rate, a 10-year CHF swap rate, and a Euro overnight rate, ... (see Volume I Section 1.7). The main reason for our interest in finding such adequate modelling approaches will again be that we need them to price interest rate derivatives (such as interest rate caps and floors, swaptions, or various types of interest rate swaps).

The standard work on modelling interest rate dynamics and pricing interest rate derivatives is [1].

Why can't we just use the Wiener model to model interest rates?

Well, in a reasonably healthy financial market, interest rates will generally range between -1% and possibly 10% (interest rates outside this range are assumed to occur only in highly exceptional cases).

In the longer run, interest rates seem to fluctuate up and down around a long-term mean somewhere in the 3% range.

What is immediately clear: The Wiener model $S(T) = S(0) \cdot e^{\mu T + \sigma \sqrt{T} w}$ definitely does not allow negative values (or, more precisely, since the exponential function is always positive, $S(T)$ is always positive if $S(0)$ is positive, and $S(T)$ is always negative if $S(0)$ is negative.) So, a priori, this model cannot be used to represent a change from positive to negative interest rates or vice versa.

Recall the expected value and standard deviation of a price $S(T)$ that is given by a Wiener model of the form $S(T) = S(0) \cdot e^{\mu T + \sigma \sqrt{T} w}$: We have (see Volume I Section 4.8)

$$E(S(T)) = S(0) \cdot e^{T(\mu + \frac{\sigma^2}{2})}$$

and

$$\sigma(S(T)) = S(0) \cdot \sqrt{e^{T(2\mu + 2\sigma^2)} - e^{T(2\mu + \sigma^2)}}.$$

Given the above remarks about the essential bounds of interest rate values, it follows that when modelling interest rates using the Wiener model, the expected value of the model would also have to be bounded, that is, $\mu \leq -\frac{\sigma^2}{2}$ would always have to be satisfied. Since, over time, $\mu < -\frac{\sigma^2}{2}$ would lead to an expected interest rate value that tends to 0, the only choice that makes sense in principle is $\mu = -\frac{\sigma^2}{2}$! But

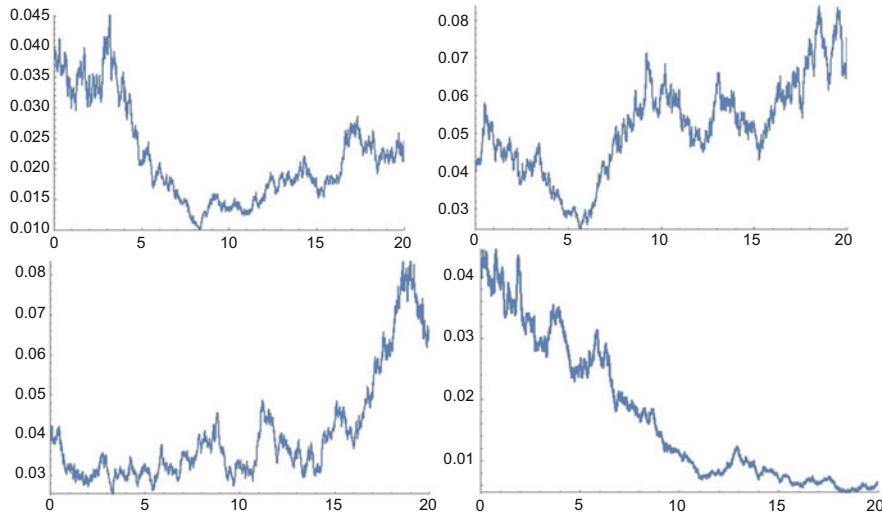


Fig. 3.1 Geometric Brownian motions with expectation 0 and 20 years to expiration

even if this condition is met, the standard deviation of the model with T will tend to infinity.

For illustration, let us choose, for example, $\mu = -\frac{\sigma^2}{2}$, $\sigma = 0.2$, and $S(0) = 0.04$, so that the Wiener model has an expected value of $S(0) = 0.04$. (Note that we use absolute values for interest rates here, not percentage values!)

The standard deviation in this case is $\sigma(S(T)) = 0.04 \cdot \sqrt{e^{T \cdot \sigma^2} - 1}$.

To make this expression a bit more instructive, we note that we always have $e^x - 1 \geq x$ and that therefore $\sigma(S(T)) = 0.04 \cdot \sqrt{e^{T \cdot \sigma^2} - 1} \geq 0.04\sqrt{T} \cdot \sigma = 0.008\sqrt{T}$.

Thus, for sufficiently large periods of T , this model allows arbitrarily large interest rate values with ever-increasing probability. However, for moderately long periods, say, up to $T = 20$ years, the standard deviation is not yet dramatically large (~ 0.03 in our example). In Fig. 3.1, we see some typical simulation paths of a Wiener model with the above parameters on a 20-year period.

In terms of the assumed value range, these simulations do indeed provide quite realistic interest rate values. But negative values are not possible, of course.

The situation changes significantly when simulations are performed over a range of, say, 100 years (or with greater volatility). In Fig. 3.2, we see two typical paths over 100 years with otherwise unchanged parameters.

The values we see here are already—at least under current conditions—clearly outside the realistic value ranges for interest rates.

So, even if using a Wiener model with expectation $S(0)$ for modelling interest rates (apart from the issue of negative interest rates) does not seem completely out

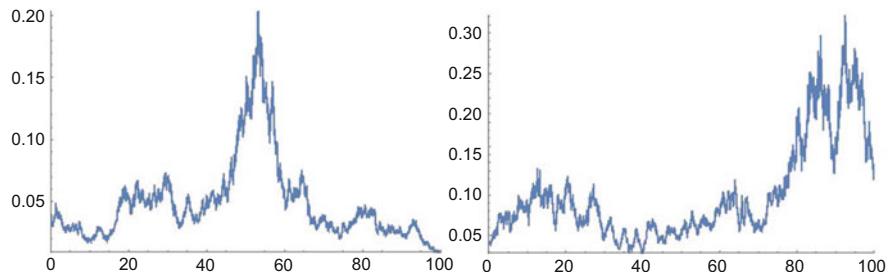


Fig. 3.2 Geometric Brownian motions with expectation 0 and 100 years to expiration

of the question at first, there are additional relevant reasons to reject interest rate modelling in a Wiener model. We will get back to these reasons later.

However, in the course of the following introduction to the problems of modelling interest rates, we will also come to realize the following—the main challenge will lie not so much in finding an appropriate modelling for **one** such interest rate but above all in addressing the following question:

If I choose a particular model with certain parameters for a specific interest rate, for example, the 6-month euro Libor, will this choice – based on the no-arbitrage principle – then influence the choice of models for other euro interest rates, such as the overnight Libor or the 10-year euro swap? Or will the choice of a model for one of these interest rates, and given the no-arbitrage principle, even dictate a mandatory model for all other interest rates? Or if certain models are chosen for one of these interest rates, will it then be possible at all to find arbitrage-free models for other interest rates? In other words, do certain models have to be excluded a priori when it comes to modelling interest rates?

So far, we have used only the Wiener model (the geometric Brownian motion) for modelling purposes. In the following, we are going to need other types of models (which will still be based on the Brownian motion, however). In introducing these models and to ensure a better intuitive understanding, we are going to use a mathematical formalism which we will explain in the following section (in its basic traits and heuristically only for the time being).

3.2 Differential Representation of Stochastic Processes: Heuristic Introduction

Readers without advanced mathematical training should not feel discouraged by the somewhat unwieldy title of this chapter. The formalism referred to above will in

fact be presented very gently in the following and will soon prove to be an extremely helpful tool. So please, embark on this journey with confidence.

If one wanted to provide a rigorously and technically fully accurate introduction to the mathematical disciplines of the “theory of stochastic processes”, “stochastic analysis”, and “stochastic differential equations”, one would have to dive deep into the mathematical foundations and would need to be proficient in a myriad of mathematical techniques and disciplines. In fact, these branches of mathematics are both extremely challenging and technically difficult fields.

Here in this section, we will simply provide an intuitive and heuristic approach to a formalism for representing stochastic processes and will explain some properties of this formalism intuitively. For readers interested in delving deeper into the basics of stochastic analysis, we recommend the books by Björk [2] or Steele [3].

All of the following should therefore not be measured by the standard of mathematical exactness, but should instead be understood as an intuitive heuristic (and thus inevitably superficial) approach to this topic:

- As before, we specify a time range $[0, T]$.
- $[0, T]$ is a continuous time range. For illustration, we will occasionally discretize this time range, subdividing it into N equal parts of length Δt and not allowing every t from $[0, T]$ to be a time point, only the subdivision points $t_0 := 0, t_1 := 1 \cdot \Delta t, t_2 := 2 \cdot \Delta t, \dots, t_{N-1} := (N - 1) \cdot \Delta t, t_N := N \cdot \Delta t = T$. In most cases, we will assume very large N and thus very small Δt . In the following, we will occasionally switch (for illustrative purposes) from the continuous time range to its discretization but pretend to be dealing with one and the same situation.
- We define a stochastic process S on $[0, T]$ to be a potential, partially, or completely random movement of a certain variable (a stock price, an interest rate, the air temperature at a certain location, ...). By $S(t)$, we denote the value of this movement (this stochastic process) at time t and assume a fixed given value $S(0)$ at time 0.
- The value $S(t)$ is not an explicit fixed value (except for $t = 0$), but can rather attain a different value for each possible realization of the random motion. The value $S(t)$ is therefore a random variable.
- For two different points in time s and t , with $0 \leq s < t \leq T$, the values $S(s)$ and $S(t)$ of the process in a particular realization are not necessarily interdependent, although in most cases they are. For example, if the value $S(s)$ of a stock price is very high at time s , then the probability that the stock price $S(t)$ is also high at time t is usually greater than if $S(s)$ is very small.

We are going to illustrate three possible dependencies between such values $S(s)$ and $S(t)$ in the following, using an example in the discretized time range $\{t_0 = 0, t_1 = \Delta t, t_2 := 2 \cdot \Delta t, t_3 := 3 \cdot \Delta t = T\}$:

Example

- (a) We consider the stochastic process consisting of the values $S(t_0), S(t_1), S(t_2), S(t_3)$ and defined as follows:

$$(t_0) = 0, \quad S(t_1) = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

$$S(t_2) = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

$$S(t_3) = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

A possible realization of this process could be 0, 1, 1, -1, for example. The random variables are completely independent of each other. For each of the four random variables, the probability distribution is fixed. The distribution of, say, $S(t_3)$ as seen from the perspective of time $t_2 = 2 \cdot \Delta t$ is the same as the one seen from time 0, regardless of the values taken by $S(t_1)$ and $S(t_2)$. An illustration of possible paths of the considered stochastic process is given in Fig. 3.3:

- (b) We consider the stochastic process with the values

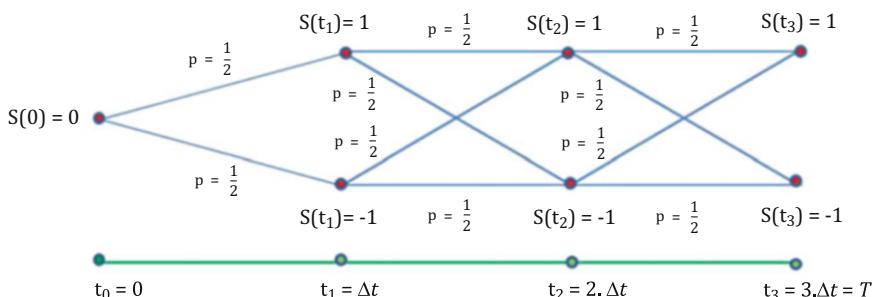


Fig. 3.3 Illustration of possible paths of the stochastic process in example (a)

$S(t_0), S(t_1), S(t_2), S(t_3)$ and defined as follows:

$$S(t_0) = 0, \quad S(t_1) = \begin{cases} S(t_0) + 1 & \text{with probability } \frac{1}{2}, \\ S(t_0) - 1 & \text{with probability } \frac{1}{2}, \end{cases}$$

$$S(t_2) = \begin{cases} S(t_1) + 1 & \text{with probability } \frac{1}{2}, \\ S(t_1) - 1 & \text{with probability } \frac{1}{2}, \end{cases}$$

$$S(t_3) = \begin{cases} S(t_2) + 1 & \text{with probability } \frac{1}{2}, \\ S(t_2) - 1 & \text{with probability } \frac{1}{2} \end{cases}$$

The possible paths of this process can be illustrated as shown in Fig. 3.4.

In example (b), the probability distribution of the values of the process at time points in the future does not remain unchanged but changes with increasing information. For example,

From the perspective of time 0, the values $S(t_2)$ have the following probability distribution (with “W” denoting “probability”):

$$W(S(t_2) = 2) = W(S(t_2) = -2) = \frac{1}{4} \text{ and } W(S(t_2) = 0) = \frac{1}{2}$$

From the perspective of time 0, the values $S(t_3)$ have the following probability distribution:

$$W(S(t_3) = 3) = W(S(t_3) = -3) = \frac{1}{8} \text{ and } W(S(t_3) = 1) =$$

$$W(S(t_3) = -1) = \frac{3}{8}$$

However:

From the perspective of time t_1 if $S(t_1) = 1$, then the values $S(t_2)$ have the following probability distribution:

$$W(S(t_2) = 2) = W(S(t_2) = 0) = \frac{1}{2}$$

From the perspective of time t_1 if $S(t_1) = -1$, then the values $S(t_2)$ have the following probability distribution:

$$W(S(t_2) = 0) = W(S(t_2) = -2) = \frac{1}{2}$$

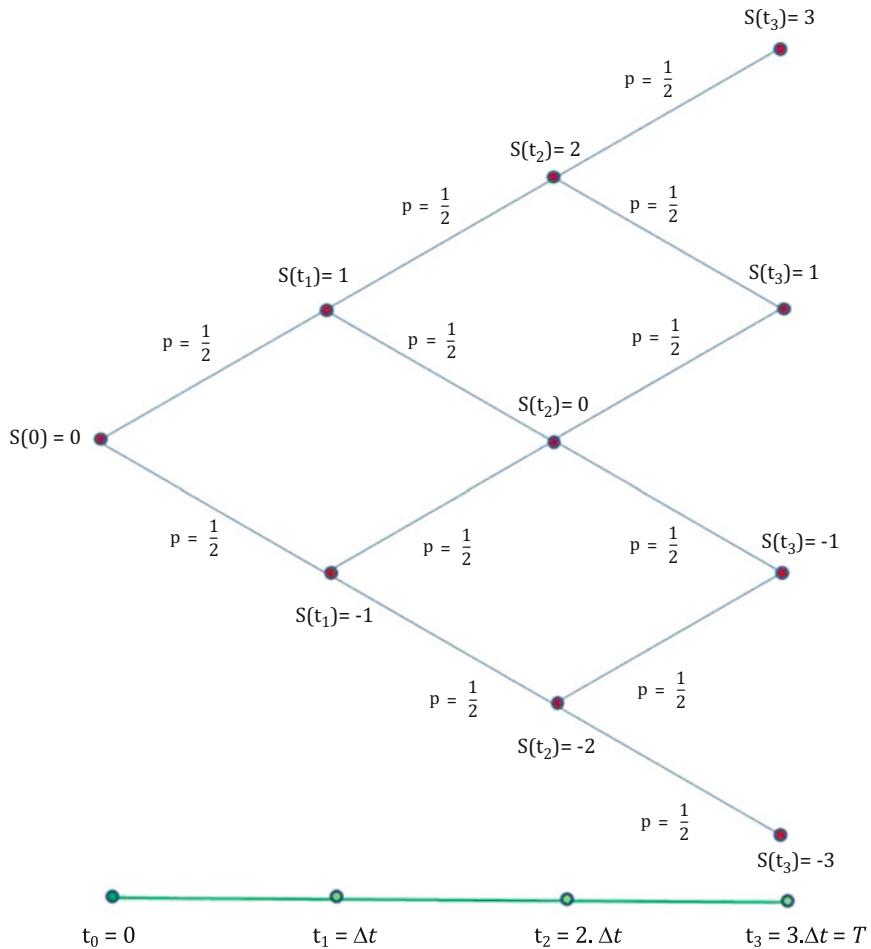


Fig. 3.4 Illustration of possible paths of the stochastic process in example (b)

From the perspective of time t_1 if $S(t_1) = 1$, then the values $S(t_3)$ have the following probability distribution:

$$W(S(t_3) = 3) = W(S(t_2) = -1) = \frac{1}{4} \text{ und } W(S(t_2) = 1) = \frac{1}{2}$$

And:

From the perspective of time t_2 if $S(t_2) = 2$, then the values $S(t_3)$ have the following probability distribution:

$$W(S(t_3) = 3) = W(S(t_2) = 1) = \frac{1}{2}$$

From the perspective of time t_2 if $S(t_2) = 0$, then the values $S(t_3)$ have the following probability distribution:

$$W(S(t_3) = 1) = W(S(t_2) = -1) = \frac{1}{2}$$

From the perspective of time t_2 if $S(t_2) = -2$, then the values $S(t_3)$ have the following probability distribution:

$$W(S(t_3) = -1) = W(S(t_2) = -3) = \frac{1}{2}$$

So, we see that, as information progresses, the probability distributions of future price values do indeed change.

One important observation: If we are currently at time t_1 , then the probability distributions for the price values at the later time points t_2 and t_3 are fixed solely by the current price value at time t_1 . If we are at time t_2 , then the probability distribution for the price values at time t_3 is fixed solely by the current price value at time t_2 . This will no longer be the case in the next example!

- (c) We consider the stochastic process consisting of the values $S(t_0), S(t_1), S(t_2)$, and $S(t_3)$ and defined as follows:

$$S(t_0) = 0, \quad S(t_1) = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

$$S(t_2) = \begin{cases} S(t_1) + 1 & \text{with probability } \frac{1}{2} \\ S(t_1) - 1 & \text{with probability } \frac{1}{2} \end{cases} \quad \text{and}$$

$$S(t_3) = \begin{cases} S(t_2) + S(t_1) + 1 & \text{with probability } \frac{1}{2} \\ S(t_2) + S(t_1) - 1 & \text{with probability } \frac{1}{2} \end{cases}$$

The possible paths of this process can be illustrated as shown in Fig. 3.5.

If, in this process, we are at the price value $S(t_2) = 0$ at time t_2 , then the probability distribution of $S(t_3)$ from this perspective is no longer fixed solely by knowing the value of $S(t_2)$. To be able to calculate the probability distribution of $S(t_3)$ in this case, we also need to know the value that $S(t_1)$ had.

The dependency structure of example (b) has its own name:

Let S be a stochastic process on a time range $[0, T]$. The process is called a “Markov process” if for all time points s and t with $0 \leq s < t \leq T$, the probability distribution for $S(t)$ from the perspective of s is determined solely by knowing the value $S(s)$.

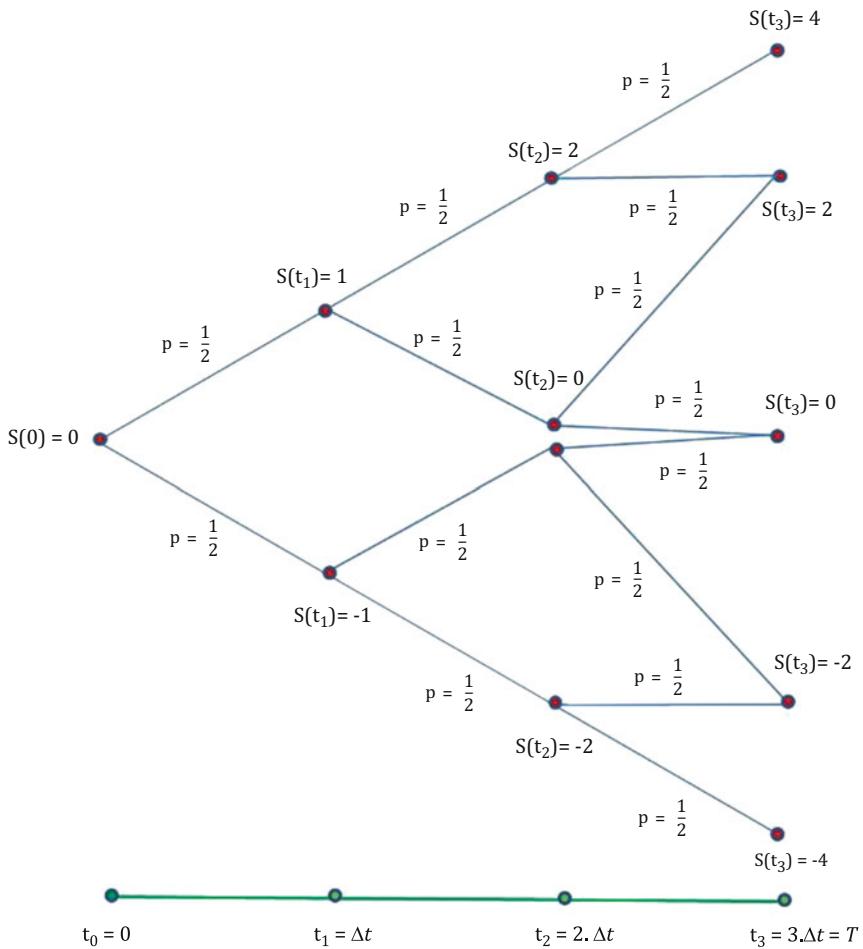


Fig. 3.5 Illustration of possible paths of the stochastic process in example (c)

The process in example (b) (on the discretized time range) is a Markov process. Trivially, of course, the process in example (a) is also a Markov process (where we do not even need to know $S(s)$).

The process in example (c), on the other hand, is not Markovian.

We can also formally express this property in the following way: For all given s and t with $0 \leq s < t \leq T$, we have:

$$S(t) = f(S(s), s, t, X_1, X_2, \dots),$$

where f is a given real-valued function and X_1, X_2, \dots are random variables that are independent from all previous values $S(u)$ with $u < t$ and with given probability distribution.

Using such Markov processes for modelling financial prices is intuitively plausible from the perspective of the efficient Market hypothesis (EMH), which assumes that every financial product is always fairly priced at any given moment. All information about this product is therefore included in the current price. So, all that is needed to obtain the probability of future prices is information about the current price. Information about previous prices is superfluous. The distribution of future prices $S(t)$ can be determined (if at all, then) purely on the basis of the current price $S(s)$.

A frequently occurring dependency structure has the general form

$$S(t) = f(S(s), s, t, w),$$

where w is one (single) standard normally distributed random variable independent of all previous values $S(u)$ with $u < t$.

In the following, we will again take the version of the time range $[0, T]$ that is discretized by very small time intervals Δt and look at two successive time points on this (discretized) time range, namely, t and $t + \Delta t$. Based on the above representation, the relation between $S(t)$ and $S(t + \Delta t)$ then has the form

$$S(t + \Delta t) = f(S(t), t, t + \Delta t, w).$$

In practice, this relationship is frequently of the form

$$S(t + \Delta t) = S(t) + \alpha \cdot \Delta t + \beta \cdot w \quad (3.1)$$

So the value $S(t + \Delta t)$ of S at time $t + \Delta t$ is calculated from the “immediately preceding” value $S(t)$ plus a value α multiplied by the length Δt of the time interval from t to $t + \Delta t$ plus a value β multiplied by the standard normally distributed w .

(Here, for now, we are going to assume fixed constants α and β in this equation. Later, we will also allow variable parameters for α and β .)

In principle, this is a reasonable and quite realistic approach (with just one flaw, which we will rectify in a moment):

For random processes in various application areas, the new value $S(t + \Delta t)$ results from the immediately preceding value $S(t)$ plus a “direction parameter” α multiplied by the time elapsed since the last observation Δt plus a (normally distributed) random term $\beta \cdot w$.

Now, what is the above-mentioned “flaw” in this approach?

- The approach should hold for “small” time intervals Δt .
- If Δt is very small and n is a not too large natural number, then $n \cdot \Delta t$ is also still a “small” time range, and so the relation (3.1) should hold for $n \cdot \Delta t$ as well. Thus

$$S(t + n \cdot \Delta t) = S(t) + \alpha \cdot n \cdot \Delta t + \beta \cdot \tilde{w} \quad (3.2)$$

with a standard normally distributed \tilde{w} .

- On the other hand, the relation (3.1) also holds for each of the n time steps of length Δt from t to $n \cdot \Delta t$; thus,

$$\begin{aligned}
S(t + n \cdot \Delta t) &= S(t + (n - 1) \cdot \Delta t) + \alpha \cdot \Delta t + \beta \cdot w_1 = \\
&= (S(t + (n - 2) \cdot \Delta t) + \alpha \cdot \Delta t + \beta \cdot w_2) + \alpha \cdot \Delta t + \beta \cdot w_1 = \\
&= S(t + (n - 2) \cdot \Delta t + \alpha \cdot 2 \cdot \Delta t) + \beta \cdot (w_1 + w_2) = \\
&= (S(t + (n - 3) \cdot \Delta t) + \alpha \cdot \Delta t + \beta \cdot w_3) + \alpha \cdot 2 \cdot \Delta t + \\
&\quad + \beta \cdot (w_1 + w_2) = \\
&= S(t + (n - 3) \cdot \Delta t) + \alpha \cdot 3 \cdot \Delta t + \beta \cdot (w_1 + w_2 + w_3) = \\
&\dots = S(t) + \alpha \cdot n \cdot \Delta t + \beta \cdot (w_1 + w_2 + w_3 + \dots + w_n) \quad (3.3)
\end{aligned}$$

with mutually independent standard normally distributed $w_1, w_2, w_3, \dots, w_n$.

If we now compare the Formulas (3.2) and (3.3), we notice an inconsistency: Instead of the standard normally distributed random variables \tilde{w} in Formula (3.2), we now have a sum $Y := w_1 + w_2 + w_3 + \dots + w_n$ of n standard normally distributed random variables. While this Y is thus still normally distributed with expected value 0, it is no longer standard normally distributed but has a variance of n . Hence, $Y = \sqrt{n} \cdot w$ with a standard normally distributed w .

Formula (3.3) thus becomes $S(t + n \cdot \Delta t) = S(t) + \alpha \cdot n \cdot \Delta t + \beta \cdot \sqrt{n} \cdot w$ and is therefore not congruent with Formula (3.2).

Meaning, original Formula (3.1) needs to be adapted such that another time-dependent factor is added to the factor β (in the same way that the time factor Δt is added in the first part $\alpha \cdot \Delta t$ on the right side of Formula (3.1)). The correct version of Formula (3.1) then is

$$S(t + \Delta t) = S(t) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w \quad (3.4)$$

with a standard normally distributed random variable w .

- If we perform the above calculation for the time interval from t to $t + n \Delta t$ once more, yet with the new formulation, we now get:

$$\begin{aligned}
S(t + n \cdot \Delta t) &= S(t + (n - 1) \cdot \Delta t) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w_1 = \\
&= \left(S(t + (n - 2) \cdot \Delta t) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w_2 \right) + \alpha \cdot \Delta t + \\
&\quad + \beta \cdot \sqrt{\Delta t} \cdot w_1 = \\
&= S(t + (n - 2) \cdot \Delta t) + \alpha \cdot 2 \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot (w_1 + w_2) = \\
&= \left(S(t + (n - 3) \cdot \Delta t) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w_3 \right) + \alpha \cdot 2 \cdot \Delta t +
\end{aligned}$$

$$\begin{aligned}
& + \beta \cdot \sqrt{\Delta t} \cdot (w_1 + w_2) = \\
& = S(t + (n-3) \cdot \Delta t) + \alpha \cdot 3 \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot (w_1 + w_2 + w_3) = \\
& \dots = S(t) + \alpha \cdot n \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot (w_1 + w_2 + w_3 + \dots + w_n) = \\
& = S(t) + \alpha \cdot n \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot Y = S(t) + \alpha \cdot (n \cdot \Delta t) + \\
& + \beta \cdot \sqrt{\Delta t} \cdot \sqrt{n} \cdot w = \\
& = S(t) + \alpha \cdot (n \cdot \Delta t) + \beta \cdot \sqrt{n \cdot \Delta t} \cdot w,
\end{aligned}$$

which corresponds exactly to Formula (3.4) (with time range $n \cdot \Delta t$ instead of Δt)!

To Summarize

In very many real-world applications, there exists, for small time ranges Δt , a relation of the following form:

$$S(t + \Delta t) = S(t) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w$$

Let us again look at such a process step by step on the time interval $[0, T]$ discretized by means of Δt (see also Fig. 3.6):

$$S(0)$$

$$S(\Delta t) = S(0) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w_1$$

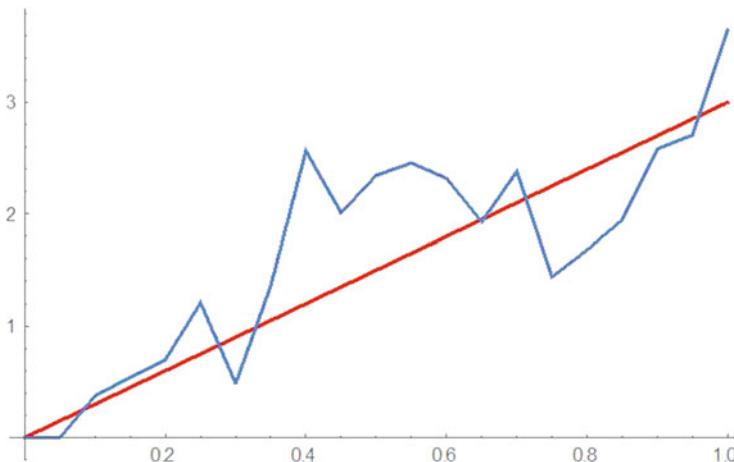


Fig. 3.6 Step-by-step evolution of the process $S(t + \Delta t) = S(t) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w$, 20 steps in $[0, 1]$

$$S(2 \cdot \Delta t) = S(0) + \alpha \cdot 2\Delta t + \beta \cdot (\sqrt{\Delta t} \cdot w_1 + \sqrt{\Delta t} \cdot w_2)$$

...

$$S(n \cdot \Delta t) = S(0) + \alpha \cdot n \Delta t + \beta \cdot (\sqrt{\Delta t} \cdot w_1 + \sqrt{\Delta t} \cdot w_2 + \dots + \sqrt{\Delta t} \cdot w_n)$$

...

$$S(T) = S(N \cdot \Delta t) = S(0) + \alpha \cdot N \Delta t + \beta \cdot (\sqrt{\Delta t} \cdot w_1 + \sqrt{\Delta t} \cdot w_2 + \dots + \sqrt{\Delta t} \cdot w_N)$$

What is happening here over time? A first part increases linearly with time with slope α (see red line in Fig. 3.6). The entire process then evolves in random fluctuation around this deterministic part.

The “random fluctuation” (multiplied by the factor β) is given step by step by

0

$$\sqrt{\Delta t} \cdot w_1$$

$$\sqrt{\Delta t} \cdot w_1 + \sqrt{\Delta t} \cdot w_2$$

...

$$\sqrt{\Delta t} \cdot w_1 + \sqrt{\Delta t} \cdot w_2 + \dots + \sqrt{\Delta t} \cdot w_n$$

...

$$\sqrt{\Delta t} \cdot w_1 + \sqrt{\Delta t} \cdot w_2 + \dots + \sqrt{\Delta t} \cdot w_n + \dots + \sqrt{\Delta t} \cdot w_N$$

Yet, recalling Volume I Section 4.15, we see that this is precisely a standard Brownian motion B on the interval $[0, T]$.

The value $\sqrt{\Delta t} \cdot w_1 + \sqrt{\Delta t} \cdot w_2 + \dots + \sqrt{\Delta t} \cdot w_n$ at time $n \cdot \Delta t$ corresponds exactly to the value $B(n \cdot \Delta t)$ of this Brownian motion at time $n \cdot \Delta t$.

Thus, we can represent $S(n \cdot \Delta t)$ in the form

$$S(n \cdot \Delta t) = S(0) + \alpha \cdot n \cdot \Delta t + \beta \cdot B(n \cdot \Delta t)$$

with a standard Brownian motion B , or if we substitute the discretized time points $n \cdot \Delta t$ by any time point t :

$$S(t) = S(0) + \alpha \cdot t + \beta \cdot B(t).$$

In light of this, let us take another look at the relationship $S(t + \Delta t) = S(t) + \alpha \cdot \Delta t + \beta \cdot \sqrt{\Delta t} \cdot w$ from Formula (3.4)

We have

$$S(t + \Delta t) = S(0) + \alpha \cdot (t + \Delta t) + \beta \cdot B(t + \Delta t)$$

and

$$S(t) = S(0) + \alpha \cdot t + \beta \cdot B(t)$$

and therefore

$$S(t + \Delta t) = S(t) + \alpha \cdot \Delta t + \beta \cdot (B(t + \Delta t) - B(t))$$

(so this is an alternative notation of Formula (3.4)), which is equivalent to

$$S(t + \Delta t) - S(t) = \alpha \cdot \Delta t + \beta \cdot (B(t + \Delta t) - B(t)). \quad (3.5)$$

It is sufficient to give this representation for arbitrarily (infinitesimally) small time ranges Δt . As we have seen before, the representation for longer time ranges can then be easily derived from this.

A commonly used notation for the above relation (3.5) for arbitrarily small time periods Δt is

$$dS(t) = \alpha \cdot dt + \beta \cdot dB(t) \quad (3.6)$$

(“The change $dS(t)$ of the process S in a tiny time step from t to $t + \Delta t$ is given by the value α multiplied by the length of the time step plus the value β multiplied by the change in the standard Brownian motion B in the period from t to $t + \Delta t$.”)

To keep our heuristic deduction, or justification, of this representation as simple as possible, we have limited ourselves to constant values α and β for now. However, in real applications, these values α and β can easily be variable and depend at least on the time t or also on the current value $S(t)$.

The general version of Formula (3.1) is therefore

$$dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dB(t) \quad (3.7)$$

Here, α and β are deterministic functions of the current time t and of the current value $S(t)$ of the process. What is more, in order for the process to be uniquely defined on a time range $[0, T]$, it is always necessary that an initial value $S(0)$ of the process is given.

Very many random processes in a wide variety of applications essentially evolve with such dynamics!

A process that evolves according to this dynamic is referred to as an **Ito process**.

$\alpha(t, S(t)) \cdot dt$ is the **drift term** of the Ito process, and $\beta(t, S(t)) \cdot dB(t)$ is the **diffusion term** of the Ito process.

The name “Ito process” derives from the Japanese mathematician Itô Kiyoshi (1915–2008), who developed stochastic analysis with groundbreaking work in the 1940s. Put simply, stochastic analysis provides the basic building blocks as well as techniques that made it possible to adapt the powerful tools of classical analysis (differential and integral calculus, theory of differential equations, calculus

of variations, etc.) so that they become suitable for use in handling and analysing stochastic processes.

The field of stochastic analysis in all its manifestations (stochastic integration, stochastic differential equations, stochastic control theory, etc.) is one of the most challenging areas of modern mathematics. A mathematically exact discussion of these areas would go beyond the scope of this book and would furthermore not align with the intentions of this book. However, step by step, we will attempt to provide a common heuristic understanding of these objects and techniques so as to be able to actually work with them in real-world applications (partly based on “recipes”).

3.3 Simulation of Ito Processes and Basic Models

Before we start using Ito processes in the following, it is necessary that we first of all gain an “intuitive” understanding of how Ito processes work and how to simulate them.

In principle, simulating an Ito process $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dB(t)$ on a time range $[0, T]$ with initial value $S(0)$ is really quite straightforward:

- Choose a discretization of the time range $[0, T]$ into N parts of equal length $\Delta t := \frac{T}{N}$.
- Start the process at time 0 with the value $S(0)$.
- Assuming that the values of the process for the time points $0, \Delta t, 2 \cdot \Delta t, \dots, n \cdot \Delta t$ have already been simulated, the value $S((n+1) \cdot \Delta t)$ of the process at time $(n+1) \cdot \Delta t$ is simulated as follows:
Calculate $\alpha(n \cdot \Delta t, S(n \cdot \Delta t))$.
Calculate $\beta(n \cdot \Delta t, S(n \cdot \Delta t))$.
Determine a standard normally distributed random number w_n (that is independent of the random numbers used so far), and set

$$\begin{aligned} S((n+1) \cdot \Delta t) := & S(n \cdot \Delta t) + \alpha(n \cdot \Delta t, S(n \cdot \Delta t)) \cdot \Delta t + \\ & + \beta(n \cdot \Delta t, S(n \cdot \Delta t)) \cdot \sqrt{\Delta t} \cdot w_n \end{aligned}$$

- We plot the points $\{n \cdot \Delta t, S(n \cdot \Delta t)\}$ thus generated for $n = 0, 1, 2, \dots, N$ and connect the successive points linearly.

Each new execution of this simulation process yields a new possible path of the stochastic process. Because of the discretization of the time range, the result will be biased of course. And the coarser the discretization, the stronger the bias. Meaning, the probability distribution of the actual simulated process at the respective points in time no longer coincides entirely with the theoretical distribution of the given process.

We perform the simulation using some basic examples in the following: (Simulations of Ito processes can of course also be performed and illustrated with our software.)

Example 1 (Brownian Motion with Drift) $dS(t) = \alpha \cdot dt + \beta \cdot dB(t)$ with constant α and β and fixed initial value $S(0)$

In a tiny time range from t to $t + \Delta t$, $S(t)$ increases linearly with slope α . However, the actual increase varies by an $N(0, \Delta t)$ distributed random variable multiplied by a fixed factor β . We already saw above that this stochastic process can also be calculated explicitly from this “differential representation”. We have

$$S(t) = S(0) + \alpha \cdot t + \beta \cdot B(t)$$

(We will see later that this explicit representation can also be obtained by “stochastic integration” of the differential representation.)

So, the process S is a Brownian motion with drift. The drift is the deterministic linear component $\alpha \cdot t$, and the Brownian motion component is the standard Brownian motion multiplied by the constant β .

Simulating this process (with parameters $S(0) = 0$, $\alpha = 3$, $\beta = 2$) according to the above representation, with division of the interval $[0, 1]$ into 100 subintervals, yields a sample path as shown in the left part of Fig. 3.7 and 30 sample paths as seen in the right part of Fig. 3.7. In both images, the deterministic drift component is shown in red.

For $\beta = 0$, this process would naturally be reduced to the deterministic function $S(t) = S(0) + \alpha \cdot t$ (i.e. the red line in Fig. 3.7).

It is relatively obvious that the process can always take both positive and negative values.

Example 2 (Geometric Brownian Motion) $dS(t) = a \cdot S(t) \cdot dt + b \cdot S(t) \cdot dB(t)$ with constant a and b and fixed initial value $S(0)$

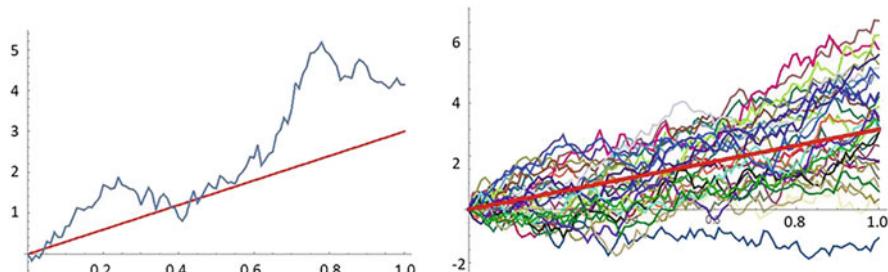


Fig. 3.7 One sample path and 30 sample paths of a simulated Brownian motion with drift, with parameters $S(0) = 0$, $\alpha = 3$, $\beta = 2$

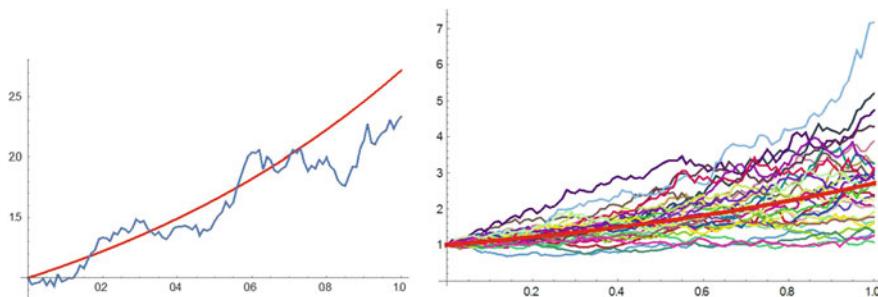


Fig. 3.8 One sample path and 30 sample paths of a simulation of the stochastic process from Example 2, with parameters $S(0) = 0$, $a = 1$, $b = 0.4$

This process differs from the process in Example 1 in that the increase in the drift term (from t to $t + \Delta t$) is not constantly equal to a , but becomes greater the larger the current value $S(t)$ of the process is.

And similarly, the variable random component (the diffusion term) is not constantly equal to b , but becomes larger as the current value $S(t)$ of the process increases.

Simulating this process (with parameters $S(0) = 1$, $a = 1$, $b = 0.4$) according to the above representation, with division of the interval $[0, 1]$ into 100 subintervals, yields a sample path as shown in the left part of Fig. 3.8 and 30 sample paths as seen in the right part of Fig. 3.8. In both images, the deterministic drift component is shown in red.

If we set $b = 0$ here, that is, if we consider the now deterministic process $dS(t) = a \cdot S(t) \cdot dt$, we see that it is in fact an ordinary differential equation.

“Dividing by dt ” yields $S'(t) = a \cdot S(t)$, and solving this differential equation with the given initial value $S(0)$ yields $S(t) = S(0) \cdot e^{at}$, and this is then also the representation of the function of the red line in Fig. 3.8.

So, the drift part of this process is an exponential function.

Can the process given by the differential representation $dS(t) = a \cdot S(t) \cdot dt + b \cdot S(t) \cdot dB(t)$ with constant a and b and fixed initial value $S(0)$ also be represented explicitly (this representation is in fact a “**stochastic differential equation**” (the equation includes both a differential expression $dS(t)$ and $S(t)$ itself)?

The answer is yes. In the next section (an excursus), we will use the Ito formula to prove that the explicit representation of the process has the following form:

$$S(t) = S(0) \cdot e^{\left(a - \frac{b^2}{2}\right) \cdot t + b \cdot B(t)}$$

A **very important information** that we should keep in mind for everything that follows is the process $S(t)$ given by the stochastic differential equation

$$dS(t) = a \cdot S(t) \cdot dt + b \cdot S(t) \cdot dB(t)$$

with constant a and b and fixed initial value $S(0)$ is precisely a **geometric Brownian motion** of the form

$$S(t) = S(0) \cdot e^{\left(a - \frac{b^2}{2}\right) \cdot t + b \cdot B(t)}.$$

The geometric Brownian motion (with positive $S(0)$) can only attain positive values.

Example 3 (Ornstein-Uhlenbeck Model) In this model, we choose a hybrid of Examples 1 and 2. It has the form

$$dS(t) = a \cdot S(t) \cdot dt + b \cdot dB(t)$$

with constant a and b and fixed initial value $S(0)$.

The drift term here corresponds to the drift term of the geometric Brownian motion in Example 2, and the diffusion term corresponds to the diffusion term of the Brownian motion with drift.

The strength of the random deviation thus remains constant regardless of the current value of the process, while the drift term, which is the one that specifies the principal direction of the motion, increases (or decreases) directly as a function of $S(t)$.

For the deterministic version (with $b = 0$), we thus get the same outcome as the geometric Brownian motion, i.e. $S(t) = S(0) \cdot e^{a \cdot t}$.

The Ornstein-Uhlenbeck model can attain both positive and negative values.

Is it possible to state the Ornstein-Uhlenbeck model explicitly? Yes, it's possible, but we would need the tool of stochastic integration for its representation, which we have not dealt with so far. We will do so later however.

Simulating this process (with parameters $S(0) = 1$, $a = 1$, $b = 0.4$) according to the above representation, with division of the interval $[0, 1]$ into 100 subintervals, yields a sample path as shown in the left part of Fig. 3.9 and 30 sample paths as seen in the right part of Fig. 3.9. In both images, the deterministic drift component is again shown in red.

Figure 3.10 compares ten sample paths (blue) of a geometric Brownian motion with ten example paths (red) of an Ornstein-Uhlenbeck model. It clearly shows a much broader average spread of the geometric Brownian motion over time compared to the paths of the Ornstein-Uhlenbeck model.

Example 4 (Mean-Reverting Ornstein-Uhlenbeck Model) This model has the form

$$dS(t) = a \cdot (m - S(t)) \cdot dt + b \cdot dB(t)$$

with constant $a > 0$, b , and m and fixed initial value $S(0)$.

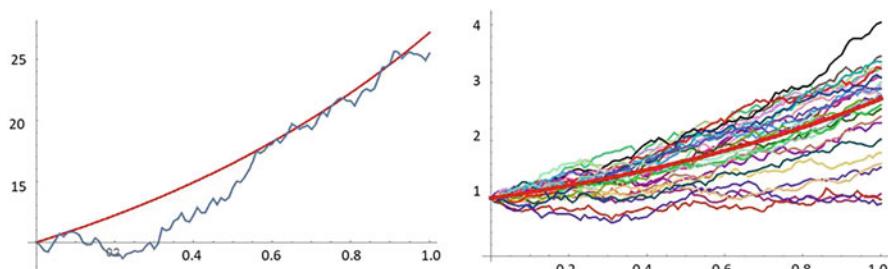


Fig. 3.9 One sample path and 30 sample paths of a simulation of the stochastic process from Example 3 (Ornstein-Uhlenbeck model), with parameters $S(0) = 0$, $a = 1$, $b = 0.4$

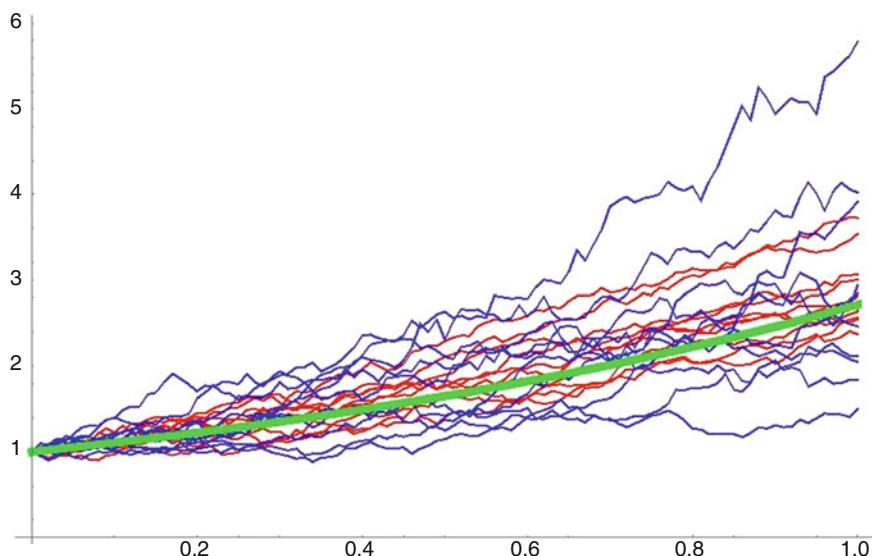


Fig. 3.10 Ten sample paths of a geometric Brownian motion (blue) and ten sample paths of an Ornstein-Uhlenbeck model (red), with parameters $S(0) = 1$, $a = 1$, $b = 0.4$

Thus, the basic form of the model is similar to the Ornstein-Uhlenbeck model (occurrence of an $S(t)$ component in the drift term and constant factor in the diffusion term).

However, the drift term can now attain both negative and positive values over time.

This is because

- If $S(t)$ is smaller than m , then the drift term is positive. A positive change in the value from $S(t)$ to $S(t + \Delta t)$ is more likely. And, in fact, it is all the more likely and probably all the more significant the more clearly $S(t)$ is smaller than m .

- If $S(t)$ is larger than m , then the drift term is negative. A negative change in the value from $S(t)$ to $S(t + \Delta t)$ is more likely. And, in fact, it is all the more likely and probably all the more significant the more clearly $S(t)$ is larger than m .

In other words, if $S(t)$ is larger than m , then $S(t)$ tends to fall, and if it is smaller than m , then $S(t)$ tends to grow. It is highly likely therefore that the value of $S(t)$ will fluctuate around the value m . **$S(t)$ exhibits “mean reversion around m ”.**

Intuitively, a larger value of b should cause stronger price fluctuations of course.

What impact does a larger a -value have? A large a -value naturally amplifies the up and down moves and should therefore cause the price to return to the mean m more quickly than a small a would.

Yet, wouldn't that also imply—so it seems at first glance—that a large a could cause the value to strongly overreact (i.e. revert to m with such strength and being so large that it actually moves far beyond m)!?

Let us look at Fig. 3.11 to illustrate the influence of the parameters a and b using a series of simulations. Before we do that, however, we first contemplate what the deterministic version of the mean-reverting Ornstein-Uhlenbeck model would look like, that is, for $b = 0$. In this case, the representation of the process $S(t)$ is

$$dS(t) = a \cdot (m - S(t)) \cdot dt$$

which yields the ordinary differential equation $S'(t) = a \cdot (m - S(t))$ with initial value $S(0)$.

The solution of this differential equation is $S(t) = m - (m - S(0)) \cdot e^{-t \cdot a}$. For someone who is not used to solving differential equations, this is easily verified by differentiating the given solution $S(t) = m - (m - S(0)) \cdot e^{-t \cdot a}$.

For t to infinity, this deterministic component converges to m .

The mean-reverting Ornstein-Uhlenbeck model can of course also be expressed explicitly, but this would again require the notion of the stochastic integral.

Let us now turn to the simulations: For each of the following simulations, we choose the initial value $S(0) = 1$, the long-term mean $m = 2$, and the time horizon $t = 100$, and we divide the time range into 10,000 parts of equal length.

Then we vary the parameters $a = 0.1, 1, 10$ and $b = 0.2, 1$. For each of the six combinations, we choose a typical simulated path and illustrate the results in Fig. 3.11.

The impact of changing a and b parameters is quite obvious:

Growing b increases the amplitude while the frequency remains essentially unchanged. Growing a obviously increases the frequency and slightly decreases the amplitude. The process can also take negative values.

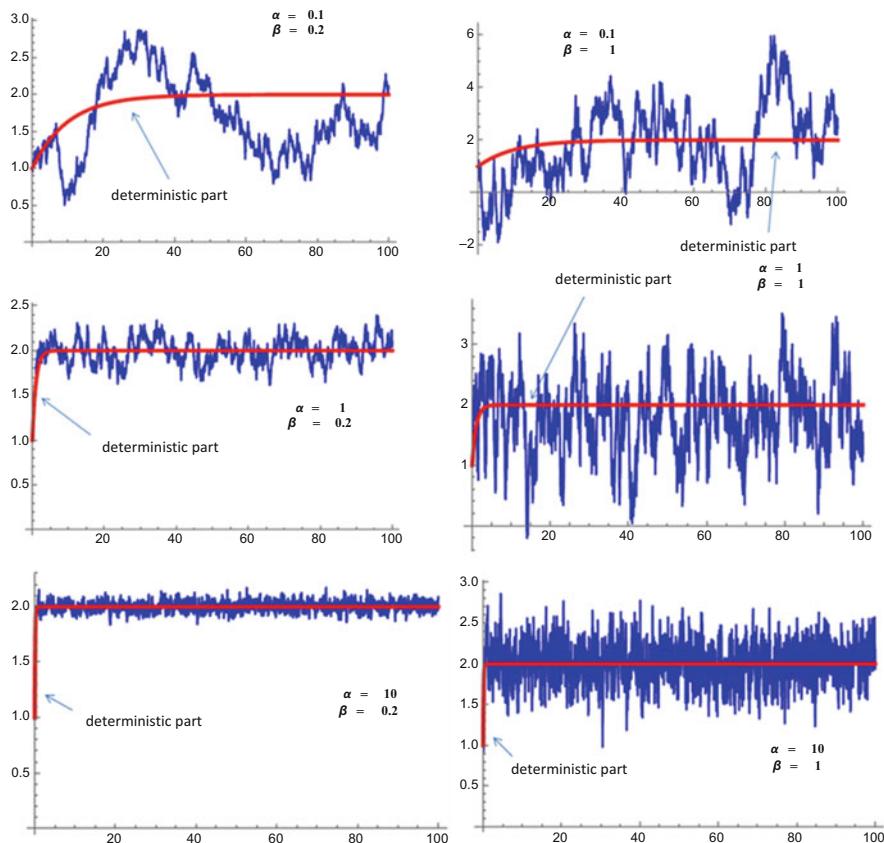


Fig. 3.11 Simulation of the mean-reverting Ornstein-Uhlenbeck model with $m = 2$, $t = 100$, $n = 10,000$ and different parameter sets for a and b

In fact, for the expected value $E(S(t))$ and the variance $V(S(t))$ of the mean-reverting Ornstein-Uhlenbeck model from the perspective of time 0 and with $S(0)$, we get

$$E(S(t)) = S(0) \cdot e^{-at} + m \cdot (1 - e^{-at})$$

$$V(S(t)) = \frac{b^2}{2a} (1 - e^{-2at}).$$

For increasing t , the expression e^{-at} converges to 0 (and does so increasingly fast the larger a is) and the expected value therefore converges to m .

The variance converges to $\frac{b^2}{2a}$, thus getting larger the larger b is (which was to be expected a priori) and the smaller a is (which was to be expected based on Fig. 3.11).

Example 5 (Cox-Ingersoll-Ross Model) For this model, we will discuss the mean-reversion version only. This version has the exact same form as the mean-reversion Ornstein-Uhlenbeck model except that the diffusion term again depends on the value of $S(t)$, though in a weaker form than is the case with the geometric Brownian motion. The strength of that dependence is not linear in $S(t)$, but increases with $\sqrt{S(t)}$.

The model has the form

$$dS(t) = a \cdot (m - S(t)) \cdot dt + b \cdot \sqrt{S(t)} \cdot dB(t)$$

with constant $a > 0$, b , and m and fixed initial value $S(0)$.

The process can be represented explicitly but is relatively complex.

An essential property of this process is that if m is also positive, it will always yield non-negative values. This is easily understood if we think of it this way: If $S(t)$ approaches 0, then the diffusion component also approaches 0, while the drift term, which is then close to the then positive value $\alpha \cdot m \cdot dt$, strongly dominates the process in the positive direction.

Figure 3.12 shows the simulations for the Cox-Ingersoll-Ross model with the same parameters as used in Fig. 3.11 for the mean-reverting Ornstein-Uhlenbeck model.

Particularly notable here is the simulation with parameters $a = 0.1$ and $b = 1$.

In this case, $S(t)$ takes the value 0 at one point and then obviously (due to the form of the model) stays at 0. This cannot happen if $2am \geq b^2$. In this case, the values $S(t)$ will always be positive.

In modelling interest rate curves, we will mainly deal with the mean-reversion versions of the Ornstein-Uhlenbeck (OU) and the Cox-Ingersoll-Ross (CIR) models (or variants thereof). For a long time—until the first occurrence of actual negative interest rates in the financial markets—the fact that the CIR model yields only positive values was considered a convincing argument for the superiority of the CIR model over the OU model when it came to interest rate modelling.

Example 6 (Exponential Growth with Limited Capacity) This model will play less of a role in our applications in the following, yet we will introduce it briefly here since it is often used in a wide variety of other applications. It is adequate especially for stochastic processes with (stochastically) exponential growth of a number of “individuals” (of whatever form) where the capacity to accommodate these individuals (the habitat) is limited by a bound M (e.g. a lake as a habitat for fish) or where the number of individuals is too large to allow the emergence of more individuals.

The model has the following form:

$$dS(t) = a \cdot S(t) \cdot (M - S(t)) \cdot dt + b \cdot S(t) \cdot dB(t)$$

with constant a , b , and positive M and fixed initial value $S(0)$.

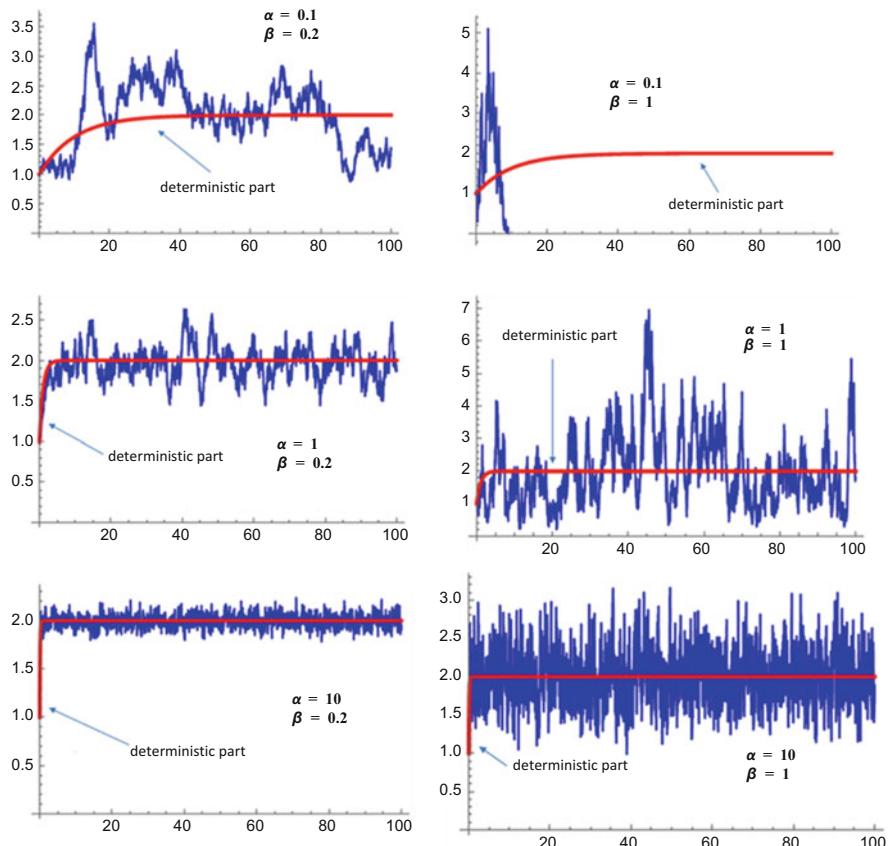


Fig. 3.12 Simulation of the mean-reverting Cox-Ingersoll-Ross model with $m = 2$, $t = 100$, $n = 10,000$ and different parameter sets for a and b

So, we are looking at a geometric Brownian motion (GBM) but with an additional factor ($M - S(t)$) in the drift term.

A small $S(t)$ gives the model the properties of a GBM with a drift constant of approximately $a \cdot M$. If $S(t)$ grows, the GBM behaviour remains essentially the same for the time being, yet the drift constant becomes smaller. If $S(t)$ approaches the capacity-bound M , growth will slow drastically and become negative if it briefly exceeds M at any one point, whereupon it will quickly drop below M .

A typical path of this model with initial value $S(0) = 1$, capacity-bound $M = 20$, $a = 0.1$, and $b = 0.3$ is shown in Fig. 3.13.

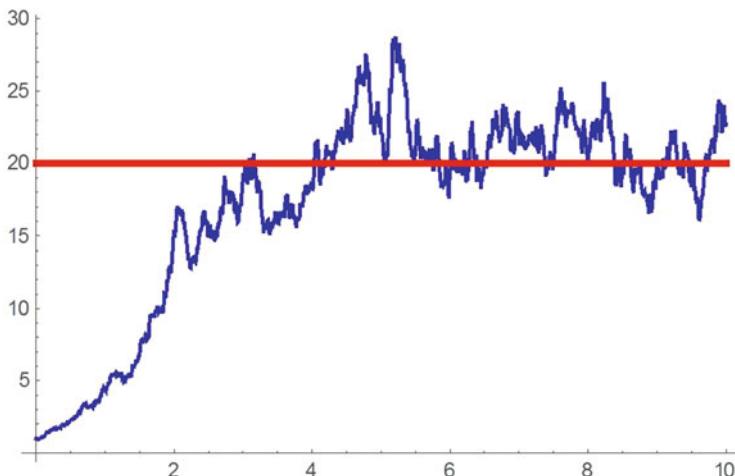


Fig. 3.13 Path of a limited capacity exponential growth model

3.4 Excursus: The Ito Formula and Differential Notation of the GBM

Probably the most important tool in stochastic analysis is the Ito formula.

As noted earlier, we can introduce the basic features of stochastic analysis only intuitively and heuristically here and therefore formulate, motivate, and apply the Ito formula only intuitively, by using it to demonstrate that the differential notation of the GBM— $S(t) = S(0) \cdot e^{(a - \frac{b^2}{2}) \cdot t + b \cdot B(t)}$ —is indeed given by $dS(t) = a \cdot S(t) \cdot dt + b \cdot S(t) \cdot dB(t)$ with initial value $S(0)$ or, conversely, that the GBM in the above form is indeed the solution of this stochastic differential equation.

Remember, an Ito process is a stochastic process $S(t)$ whose dynamics is given by $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dB(t)$ with fixed given initial value $S(0)$.

If a function f is applied to such a process S , then $Y(t) := f(S(t))$ is also a stochastic process (e.g. $Y(t) = (S(t))^2$ or $Y(t) = \log(1 + \sin^2(S(t))), \dots$).

Or, more generally, let g be any function of two variables t and x , and then $Y(t) := g(t, S(t))$ is again a stochastic process (e.g. $Y(t) = (t - S(t))^2$, or $Y(t) = S(t) \cdot \log(t + \sin^2(t \cdot S(t))), \dots$).

This begs the question: Is the new process $Y(t)$ also an **Ito process**? (Or under what conditions is $Y(t)$ an Ito process?)

(continued)

Meaning, is the dynamics of the new process $Y(t)$ again of the form $dY(t) = \gamma \cdot dt + \delta \cdot dB(t)$ with any functions γ and δ ?

And if so, what would these functions γ and δ be, that is, what is the “**Ito representation**” of Y ?

To answer that question, we need to find a way to determine $dY(t)$.

Let us be brazenly heuristic here and ask how would we proceed in classical analysis in this case? We would try to differentiate $Y(t) = g(t, S(t))$ with respect to t . To do that, we need to apply the chain rule from classical analysis (we recall for this purpose that $g(t, x)$ is a function of the variables t and x). And in the following, we denote by g_t the partial derivative of g with respect to the first variable t and by g_x the partial derivative of g with respect to the second variable x):

$$\begin{aligned}\frac{dY(t)}{dt} &= \frac{dg(t, S(t))}{dt} = g_t(t, S(t)) \cdot \frac{dt}{dt} + g_x(t, S(t)) \cdot \frac{dS(t)}{dt} = \\ &= g_t(t, S(t)) + g_x(t, S(t)) \cdot \frac{dS(t)}{dt}\end{aligned}$$

We remain brazen and “multiply by dt ”, which gives us

$$dY(t) = g_t(t, S(t)) \cdot dt + g_x(t, S(t)) \cdot dS(t).$$

Substituting the Ito representation $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dB(t)$ of $S(t)$ in this equation, we get

$$\begin{aligned}dY(t) &= g_t(t, S(t)) \cdot dt + g_x(t, S(t)) \cdot (\alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dB(t)) = \\ &= (g_t(t, S(t)) + g_x(t, S(t)) \cdot \alpha(t, S(t))) \cdot dt + \\ &\quad + g_x(t, S(t)) \cdot \beta(t, S(t)) \cdot dB(t)\end{aligned}$$

And with this, it would appear that we have found an Ito representation for $Y(t)$

(with $\gamma = (g_t(t, S(t)) + g_x(t, S(t)) \cdot \alpha(t, S(t)))$ and with $\delta = g_x(t, S(t)) \cdot \beta(t, S(t))$ or, in a short form, $\gamma = g_t + g_x \cdot \alpha$ and $\delta = g_x \cdot \beta$).

But it is really just apparently that we have found an Ito representation, because in stochastic analysis, we cannot proceed in the same way as in classical analysis. More importantly, as we will see in a moment, this representation would even be wrong!

However, proceeding in this way gave us an inkling that it should indeed be possible to arrive at an Ito representation of $Y(t)$ with similar—but then actually admissible—manipulations.

Indeed, we have

(continued)

Theorem 3.1 (Ito Formula) Let S be an Ito process of the form

$$dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dB(t) \text{ with fixed initial value } S(0).$$

Let

$$\begin{aligned} g : \mathbb{R} \times \mathbb{R} &\rightarrow \mathbb{R} \\ (t, x) &\mapsto g(t, x) \end{aligned}$$

be a function that is once continuously differentiable with respect to t and twice continuously differentiable with respect to x .

Then the process $Y(t) := g(t, S(t))$ is again an Ito process, and we get

$$\begin{aligned} dY(t) &= \\ &= \left(g_t(t, S(t)) + g_x(t, S(t)) \cdot \alpha(t, S(t)) + g_{xx}(t, S(t)) \cdot \frac{(\beta(t, S(t)))^2}{2} \right) \cdot dt + \\ &\quad + g_x(t, S(t)) \cdot \beta(t, S(t)) \cdot dB(t). \end{aligned}$$

$$\text{In short, } dY = \left(g_t + g_x \cdot \alpha + g_{xx} \cdot \frac{\beta^2}{2} \right) \cdot dt + g_x \cdot \beta \cdot dB.$$

If we compare the naive, classical approach sketched out above with the actual Ito formula, we see that the stochastic (Ito) version has an additional term $g_{xx} \cdot \frac{\beta^2}{2} \cdot dt$. This is typical of results in stochastic analysis: They are often very similar to the analogous results from classical analysis yet require additional terms in most cases.

As promised, we will now apply Ito formula to the geometric Brownian motion. In fact, we will do so in two ways.

Example

(a) We start with the explicit representation of the GBM

$$Y(t) = Y(0) \cdot e^{\left(a - \frac{b^2}{2}\right) \cdot t + b \cdot B(t)}.$$

This process Y is a function $g(t, X(t))$, where $g(t, x) = Y(0) \cdot e^{\left(a - \frac{b^2}{2}\right) \cdot t + b \cdot x}$ and $X(t) = B(t)$.

This $X(t)$ is an Ito process, since $dX(t) = 0 \cdot dt + 1 \cdot dB(t)$.

Therefore, in the terminology of the Ito formula above, $\alpha = 0$ and $\beta = 1$.

(continued)

The function g can be differentiated arbitrarily often with respect to both the first and the second variable. The Ito formula now states that the process Y is thus an Ito process, and we can also use the Ito formula to calculate the differential representation of Y —to do this, we first need the derivatives of g :

$$g_t = \left(a - \frac{b^2}{2} \right) \cdot Y(0) \cdot e^{\left(a - \frac{b^2}{2} \right) \cdot t + b \cdot x} = \left(a - \frac{b^2}{2} \right) \cdot g$$

$$g_x = b \cdot Y(0) \cdot e^{\left(a - \frac{b^2}{2} \right) \cdot t + b \cdot x} = b \cdot g$$

$$g_{xx} = b^2 \cdot Y(0) \cdot e^{\left(a - \frac{b^2}{2} \right) \cdot t + b \cdot x} = b^2 \cdot g$$

Therefore,

$$g_t + g_x \cdot \alpha + g_{xx} \cdot \frac{\beta^2}{2} = g_t + g_{xx} \cdot \frac{1}{2} = a \cdot g = a \cdot Y$$

and

$$g_x \cdot \beta = g_x = b \cdot g = b \cdot Y.$$

The Ito representation for the GBM Y is therefore in shorthand notation

$$dY = a \cdot Y \cdot dt + b \cdot Y \cdot dB$$

and in full notation

$$dY(t) = a \cdot Y(t) \cdot dt + b \cdot Y(t) \cdot dB(t),$$

which is exactly what we wanted to show.

- (b) We will now look at basically the same example again, yet in this case, we are not going to start with the explicit process (the solution to the stochastic differential equation) but with the stochastic differential equation itself $dX(t) = a \cdot X(t) \cdot dt + b \cdot X(t) \cdot dB(t)$. So here, $\alpha = a \cdot X(t)$ and $\beta = b \cdot X(t)$.

In anticipation of the solution to this equation, we are now going to consider a new process $Y(t) := \log X(t)$.

So, we have $Y(t) = g(t, X(t))$, where g is the function $g(t, x) = \log x$.

(continued)

In its domain of definition, g can be differentiated arbitrarily often, and we have $g_t = 0$, $g_x = \frac{1}{x}$, and $g_{xx} = -\frac{1}{x^2}$.

The Ito formula states that Y is again an Ito process, and we can calculate the representation of Y : We have

$$g_t + g_x \cdot \alpha + g_{xx} \cdot \frac{\beta^2}{2} = \frac{1}{x} \cdot a \cdot x - \frac{1}{x^2} \cdot \frac{(bx)^2}{2} = a - \frac{b^2}{2}$$

and

$$g_x \cdot \beta = \frac{1}{x} \cdot b \cdot x = b,$$

and thus

$$dY(t) = \left(a - \frac{b^2}{2} \right) \cdot dt + b \cdot dB(t).$$

Y is therefore a Brownian motion with drift (see previous section, Example 1), of which we know the explicit representation:

$$Y(t) = Y(0) + \left(a - \frac{b^2}{2} \right) \cdot t + b \cdot B(t).$$

And with that, we can then also explicitly represent the process $X(t)$. After all, $Y(t) = \log X(t)$, so

$$\begin{aligned} X(t) &= e^{Y(t)} = e^{Y(0) + \left(a - \frac{b^2}{2} \right) \cdot t + b \cdot B(t)} = e^{Y(0)} \cdot e^{\left(a - \frac{b^2}{2} \right) \cdot t + b \cdot B(t)} = \\ &= X(0) \cdot e^{\left(a - \frac{b^2}{2} \right) \cdot t + b \cdot B(t)}, \end{aligned}$$

which is exactly what we wanted to show.

3.5 Interest Rate Modelling Using Mean-Reverting Ornstein-Uhlenbeck

We now continue from where we left off in Sect. 3.1. Back then we noted that there are other reasons (besides the fact that the GBM can never attain negative values and that it cannot be used to model mean reversion) which make interest rate modelling using GBM seem inadequate.

Let us now look at the differential representation of the GBM:

$dS(t) = a \cdot S(t) \cdot dt + b \cdot S(t) \cdot dB(t)$. So, the random component in the GBM, the diffusion term, is proportional to the value $S(t)$ of the process. This means that for large values of $S(t)$, large local fluctuations are to be expected when measured in absolute terms. For small values of $S(t)$, small local fluctuations are to be expected when measured in absolute terms. If $S(t)$ has a value of 10, then local fluctuations can be expected to be about ten times larger than if $S(t)$ has a value of 1. Is such a property realistic when it comes to interest rates? When interest rates are in the 10% range, can we expect local fluctuations to be ten times greater than when interest rates are in the 1% range? Not really.

Let us run a very simple test: As an example, we take the historical daily closing prices of the Euro Overnight Eonia Index from 1999 to 2014. These almost 4000 values are plotted in red in Fig. 3.14.

Let us assume that this interest rate follows a GBM of the form $dS(t) = a \cdot S(t) \cdot dt + b \cdot S(t) \cdot dB(t)$.

For the daily returns $ren(t) := \frac{S(t+\frac{1}{250}) - S(t)}{S(t)}$ (we assume 250 days per trading year, so $dt = \frac{1}{250}$), we would then get approximately $ren(t) \sim a \cdot dt + b \cdot dB(t)$. These daily returns should thus be essentially independent of the magnitude of $S(t)$. For better graphical representation, we normalize the value: We consider the normalized value $\frac{rent(t)}{\sqrt{dt}} \sim a \cdot \sqrt{dt} + b \cdot w$ with standard normally distributed w . This value should then also fluctuate relatively uniformly around the fixed value $a \cdot \sqrt{dt}$ regardless of the value of $S(t)$. In Fig. 3.14, the values $\frac{rent(t)}{\sqrt{dt}}$ are plotted in blue, and it is clear to see that smaller values of $\frac{rent(t)}{\sqrt{dt}}$ tend to occur with larger values of $S(t)$ and vice versa.

If we however assumed that the interest rate follows an Ornstein-Uhlenbeck model, that is, of the form $dS(t) = a \cdot S(t) \cdot dt + b \cdot dB(t)$, then $\frac{rent(t)}{\sqrt{dt}}$ would have to be essentially of the form $\frac{rent(t)}{\sqrt{dt}} \sim a \cdot \sqrt{dt} + b \cdot w \cdot \frac{1}{S(t)}$, meaning that it would again fluctuate around the fixed value $a \cdot \sqrt{dt}$. However, the amplitude of that fluctuation would then have to be larger for small $S(t)$ and smaller for large $S(t)$, which, as we have seen in Fig. 3.14, is indeed the case.

In fact, if we assumed interest rate dynamics according to an Ornstein-Uhlenbeck model, the value $\left(\frac{rent(t)}{\sqrt{dt}} - a \cdot \sqrt{dt}\right) \cdot S(t) \sim b \cdot w$ would be expected to fluctuate relatively uniformly around 0 regardless of the magnitude of $S(t)$. For the value $a \cdot \sqrt{dt}$, we are going to use an approximate estimate based on the historical data and obtain an estimated value of 0.0125.

In Fig. 3.15, the values $\left(\frac{rent(t)}{\sqrt{dt}} - a \cdot \sqrt{dt}\right) \cdot S(t)$ have again been plotted in blue and are shown in comparison with the index dynamics $S(t)$ plotted in red. As we can see, there is indeed much less dependence between the two values than in Fig. 3.14.

This simple observation too strongly suggests that an Ornstein-Uhlenbeck model is much more suited to modelling interest rate movements than a GBM. For this reason, we will use only (mean-reverting) Ornstein-Uhlenbeck models and variants thereof in our introduction to interest rate modelling.

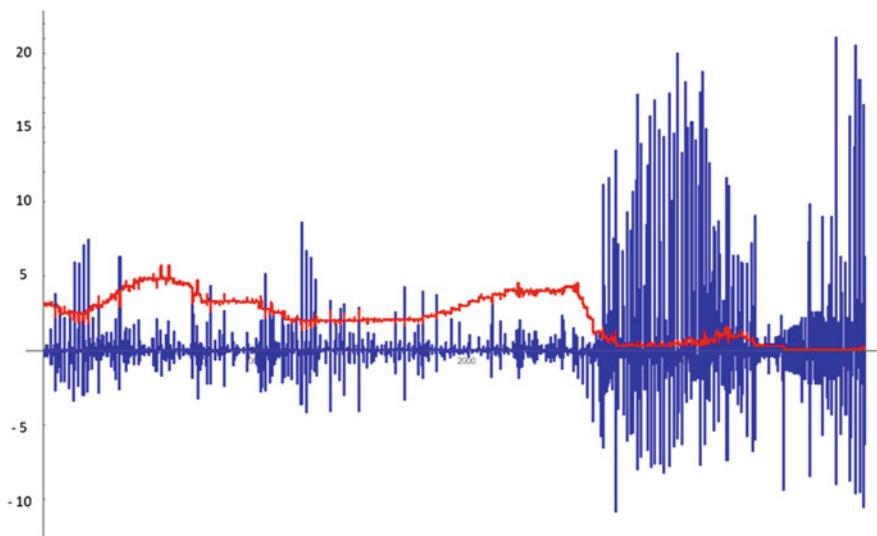


Fig. 3.14 Euro Overnight Eonia Index from 1999 to 2014 (red) compared with normalized relative daily returns (blue)

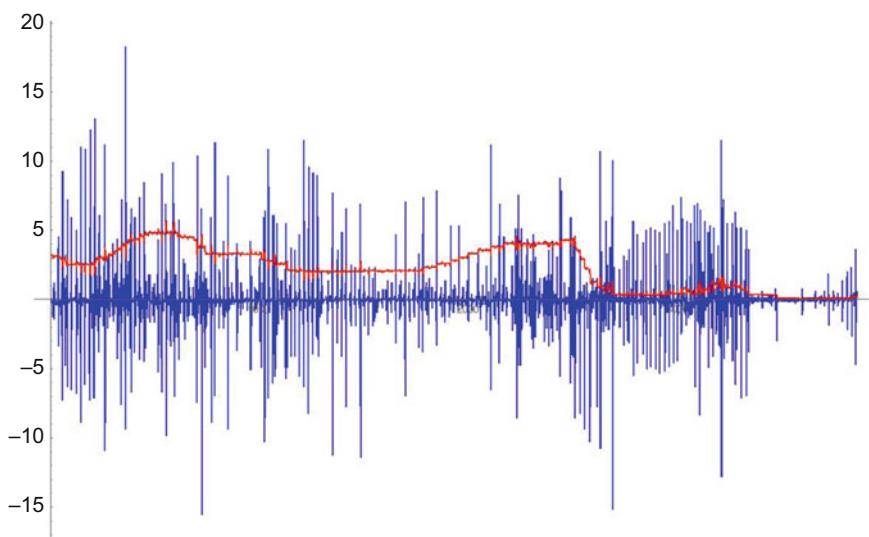


Fig. 3.15 Euro Overnight Eonia Index from 1999 to 2014 (red) compared with normalized daily returns (blue)

We now go a bit further in our heuristic considerations with regard to modelling the Euro Overnight Eonia Index using a mean-reverting Ornstein-Uhlenbeck model (MR-OU) $dS(t) = a \cdot (m - S(t)) \cdot dt + b \cdot dB(t)$. Using a couple of—once more **heuristic!**—considerations, we want to argue an explicit choice of the parameters a , m , and b in the mean-reverting Ornstein-Uhlenbeck model for modelling the Euro Overnight Eonia Index:

- For m , we could simply choose the long-term average of the Eonia Index from 1999 to 2014. Based on these historical data, we get $m \sim 2.25$.
- If the Eonia Index really followed an MR-OU, that is, if $dS(t) \sim a \cdot (2.25 - S(t)) \cdot dt + b \cdot dB(t)$ and if we took the expected value on both sides, we would get

$$E(dS(t)) \sim E(a \cdot (2.25 - S(t)) \cdot dt) + E(b \cdot dB(t)).$$

Based on the historical data, we obtain $E(dS(t)) \sim -0.0008$.

And we have $E(b \cdot dB(t)) = b \cdot E(\sqrt{dt} \cdot w) = 0$, since w is a standard normally distributed random variable. So, we conclude $E(a \cdot (2.25 - S(t)) \cdot dt) \sim -0.0008$

- Let us now again assume that $dS(t) = a \cdot (2.25 - S(t)) \cdot dt + b \cdot dB(t)$. We conclude

$$\begin{aligned} E((dS(t) - a \cdot (2.25 - S(t)) \cdot dt)^2) &= E((b \cdot dB(t))^2) = b^2 \cdot E((\sqrt{dt} \cdot w)^2) \\ &= b^2 \cdot dt \cdot E(w^2) = b^2 \cdot dt \end{aligned}$$

and therefore

$$b = \frac{1}{\sqrt{dt}} \cdot \sqrt{E((dS(t) - a \cdot (2.25 - S(t)) \cdot dt)^2)}.$$

- We now turn to the expected value under the root

$$\begin{aligned} E((dS(t) - a \cdot (2.25 - S(t)) \cdot dt)^2) &= \\ &= E((dS(t))^2) - 2 \cdot a \cdot dt \cdot E(dS(t) \cdot (2.25 - S(t))) \\ &\quad + a^2 \cdot dt^2 \cdot E((2.25 - S(t))^2) \end{aligned}$$

From the historical data, we get, for each of the expected values in the last line,

$$E((dS(t))^2) \sim 0.01315$$

$$E(S(t) \cdot (2.25 - S(t))) \sim 0.006125$$

$$E((2.25 - S(t))^2) \sim 2.206$$

We thus get (by substituting these values and $dt = \frac{1}{250}$)

$$E((dS(t) - a \cdot (2.25 - S(t)) \cdot dt)^2) \sim 0.01315 - a \cdot 0.00005 + a^2 \cdot 0.000035$$

In later modelling applications we will always work with values of a smaller than 5. The last two summands in the last formula are therefore practically negligible, and we obtain $E((dS(t) - a \cdot (2.25 - S(t)).dt)^2) \sim 0.01315$ and thus

$$b \sim \frac{1}{\sqrt{dt}} \cdot \sqrt{E((dS(t) - a \cdot (2.25 - S(t)) \cdot dt)^2)} \sim \mathbf{1.813}.$$

(while being well aware of course that $E(\sqrt{X})$ does not give the same result as $\sqrt{E(X)}$!)

- So, we model $S(t)$ using $dS(t) \sim a \cdot (2.25 - S(t)) \cdot dt + 1.813 \cdot dB(t)$ and just need to estimate a suitable parameter a .

A first idea would be to estimate a using $a \sim E\left(\frac{dS(t)-1.813 \cdot dB(t)}{(2.25-S(t)) \cdot dt}\right)$.

This turns out to be extremely unstable however (since the denominator often takes values very close to 0).

Another approach would be the following: We want to determine a value for a such that in many simulations of the MR-OU model with this a as a parameter, the expected value “ E_{sim} ” and standard deviation “ $S_{t,sim}$ ” of the simulated price match the expected value “ E_r ” and standard deviation “ $S_{t,r}$ ” of the real historical price as closely as possible.

We conducted the search for a suitable a in the following way: We ran a in steps of 0.1 through all values from 0.1 to 5. For each value of a , we ran 100 simulations and used as deviation for each a the average value of $f(a) = (E_{sim} - E_r)^2 + (S_{t,sim} - S_{t,r})^2$.

This resulted in the following graph for $f(a)$ (see Fig. 3.16):

Thus, the picture suggests a parameter choice of $a = \mathbf{1.2}$.

- Based on this calibration approach, the model

$$dS(t) = 1.2 \cdot (2.25 - S(t)) \cdot dt + 1.1813 \cdot dB(t)$$

would seem suited for modelling the Euro Overnight Eonia Index.

In the following (in Fig. 3.17), we are going to show some example paths (each in blue) that were created through simulation using exactly this model over a simulation period of approximately 15 years, and we will compare them with the actual development in the years 1999 to 2014 (red).

We can see here a basically similar behaviour of simulated and historical prices. Yet the (red) historical curve exhibits distinct phases of more constant values than the simulated paths.

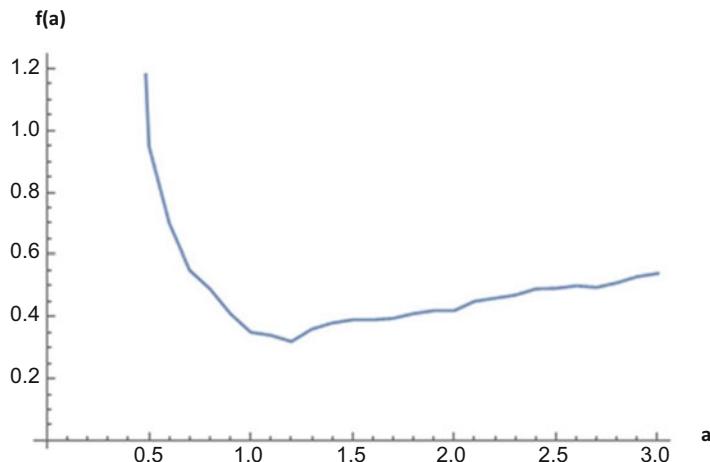


Fig. 3.16 “Simulation error” $f(a)$ as a function of the choice of parameter a

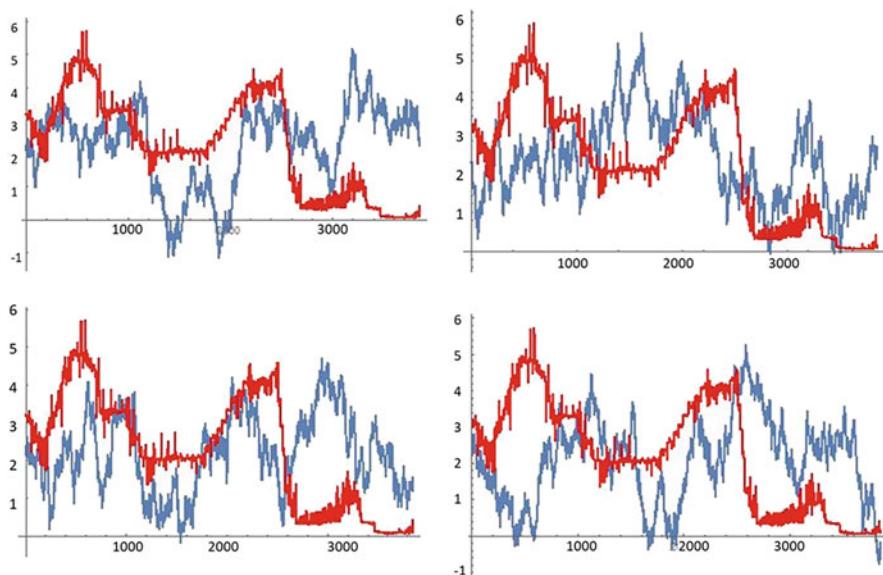


Fig. 3.17 Typical simulation paths (blue) in the model $dS(t) = 1.2 \cdot (2.25 - S(t)) \cdot dt + 1.1813 \cdot dB(t)$ compared with actual Euro Overnight Eonia paths (red)

However, this characteristic definitely cannot be replicated with the MR-OU model (no matter how subtly the parameters are calibrated).

In fact, the first-order autocorrelation of the daily returns of the Euro Overnight Eonia has a clearly positive value of 0.226 in the above period, while the simulated paths consistently show a first-order autocorrelation very close to 0.

Nevertheless, we will essentially use OU model variants to model interest rate curves in the following. We will, however, also see in the following that when it comes to the valuation of interest rate derivatives, what matters is not that we model a real interest rate evolution as closely as possible, but rather that we determine the parameters of a “risk-neutral” evolution as precisely as possible.

3.6 Examples of Interest Rate Derivatives and a Principal Methodology for Pricing such Derivatives

With respect to interest rate derivatives, our goal for this section is to present a basic valuation model for interest rate derivatives. Using the following techniques, we want readers to learn **one** method (among several possible ones) that they can use on their own to price interest rate derivatives and also to understand this valuation approach.

Correct valuation of interest rate derivatives is an extremely complex undertaking, and there are various ways to approach it.

The method we have chosen here is based on short-rate models, and we will limit our discussion to the mean-reverting Vasicek model and the mean-reverting Hull-White model. Both models are variants of an MR-OU process, the use of which we motivated in the previous section:

- We choose a model to replicate a shortest-term interest rate (a “spot rate”) $r(t)$ (e.g. the Euro Overnight Eonia Index).
- This model depends on the specific choice of various parameters that occur in it.
- Other types of interest rates can be expressed essentially as functions of this spot rate.
- It will turn out, however, that these parameters should not be “freely estimated” (e.g. with a view to ensuring that the model is closely calibrated to the historical spot rate performance). Instead, the parameters must be chosen such that they are consistent with the current values of other types of interest rates.
- Once the model has been “correctly” calibrated in this sense, you will also have a “correct model” for all other types of interest rates.
- A derivative on one of these interest rates is then simply the (correctly (!)) discounted expected payoff of that derivative with respect to these “correct” models.

This program is what we will analyse in detail in the following sections.

Now, which types of interest rate derivatives are the most relevant and the most common ones in the market?

- **Termination rights** for loans:

A loan in the amount of 500,000 EUR for a loan period of 20 years at a fixed interest rate of 4% p.a. can be such that either

- (A) 20,000 EUR have to be paid at the end of each year of the loan period, and the principal amount of 500,000 has to be repaid at the end of the period.
- (B) The borrower has the option at the end of each year of the loan period to opt for a final interest payment of 20,000 EUR and immediate repayment of the 500,000 EUR principal (right of termination at any time).

Variant (b) is of course a better deal for the borrower than variant (a). The borrower has a (termination) option in addition to the loan. This is a Bermudan option (can be exercised at various selected time points but not “anytime”). Normally, a loan of type (b) will therefore cost more than a loan of type a) (the additional cost being the “price of the option”). Meaning, normally, the loan interest rate will be somewhat higher for variant (b) than for variant (a). What is the “fair difference” between the two interest rates?

- **Interest caps** for loans:

A loan in the amount of 500,000 EUR for a loan period of 20 years can be set up such that at the beginning of each half year during that period, the 6-month euro Libor is determined and at the end of each half year during that period, interest on the loan becomes payable in the amount of this initially determined 6-month Libor + 0.5% per annum. An interest cap limits (“caps”) the maximum interest to be paid. In the above example for instance, an interest cap could be such that at each semi-annual payment date, no more than a maximum loan interest of 4% p.a. is ever payable. What is the fair price of such a cap?

- **Interest floors** for savings accounts:

A 5-year restricted savings account with a 100,000 EUR initial deposit could be set up such that it accrues interest in the following way: The 1-year euro swap rate is determined at the beginning of each year during that 5-year period. At the end of each year, interest equal to this initially determined swap rate minus 0.5% is credited to the savings account. Yet no less than 0.25% interest is paid out in any case. This threshold is an interest floor that keeps interest rates above a minimum level.

- **A simple interest rate swap:**

Gives the holder the right to exchange, at a time T in the future, a variable rate loan (e.g. with semi-annual interest at 6-month Libor) for a loan with the same loan period and the same principal with fixed semi-annual interest (of, for example, 4% p.a.).

- **Complex interest rate swaps**

In our case studies, we will get to know much more complex real-market versions of interest rate swaps, where, for example, the amount of the “fixed interest rate” depends on various conditions.

Devising the program for the valuation of such products as described above, which requires substantial preliminary work, is what we will dedicate ourselves to in the following sections.

3.7 Basic Concepts of Frictionless Interest Rate Markets: Zero-Coupon Bonds and Interest Rates

In the following, we are going to define various terms related to interest rate markets, some of which we have already presented in earlier contexts, yet now we will do so in a more formal way, using mathematical language:

A **zero-coupon bond with maturity T** is a financial product that guarantees us payment of 1 euro at time T . The **price** at time t of a zero-coupon bond with maturity T is denoted by $p(t, T)$.

We assume a frictionless zero-coupon bond market. That is, we assume the following:

- For every T in the future, there is a zero-coupon bond in the market with maturity T .
- Each of these zero-coupon bonds can be traded as defined earlier for frictionless markets (infinitely and arbitrarily divisible, no transaction costs, both long and short, etc.).
- We assume that there is no credit risk, which means future payment from a zero-coupon bond is always guaranteed and certain. It follows in particular that $p(T, T) = 1$ for all T .
- It is intuitively evident that the price $p(t, T)$ changes only slightly or even “smoothly” (no “salient points” or “jumps”) when t remains fixed and when T increases or decreases slightly.

We assume in a frictionless bond market, the price function $p(t, T)$ is a differentiable function in T . For fixed t and variable $T > t$, we refer to the function $p(t, T)$ as the **bond price curve at time t (“term structure at t ”)**.

(Caveat: A change in t (while T remains fixed) can lead to erratic changes in $p(t, T)$ (even to jumps in the case of suddenly changing market interest rates). Therefore, nothing can be said a priori about the analytical properties of $p(t, T)$ as a function in t and no meaningful assumptions can be made. $p(t, T)$ is a stochastic process when T is fixed. One can nonetheless try, of course, to find a suitable stochastic model for $p(t, T)$ with changing t . The only thing that can be said definitively is $\lim_{t \rightarrow T} p(t, T) = 1$.)

We already asked some explicit questions regarding the price function $p(t, T)$ above:

- What is a suitable stochastic model for $p(t, T)$ with fixed T and varying t ?
- Due to no arbitrage, does there have to be a certain dependence between the values $p(t, T_1)$ and $p(t, T_2)$ for fixed t and different T_1 and T_2 ? And if so, what

kind of dependence? What can be concluded from that with regard to the form of the function $p(t, T)$ as a function of T ?

Zero-coupon bonds allow us to invest cash risk-free at a specific risk-free interest rate for a specific period from now until a point in time T (i.e. for the interval $[0, T]$). To invest an amount of K euros until time T in the future, we use that amount of K EUR to buy $\frac{K}{p(0, T)}$ units of the bond with maturity T . At time T , we will then receive a payout of $\frac{K}{p(0, T)}$ euros. The cash amount of K euros has therefore evolved over T years to $\frac{K}{p(0, T)}$ euros.

If we compare this outcome with the outcome in the case of a **one-time (simple) interest application** or in the case of **continuous interest application** over the course of T years at a per-annum interest rate of x , we get from

$$K \cdot (1 + x \cdot T) = \frac{K}{p(0, T)}$$

the simple per-annum interest rate

$$x = \frac{1}{T} \cdot \left(\frac{1}{p(0, T)} - 1 \right)$$

and from the comparison

$$K \cdot e^{xT} = \frac{K}{p(0, T)}$$

the continuous per-annum interest rate

$$x = \frac{1}{T} \cdot \log \frac{1}{p(0, T)}.$$

Yet zero-coupon bonds also allow us to generate **already now (at time 0)** a risk-free interest rate on cash investments for any time interval $[S, T]$ in the future.

If we want to lock in fixed interest rates already now for an investment of K euros at time S in the future until T in the future (i.e. for the time interval $[S, T]$), we can proceed as follows:

Right now, at time 0, we sell exactly K zero-coupon bond units maturing in S (S Bond) and receive $K \cdot p(0, S)$ euros. With this money, we buy exactly $\frac{K \cdot p(0, S)}{p(0, T)}$ zero-coupon bond units maturing in T (T bond).

At time S , we have to pay out K euros (maturity of the K units of the S Bond), and to make that payment, we use the K euros which we want to invest at time S .

At time T , we receive a payout of $\frac{K \cdot p(0, S)}{p(0, T)}$ euros (maturity of $\frac{K \cdot p(0, S)}{p(0, T)}$ units of the T bond).

Conclusion: At time 0, we already made definitive arrangements for a certain cash amount of K euros to change from K euros at the future point in time S to $\frac{K \cdot p(0, S)}{p(0, T)}$ euros at an even further into the future point in time T .

If we compare this outcome again, as above, with a simple interest outcome $K \cdot (1 + x \cdot (T - S))$ and a continuous interest outcome $K \cdot e^{x \cdot (T - S)}$ for that period, we get the simple interest rate $x = \frac{1}{T - S} \cdot \left(\frac{p(0, S)}{p(0, T)} - 1 \right)$ and the continuous interest rate $x = \frac{1}{T - S} \cdot \log \frac{p(0, S)}{p(0, T)}$.

Of course, from the last formulas, we would then obtain the former ones again if we specifically set $S = 0$. And of course we can argue from the viewpoint of any present time t instead of time 0.

We summarize these considerations in the form of a definition:

Definition 3.2

- The **simple forward interest rate** (hereinafter, **Libor forward interest rate**) for $[S, T]$ from the perspective of t is given by

$$L(t; S, T) := \frac{1}{T - S} \cdot \left(\frac{p(t, S)}{p(t, T)} - 1 \right)$$

- The **continuous forward interest rate** for $[S, T]$ from the perspective of t is given by

$$R(t; S, T) := \frac{1}{T - S} \cdot \log \frac{p(t, S)}{p(t, T)}$$

- The **simple spot interest rate** for $[t, T]$ (hereinafter, **Libor spot interest rate**) for $[t, T]$ is given by

$$L(t, T) := \frac{1}{T - t} \cdot \left(\frac{1}{p(t, T)} - 1 \right)$$

- The **continuous spot interest rate** for $[t, T]$ is given by

$$R(t, T) := \frac{1}{T - t} \cdot \log \frac{1}{p(t, T)}$$

In the continuous versions of the above formulas, we will now consider arbitrarily short periods of time, that is, we let S go to T and thereby obtain the shortest-term

interest rate (short rate) at time T that we can generate from time t . We get

$$\begin{aligned} f(t, T) &:= \lim_{S \rightarrow T} R(t; S, T) = \lim_{S \rightarrow T} \left(\frac{1}{T - S} \cdot \log \frac{p(t, S)}{p(t, T)} \right) = \\ &= - \lim_{S \rightarrow T} \left(\frac{\log p(t, T) - \log p(t, S)}{T - S} \right) = - \frac{\partial \log p(t, T)}{\partial T} = \\ &= - \frac{p'(t, T)}{p(t, T)} \end{aligned}$$

where $p'(t, T)$ denotes the derivative of $p(t, T)$ with respect to T .

Specifically, $r(t) := f(t, t)$ is the shortest-term current interest rate that we can generate at time t using zero-coupon bonds. We summarize these considerations again in a definition.

Definition 3.3

- The **forward spot rate** for the time T from the viewpoint of t is given by

$$f(t, T) := - \frac{\partial \log p(t, T)}{\partial T} = - \frac{p'(t, T)}{p(t, T)}$$

- The **short rate** at time t is given by $r(t) := f(t, t) = - \frac{p'(t, T)}{p(t, T)} \Big|_{T=t}$

If an investor wants to invest money (e.g. 1 euro) risk-free for an indefinite period of time as from now (time 0), that is, in such a way that the money can be withdrawn again at any time without incurring any loss (as is the case with an unrestricted savings account, for example), then that investor can (theoretically) proceed as follows:

At time 0, the investor invests the amount of 1 euro for the “infinitesimally short” time period $[0, dt]$ in a dt bond. By the time dt , the amount of 1 euro has grown to $e^{r(0) \cdot dt}$ euro.

At the time dt , the investor invests the amount of $e^{r(0) \cdot dt}$ euros for the “infinitesimally short” time period $[dt, 2 \cdot dt]$ in $e^{r(0) \cdot dt}$ units of a $2 \cdot dt$ bond. By the time dt , the amount has grown from $e^{r(0) \cdot dt}$ euros to $e^{r(0) \cdot dt + r(dt) \cdot dt}$ euros.

We continue “rolling” in this way until an arbitrary point in time t .

If we represent t in the form $t = M \cdot dt$, then the investment will have grown to $e^{r(0) \cdot dt + r(dt) \cdot dt + r(2 \cdot dt) \cdot dt + \dots + r(M \cdot dt) \cdot dt}$ euros by the time t . For $dt \rightarrow 0$, the “Riemann sum” in the exponent converges to the integral $\int_0^t r(s) ds$, and so by the time t , the invested amount has grown to $B_t := e^{\int_0^t r(s) ds}$.

Definition 3.4

- The process $B_t := e^{\int_0^t r(s) ds}$ is referred to as the *money account process*.

Of course, $B_0 = 1$, and for the derivative of B with respect to t , we have $B'_t = r(t) \cdot B_t$. If a constant value $r(s) = r$ is assumed for $r(s)$, then $B_t = e^{r \cdot t}$.

At this point, we need to introduce a change regarding the notation used so far. It has become customary in the financial mathematics literature to refer to this money account process by B . We will encounter this money account process time and again in the future. However, we have previously used the letter B to denote the Brownian motion. In the financial mathematics literature, a widely established notation for that is W . Therefore, we will also use W to denote the Brownian motion from now on.

Now that we have clarified this notation:

We have $B'_t = r(t) \cdot B_t$. Using the differential notation introduced above, we can also write $dB_t = r(t) \cdot B_t \cdot dt$.

(Although the Brownian motion does not yet figure here, it was necessary to clarify that notation in order to avoid any ambiguity with regard to B .)

Finally, it is important to note the following relationship between bond prices $p(t, T)$ and forward spot rates $f(t, T)$:

Due to the definition of $f(t, u)$ (for the time being, we deliberately choose the parameter u instead of T), we have (as already mentioned in our explanation of $f(t, T)$)

$$f(t, u) = -\frac{\partial \log p(t, u)}{\partial u}$$

We apply the integral $\int_s^T du$ to both sides of this equation, where $t \leq s \leq T$. This gives us

$$\int_s^T f(t, u) du = \int_s^T -\frac{\partial \log p(t, u)}{\partial u} du = \log p(t, s) - \log p(t, T),$$

hence

$$\log p(t, T) = \log p(t, s) - \int_s^T f(t, u) du$$

and thus

$$p(t, T) = p(t, s) \cdot e^{-\int_s^T f(t, u) du}.$$

We will frequently revisit this relationship in the following. Specifically, if we set $s = t$, we get

$$p(t, T) = e^{-\int_t^T f(t, u) du}.$$

Having clarified these essential notations, let us now recall again the basic task that we set out to handle in this section and how we can solve it in principle:

We want to appropriately model all the main basic interest rate and bond price types defined in this chapter (we do not need any other types!), that is, $p(t, T)$, $f(t, T)$, $r(t)$, $L(t, S, T)$, $R(t, S, T)$, etc..

An important insight that follows directly from the definitions: Once we have modelled the bond price $p(t, T)$ for all T , then we can use that to express all other interest rates. Thus, it is completely sufficient to model the bond price $p(t, T)$ for all T .

However, since we can represent each $p(t, T)$ in the form $p(t, T) = e^{-\int_t^T f(t,u)du}$, it would also suffice to model the forward spot rate $f(t, T)$ for all T .

If we now want to model all $p(t, T)$ (or all $f(t, T)$), we have to ask ourselves the following question:

Can we really model the process $p(t, T)$ (or $f(t, T)$) for each T freely? Or must certain relations between the different $p(t, T_1)$ and $p(t, T_2)$ hold here, since otherwise arbitrage would be possible (or because contradictions of other types might occur)? What kind of relations would that be? What do we need to look out for here?

Or we could—and should—also ask ourselves the following question:

Would it suffice to model only a few (or even only *one*) bond prices $p(t, T)$ (or forward spot rates $f(t, T)$), from which all others could then be calculated based on no-arbitrage arguments?

In the extreme case, this question could even be

Would it suffice to model only the short rate $r(t)$ and then calculate all other bond prices (etc.) from that using no-arbitrage arguments?

We will answer these questions in the next but one section.

Before that, we want to derive some more basic concepts and relationships.

3.8 Fixed and Floating Rate Coupon Bonds

We are first going to look at **fixed coupon bonds**.

- With this type of bonds, the bond holder receives coupons in fixed predetermined amounts of $c_1, c_2, c_3, \dots, c_n$ at fixed times $T_1 < T_2 < T_3 < \dots < T_n$.
- At T_n , the bond holder also receives a principal amount of K .
- T_0 (with $T_0 < T_1$) is the initial issue date of the bond.

Such a fixed-coupon bond can obviously be represented by a combination of zero-coupon bonds with n different maturities $T_1, T_2, T_3, \dots, T_n$. It yields exactly the same cash flows as the combination of

c_1 units of T_1 zero-coupon bonds
 c_2 units of T_2 zero-coupon bonds
 ...
 c_{n-1} units of T_{n-1} zero-coupon bonds
 and
 $K + c_n$ units of T_n zero-coupon bonds

Therefore, due to no-arbitrage arguments, the **fair price $p(t)$ of the fixed-coupon bond** for a time t with $T_0 \leq t < T_1$ is given by

$$\begin{aligned} p(t) &= c_1 \cdot p(t, T_1) + c_2 \cdot p(t, T_2) + c_3 \cdot p(t, T_3) + \dots + c_{n-1} \cdot p(t, T_{n-1}) + \\ &\quad + c_n \cdot p(t, T_n) + K \cdot p(t, T_n) = \\ &= \sum_{i=1}^n c_i \cdot p(t, T_i) + K \cdot p(t, T_n). \end{aligned}$$

Usually the coupon amount c_i at time T_i is given as a simple interest rate per annum for the time interval $[T_{i-1}, T_i]$, for example, as interest rate r . This interest rate then applies to the principal K . The coupon payment in this case is $c_i = r \cdot (T_i - T_{i-1}) \cdot K$, and the above formula becomes $p(t) = \sum_{i=1}^n r \cdot (T_i - T_{i-1}) \cdot K \cdot p(t, T_i) + K \cdot p(t, T_n)$.

Also, in most cases, the time intervals between successive coupon payments are of (essentially) equal length, approximately of length δ (e.g. $\delta = 1$, so 1 year, or $\delta = \frac{1}{2}$, so half a year). The actual coupon payment in this case is $c_i = r \cdot \delta \cdot K$, and then the formula for $p(t)$ is

$$p(t) = K \cdot \left(r \cdot \delta \cdot \sum_{i=1}^n p(t, T_i) + p(t, T_n) \right).$$

Coupon bonds with variable coupons (“floating rate bonds”) come in various forms. We will limit our discussion here to the most commonly encountered real-market version.

- With this type of bonds, the bond holder receives coupons in variable (but obviously predefined) amounts of $c_1, c_2, c_3, \dots, c_n$ at fixed times $T_1 < T_2 < T_3 < \dots < T_n$.
- At T_n , the bond holder also receives a principal amount of K .
- Time T_0 (with $T_0 < T_1$) is the initial issue date of the bond.

In the most common version of a floating rate bond, the coupon c_i is determined as follows:

The coupon c_i corresponds (!) to the simple LIBOR forward interest rate $L(T_{i-1}, T_i)$ for the time range between the most recent preceding payment date T_{i-1} and the current payment date T_i . Recall that in financial-mathematical terms,

the fair version of the LIBOR is

$$L(T_{i-1}, T_i) = \frac{1}{T_i - T_{i-1}} \cdot \left(\frac{1}{p(T_{i-1}, T_i)} - 1 \right).$$

The coupon c_i is thus equal to $K \cdot (T_i - T_{i-1}) \cdot L(T_{i-1}, T_i) = K \cdot \left(\frac{1}{p(T_{i-1}, T_i)} - 1 \right)$.

So, the amount of payment at time T_i is already known at time T_{i-1} (but not earlier).

Typical floating rate bonds are, for example, bonds that pay the previous period's (actual (!)) 6-month LIBOR every 6 months.

What is the fair value of such a floating rate bond? Does this fair value even exist? Since the amount of future payments is not known a priori, it is not self-evident that a fair price for this product must necessarily exist.

But if this fair value exists, then let us denote it again by $p(t)$ for a time t with $T_0 \leq t < T_1$.

For the following considerations, we can just set $K = 1$ (otherwise, the fair price $p(t)$ derived below is simply multiplied by K).

The coupon paid in T_i is therefore $\frac{1}{p(T_{i-1}, T_i)} - 1$.

The second part “ -1 ” of this payment corresponds to a short position in a T_i -bond at time t and therefore has the value $-p(t, T_i)$ at time t .

The first part “ $\frac{1}{p(T_{i-1}, T_i)}$ ” of this payment corresponds to a long position in $\frac{1}{p(T_{i-1}, T_i)}$ units of a T_i -bond at time t and therefore has the value $p(t, T_i) \cdot \frac{1}{p(T_{i-1}, T_i)}$ at time t .

Unfortunately, this fact does not help us to determine $p(t)$ since at time t , we do not yet know the value of $p(T_{i-1}, T_i)$.

We will therefore determine the value of the payment in the amount of $\frac{1}{p(T_{i-1}, T_i)}$ at time T_i from the perspective of time t in a different way, namely, by replicating the payment using an alternative procedure:

- To do this, we buy one unit of a T_{i-1} -bond at the price of $p(t, T_{i-1})$ at time t .
- At time T_{i-1} , we receive payment of 1 EUR.
- We use this 1 EUR payment to buy exactly $\frac{1}{p(T_{i-1}, T_i)}$ units of a T_i -bond at time T_{i-1} (at that point, the price of one unit is obviously exactly $p(T_{i-1}, T_i)$).
- At time T_i , we then receive a payout of exactly $\frac{1}{p(T_{i-1}, T_i)}$!

In this way, we already secured payment of exactly $\frac{1}{p(T_{i-1}, T_i)}$ EUR at time T_i , at the price of $p(t, T_{i-1})$ EUR (at time t).

Thus, the value of a payment in the amount of $\frac{1}{p(T_{i-1}, T_i)}$ EUR at time T_i from the perspective of time t has exactly the value $p(t, T_{i-1})$.

In other words, **from the perspective of time t** , the **coupon c_i** that is paid out at time T_i has **the fair value $p(t, T_{i-1}) - p(t, T_i)$** . This holds for $i \geq 2$.

Since $T_0 \leq t < T_1$ is assumed, the first coupon payment in T_1 is already known at time t . It's $c_1 = L(T_0, T_1) = \frac{1}{p(T_0, T_1)} - 1$. At time t , the coupon c_1 thus has the value $p(t, T_1) \cdot \left(\frac{1}{p(T_0, T_1)} - 1 \right)$.

So the fair value of the floating rate bond at time t is the sum of the fair values of all the individual payments (coupons c_i at time T_i and principal 1 at time T_n):

$$\begin{aligned} p(t) &= p(t, T_1) \cdot \left(\frac{1}{p(T_0, T_1)} - 1 \right) + \\ &\quad + \sum_{i=2}^n (p(t, T_{i-1}) - p(t, T_i)) + 1 \cdot p(t, T_n) = \\ &= \frac{p(t, T_1)}{p(T_0, T_1)}. \end{aligned}$$

To summarize,

For any point in time t with $T_0 \leq t < T_1$, the fair price $p(t)$ of the above floating rate bond is given by $p(t) = \frac{p(t, T_1)}{p(T_0, T_1)}$.

Specifically, $p(T_0) = 1$.

3.9 Interest Rate Swaps

We define an interest rate swap as a financial product that provides for an exchange of a fixed coupon bond (with fixed interest rate R) for a floating rate bond (limited here to floating rate bonds as set out above).

Often, the fixed interest rate R of the fixed coupon bond is defined in such a way that the fair price of the swap at time T_0 is 0. We refer to this as the fair interest rate R of a swap.

The fair interest rate R of a swap can be easily determined based on the above preparatory work.

We define the swap such that we will receive variable interest payments and pay fixed interest payments.

And we are again going to assume a principal amount $K = 1$ and, for simplicity (and in line with the reality of the vast majority of swap agreements), fixed periods of length δ between two payment dates.

From our perspective, the cash flow from a swap then consists of payments at times T_i for $i = 1, 2, \dots, n$ of the form $\delta \cdot (L(T_{i-1}, T_i) - R)$.

Repayment of the principal at time T_n takes place in the same form on both sides and therefore balances out, so it does not need to be considered further.

Based on the considerations in the previous chapter, the fair value of each such cash flow $\delta \cdot (L(T_{i-1}, T_i) - R)$ from the perspective of time t is therefore given by $p(t, T_{i-1}) - p(t, T_i) - \delta \cdot R \cdot p(t, T_i)$.

Therefore, for the **fair value $\Pi(t)$ of the swap at a point in time t** with $T_0 \leq t < T_1$, we have

$$\begin{aligned}\Pi(t) &= \sum_{i=2}^n (p(t, T_{i-1}) - p(t, T_i) - \delta \cdot R \cdot p(t, T_i)) + \\ &+ \left(\frac{1}{p(T_0, T_1)} - 1 \right) \cdot p(t, T_1) - \delta \cdot R \cdot p(t, T_1) = \\ &= \frac{p(t, T_1)}{p(T_0, T_1)} - p(t, T_n) - \delta \cdot R \cdot \sum_{i=1}^n p(t, T_i)\end{aligned}$$

Specifically, at time $t = T_0$:

$$\Pi(T_0) = 1 - p(T_0, T_n) - \delta \cdot R \cdot \sum_{i=1}^n p(T_0, T_i).$$

The **fair interest rate R of a swap at time T_0 —the “swap rate” for this swap**—is thus given by the equation

$$0 = \Pi(T_0) = 1 - p(T_0, T_n) - \delta \cdot R \cdot \sum_{i=1}^n p(T_0, T_i) \text{ as}$$

$$R = \frac{1 - p(T_0, T_n)}{\delta \cdot \sum_{i=1}^n p(T_0, T_i)}.$$

3.10 Valuation of Bond Prices and Interest Rate Derivatives in a Short-Rate Approach

Out of several possible approaches to modelling a variety of interest rate types and to valuing interest rate derivatives, we will, as previously announced, only present the short-rate approach here.

Thus, in the following, we are going to assume that we have chosen a model for a short-rate $r(t)$. How we choose that model is a question we will discuss in the next chapter, since it is not relevant here. We will however assume that $r(t)$ is an Ito process, and we denote the parameters of this Ito process by

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

Remember the basic question that we asked earlier: Does one such model suffice, based on the no-arbitrage principle, to determine the fair prices $p(t, T)$ of all T bonds at time t and thus also the fair prices at time t of all interest rate derivatives expiring in T ?

We also answered the question back then, and the answer is “No”. It will be necessary to adopt a model for at least one more product.

How bond prices and interest rate derivative prices are determined under these premises will be derived below in outlines only. We will do this purely formally and not worry about the legitimacy of the following transformations, the existence of the following derivatives, and the computations with infinitesimal quantities or applications of the Ito formulas. The mathematically exact proof for this approach would require an extensive and intricate argumentation and a technically perfect basis in the theory of stochastic processes and stochastic differential equations and exceeds the scope and intentions of this book project.

For this heuristic derivation, we assume two bonds with different maturities S and T . We denote their fair prices at time t by $p(t, T)$ and $p(t, S)$.

Furthermore, we recall that we could also invest in the money account process $B_t = e^{\int_0^t r(s)ds}$, that is, in $dB_t = r(t) \cdot B(t) \cdot dt$!

Our fundamental approach is that the respective bond prices are functions of the current time t , the maturity T , and the current value of the spot rate $r(t)$.

We therefore write $p(t, T) := F(t, r(t), T)$ and $p(t, S) := F(t, r(t), S)$.

Since T and S are fixed values that will not change in the following, we will use the following notations below:

$F^T(t, r) := F(t, r, T)$ and $F^S(t, r) := F(t, r, S)$

By F_t^T and F_r^S and by F_r^T and F_r^S , we denote the derivatives of F^T and of F^S with respect to the first variable t and the second variable r .

F^T and F^S are functions of t and of the Ito process $r(t)$ and because of the Ito formula they are themselves Ito processes. The representation of F^T and F^S as Ito processes is also obtained by applying the Ito formula. That's what we are going to do now. First, we recall again the Ito representation of $r(t)$, which we will use in the following in its shorthand representation $dr(t) = \mu \cdot dt + \sigma \cdot dW(t)$. (Always keeping in mind, however, that μ and σ are generally not constant, but may depend on t and $r(t)$.)

Using the Ito formula, we obtain

$$dF^T = \left(F_t^T + \mu \cdot F_r^T + \frac{\sigma^2}{2} F_{rr}^T \right) dt + \sigma \cdot F_r^T \cdot dW(t)$$

and

$$dF^S = \left(F_t^S + \mu \cdot F_r^S + \frac{\sigma^2}{2} F_{rr}^S \right) dt + \sigma \cdot F_r^S \cdot dW(t)$$

(continued)

We rearrange and rename these two representations somewhat and get

$$dF^T = F^T \cdot \alpha_T \cdot dt + F^T \cdot \sigma_T \cdot dW(t) \text{ and } dF^S = F^S \cdot \alpha_S \cdot dt + F^S \cdot \sigma_S \cdot dW(t), \quad (3.8)$$

where

$$\alpha_T = \frac{F_t^T + \mu \cdot F_r^T + \frac{\sigma^2}{2} \cdot F_{rr}^T}{F^T} \text{ and } \alpha_S = \frac{F_t^S + \mu \cdot F_r^S + \frac{\sigma^2}{2} \cdot F_{rr}^S}{F^S} \text{ and} \quad (3.9)$$

$$\sigma_T = \frac{\sigma \cdot F_r^T}{F^T} \text{ resp. } \sigma_S = \frac{\sigma \cdot F_r^S}{F^S}. \quad (3.10)$$

Now, we create a dynamic portfolio V from the T bond and the S bond. Let this portfolio be defined such that at any point in time t , we denote our current total assets in the portfolio by $V(t)$ and invest $u(t) \cdot V(t)$ of these assets in the T bond and invest the remainder, hence $(1 - u(t)) \cdot V(t)$, in the S -bond.

This means

At any time t , we hold exactly (in shorthand) $\frac{u \cdot V}{F^T}$ units of the T bond and $\frac{(1-u) \cdot V}{F^S}$ units of the S bond.

The trading strategy for creating this portfolio is, in principle, such that we always invest only as much money in the T bond and in the S bond as there is value V in the portfolio. The **strategy** is therefore **self-financing**. No additional money is added to or withdrawn from the portfolio at any time during the strategy's life.

We will determine the specific value for $u(t)$, that is, the exact definition of the dynamic portfolio, at a later time.

Right now, we ask ourselves: How does the value $V(t)$ of the portfolio change in an infinitesimally small time interval from t to $t + dt$? In other words, what does $dV(t)$ look like? The answer is simple:

The proportion invested in the T bond changes by $\frac{u \cdot V}{F^T}$ times dF^T , and the proportion invested in the S bond changes by $\frac{(1-u) \cdot V}{F^S}$ times dF^S .

So we have

$$dV = \frac{u \cdot V}{F^T} \cdot dF^T + \frac{(1-u) \cdot V}{F^S} \cdot dF^S = V \cdot \left(u \cdot \frac{dF^T}{F^T} + (1-u) \cdot \frac{dF^S}{F^S} \right) \quad (3.11)$$

From Formula (3.8), we know that

$$\frac{dF^T}{F^T} = \alpha_T \cdot dt + \sigma_T \cdot dW(t) \text{ and } \frac{dF^S}{F^S} = \alpha_S \cdot dt + \sigma_S \cdot dW(t).$$

(continued)

Substituting this in Formula (3.11), we get

$$\begin{aligned} dV &= V \cdot (u \cdot (\alpha_T \cdot dt + \sigma_T \cdot dW(t)) + (1-u) \cdot (\alpha_S \cdot dt + \sigma_S \cdot dW(t))) = \\ &= V \cdot ((u \cdot \alpha_T + (1-u) \cdot \alpha_S) \cdot dt + \\ &\quad + (u \cdot \sigma_T + (1-u) \cdot \sigma_S) \cdot dW(t)). \end{aligned} \quad (3.12)$$

And now we make an explicit choice for the special form of our dynamic portfolio. That is, we choose u explicitly. And we choose u in such a way that the random component in the portfolio's dynamics (i.e. the part that is governed by the Brownian motion) is eliminated. That means:

$$u \cdot \sigma_T + (1-u) \cdot \sigma_S = 0 \text{ and therefore } u = \frac{\sigma_S}{\sigma_S - \sigma_T} \text{ and } 1-u = \frac{-\sigma_T}{\sigma_S - \sigma_T}.$$

Substituting this choice for u in the Formula (3.12) for dV , we get

$$dV = V \cdot \left(\frac{\sigma_S}{\sigma_S - \sigma_T} \cdot \alpha_T - \frac{\sigma_T}{\sigma_S - \sigma_T} \cdot \alpha_S \right) \cdot dt.$$

The dynamics $V(t)$ of this type of portfolio V is thus deterministic.

So, in this environment, we now have *two* ways to invest our money deterministically, namely, the money account process B and portfolio V !

Now, the thing here is that (this is a very important principle and one that will be used frequently in the following)

Whenever there are two possibilities P and Q for a deterministic investment with dynamics of the form $dP(t) := p(t) \cdot P(t) \cdot dt$ and $dQ(t) := q(t) \cdot Q(t) \cdot dt$, it is a requirement that $p(t) = q(t)$!

The (heuristic) rationale for this is

If at a time t , we had, for example, $p(t) > q(t)$, then we could short the asset Q for a tiny period from t to $t+dt$ and invest the money received in asset P and thus make a safe profit without investing any money, and that contradicts the no-arbitrage principle.

Applying this principle to our two deterministic investments B and V , we get

$$\frac{\sigma_S}{\sigma_S - \sigma_T} \cdot \alpha_T - \frac{\sigma_T}{\sigma_S - \sigma_T} \cdot \alpha_S = r$$

(continued)

We rearrange this equation somewhat

$$\begin{aligned} \frac{\sigma_S}{\sigma_S - \sigma_T} \cdot \alpha_T - \frac{\sigma_T}{\sigma_S - \sigma_T} \cdot \alpha_S &= r \\ \Leftrightarrow \sigma_S \cdot \alpha_T - \sigma_T \cdot \alpha_S &= r \cdot \sigma_S - r \cdot \sigma_T \\ \Leftrightarrow \sigma_S \cdot (\alpha_T - r) &= \sigma_T \cdot (\alpha_S - r) \\ \Leftrightarrow \frac{\alpha_T(t) - r(t)}{\sigma_T(t)} &= \frac{\alpha_S(t) - r(t)}{\sigma_S(t)} \end{aligned}$$

Now, this is a highly remarkable equation: Both sides have variables that are strikingly reminiscent of a Sharpe ratio (trend term minus interest rate divided by volatility). The left side shows exactly this expression for the T bond, and the right side shows exactly this expression for the S bond. We refer to this expression (trend term minus interest rate divided by volatility) as the bond's "**market price of risk**", and we denote it by $\lambda_T(t)$ resp. by $\lambda_S(t)$.

The essential insight from the last formula is the following: Regardless of how S or T was chosen, the market price of risk at time t will always have the same value, irrespective of the value of S and of T !

So this market price of risk is a function of the entire bond market and will always have the same value regardless of what bond we are analysing. We can thus write this market price of risk function simply as a function $\lambda(t)$ of time (and irrespective of S or T).

To summarize, if we analyse a frictionless bond market and have modelled a short-rate $r(t)$ in some way as an Ito process, then there is a function $\lambda(t)$ (which we refer to as the "market price of risk" of that bond market) so that we get

If, for any U bond in our bond market, we denote the dynamics of its price movements by $dF^U = F^U \cdot \alpha_U \cdot dt + F^U \cdot \sigma_U \cdot dW(t)$, then we always have $\lambda(t) = \frac{\alpha_U(t) - r(t)}{\sigma_U(t)}$ regardless of the value of U .

For everything that follows, for determining fair bond prices just as for determining fair prices of interest rate derivatives in this bond market, we will need to know $\lambda(t)$ (at least implicitly).

Yet if we assume that we know how to model the U bond for *one* specific U , then we already know the value of $\lambda(t)$.

Let us now assume that we know the market price of risk $\lambda(t)$:

Then for any T

$$\frac{\alpha_T(t) - r(t)}{\sigma_T(t)} = \lambda(t)$$

(continued)

Here, we substitute $\alpha_T(t)$ and $\sigma_T(t)$ by their representation from Formulas (3.9) and (3.10); thus, $\alpha_T = \frac{F_t^T + \mu \cdot F_r^T + \frac{\sigma^2}{2} \cdot F_{rr}^T}{F^T}$ and $\sigma_T = \frac{\sigma \cdot F_r^T}{F^T}$.

If we then rearrange that expression, we get the following equation for a T bond's fair value F^T :

$$F_t^T + (\mu - \lambda \cdot \sigma) \cdot F_r^T + \frac{\sigma^2}{2} \cdot F_{rr}^T - r \cdot F^T = 0.$$

We note in addition that for $F^T(T, r)$, that is, for the T bond's price at time T , we always have $F^T(T, r) = 1$ regardless of the value r (and we recall that the values μ and σ that occur here are the parameters occurring in the modelling of $r(t)$).

Thus, we have derived a partial differential equation with a final value condition for the fair price of any arbitrary T bond. This partial differential equation can be solved (using, for example, the **Feynman-Kac formula**, which we will address later), and that solution is again a representation in a form that we have already come to know, namely, the expected value of the bond's discounted payoff under an artificial probability measure. We give the solution here and then discuss it again from this angle. We have

$$F(t, r, T) = E \left(e^{-\int_t^T \check{r}(s) ds} \mid \check{r}(t) = r \right),$$

where $\check{r}(s)$ follows the dynamics $d\check{r}(s) = (\mu - \lambda \cdot \sigma) \cdot ds + \sigma \cdot dW(s)$.

So, the expected value is now taken from $e^{-\int_t^T \check{r}(s) ds} = e^{-\int_t^T \check{r}(s) ds} \times 1$, that is, from the T bond's discounted payoff 1. The random component in this expression whose expected value we want to determine here is, however, mainly the discount factor and not a potentially varying payoff!

What is more, here too, the expected value is not taken with respect to the actual short-rate $r(s)$, as one might (naively) assume, but yet again with respect to a slightly modified "artificial interest rate" $\check{r}(s)$, whose dynamics differs from the dynamics of $r(s)$ only in that instead of the trend term μ we now need to set the trend term $(\mu - \lambda \cdot \sigma)$.

Using an analogous (heuristic) approach, we could also have derived a **more general result for any derivatives on the short rate $r(t)$** . We want to formulate this result in full detail here. Our above result on bond prices is then a simple special case of this general result. We have the following:

Theorem 3.5 *Let $r(t)$ be a short rate with modelling $dr(t) = \mu(t, r(t)) \cdot dt + \sigma(t, r(t)) \cdot dW(t)$. Let $\lambda(t)$ be the market price of risk of the relevant bond market. Let D be a derivative on $r(t)$ with expiration in T and payoff function $\Phi(r(T))$.*

Then the fair price process $F(t, r)$ (the fair price of the derivative at time t if $r(t) = r$) satisfies the partial differential equation:

$$F_t + (\mu - \lambda \cdot \sigma) \cdot F_r + \frac{\sigma^2}{2} \cdot F_{rr} - r \cdot F = 0.$$

with boundary condition $F(T, r) = \Phi(r)$.

Furthermore, F has the explicit representation:

$$F(t, r) = E \left(e^{-\int_t^T \check{r}(s) ds} \cdot \Phi(\check{r}(T)) \middle| \check{r}(t) = r \right),$$

where $\check{r}(s)$ follows the dynamics $d\check{r}(s) = (\mu - \lambda \cdot \sigma) \cdot ds + \sigma \cdot dW(s)$.

With this result, we can now value any derivatives on $r(t)$. The prerequisite is that we have chosen a specific modelling for $r(t)$ and that we have calculated (or suitably estimated) the market price of risk, that is, $\lambda(t)$, from a bond's market prices. We will discuss two specific models for $r(t)$ in the following sections, with a particular focus on how $\lambda(t)$ can be determined (approximately) in these cases. (Note that we will in fact not determine $\lambda(t)$ there, but will proceed directly to determining the total risk-neutral drift term $(\mu - \lambda \cdot \sigma)$, which is what is actually needed.)

The expected value of the integral that yields the fair value can then be determined explicitly in some cases and can in any case be calculated approximately by Monte Carlo simulation.

Note: We are dealing with derivatives on the spot rate r here and not derivatives on bond prices!

Derivatives with expiration T on \tilde{T} bonds, where the maturity \tilde{T} of those bonds is of course greater than T , cannot be valued directly using this theorem:

For example, let Ψ be the payoff function of such a derivative. Then the payoff of the derivative is given by $\Psi(p(T, \tilde{T}))$.

Examples of such derivatives on a bond price would be (for a fixed h , for instance, $h = \frac{1}{2}$) a Libor interest rate $L(T, T+h) = \frac{1}{h} \cdot \left(\frac{1}{p(T, T+h)} - 1 \right)$ or a call option on such a Libor interest rate, that is, a derivative with payoff $\max(L(T, T+h) - K, 0) = \max\left(\frac{1}{h} \cdot \left(\frac{1}{p(T, T+h)} - 1 \right) - K, 0\right)$ or a put option on such a Libor interest rate, that is, a derivative with payoff $\max(K - L(T, T+h), 0) = \max\left(K - \frac{1}{h} \cdot \left(\frac{1}{p(T, T+h)} - 1 \right), 0\right)$.

Such a call option is called a **caplet** and such a put option is called a **floorlet**.

It would be tempting here to value this derivative in the following way: Given the formula for fair bond prices that we derived above, we have

$$\Psi(p(T, \tilde{T})) = \Psi \left(E \left(e^{-\int_T^{\tilde{T}} \check{r}(s) ds} \right) \middle| \check{r}(T) = r(T) \right). \quad (3.13)$$

Thus, the payoff would be a function of $r(T)$, and we could then value the derivative using the above valuation formula. Hence

$$F(t, r) = E \left(e^{-\int_t^T \check{r}(s) ds} \cdot \Psi \left(E \left(e^{-\int_T^{\tilde{T}} \check{r}(s) ds} \right) \right) \middle| \check{r}(t) = r \right). \quad (3.14)$$

We could now approximate this value by simulating paths for $\check{r}(s)$ from t to \tilde{T} using Monte Carlo simulation.

There is a flaw in this line of reasoning, however: In Formula (3.13), \check{r} is calculated from the perspective of time T , using the parameters $(\mu - \lambda \cdot \sigma)(T)$ and $\sigma(T)$ that apply at this time, while in Formula (3.14), it is calculated from the perspective of time t with the parameters $(\mu - \lambda \cdot \sigma)(t)$ and $\sigma(t)$ that apply at that time. One could possibly assume that the volatility σ remains constant and thus unchanged over the entire period. However, this assumption is no longer permissible for $\mu - \lambda \cdot \sigma$.

The correct valuation of bond derivatives is a more complicated matter, and we cannot delve deeper into the details here.

We will only do the following: We are going to present two basic models for short rates $r(t)$, namely, the Vasicek model and the Hull-White model, and show how the risk-neutral parameters can be estimated in these models and then how the associated bond prices can be explicitly calculated in these models.

In a subsequent section, we will give just the formulas, without proof, for valuing caplets and floorlets in these two short-rate models (and hence the formulas for valuing interest rate caps and interest rate floors).

3.11 The Mean-Reverting Vasicek Model and the Hull-White Model for the Short Rate

The two models for modelling the short rate, which we will discuss in more details below, are the **mean-reverting Vasicek model**

$$dr(t) = (b - a \cdot r(t))dt + \sigma \cdot dW(t)$$

with constant parameters a , b , and σ

as well as the **Hull-White model** (extended Vasicek)

$$dr(t) = (\theta(t) - a \cdot r(t))dt + \sigma \cdot dW(t)$$

with the time-dependent parameter $\theta(t)$.

In the Hull-White model, the parameters a and σ are also occasionally considered functions of time, yet we will limit our presentation in the following to constant a and σ .

To model the short rate $r(t)$, the parameters a , b , and σ (or a , $\theta(t)$, and σ) have to be calibrated in such a way that $r(t)$ matches a real interest rate curve as closely as

possible. In contrast, to calculate fair bond prices in the interest rate environment of this short rate, the risk-neutral version of the parameters must be calibrated in such a way that the short rate provides the “best possible compatibility with the current bond and derivative prices in the market”. We will see what exactly that means in the next section.

We recall from the previous section: If $r(t)$ has the “real” modelling

$dr(t) = (b - a \cdot r(t))dt + \sigma \cdot dW(t)$ or $dr(t) = (\theta(t) - a \cdot r(t))dt + \sigma \cdot dW(t)$, then the risk-neutral version $\check{r}(s)$ has a modelling of the form $d\check{r}(s) = (\mu - \lambda \cdot \sigma) \cdot ds + \sigma \cdot dW(s)$.

The diffusion term σ thus remains unchanged.

The drift term, however, can change completely: $(\mu - \lambda \cdot \sigma)$ no longer has to be of the basic structure $B - A \cdot r(t)$ or $\psi(t) - A \cdot r(t)$ with some parameters A, B , and ψ , but can be of a completely different form.

It is, however, possible to estimate the parameter σ for a “real model” in exactly the same way as for a “risk-neutral” model.

This can be done as we suggested and demonstrated in Sect. 3.6. (After all, the Vasicek model (and the Hull-White model) is nothing other than a mean-reverting Ornstein-Uhlenbeck process (or a somewhat generalized OU process).)

Calibration of the drift parameter in a “real model” can be done in the same way as given in Sect. 3.6.

What remains to be done is to calibrate the drift parameter $\mu - \lambda \cdot \sigma$ in the “risk-neutral” model. Yet the task here is *not*:

“Suppose that the **real model** has a modelling of the form

$dr(t) = (b - a \cdot r(t))dt + \sigma \cdot dW(t)$ or $dr(t) = (\theta(t) - a \cdot r(t))dt + \sigma \cdot dW(t)$, what then is the drift term $\mu - \lambda \cdot \sigma$ of the associated risk-neutral model?”

But rather:

“Suppose that the **risk-neutral model** has a modelling of the form

$dr(t) = (b - a \cdot r(t))dt + \sigma \cdot dW(t)$ or $dr(t) = (\theta(t) - a \cdot r(t))dt + \sigma \cdot dW(t)$, what then are the parameters $a, b, \theta(t)$ of the risk-neutral model?”

Before we can answer this question, we need a somewhat technical auxiliary mathematical result, which we will derive in the next section and formulate in such a way that readers who do not wish to read the derivation can understand it and use it later.

3.12 Affine Model Structures of Bond Prices

We recall the notation $p(t, T)$ for the price of a T bond at time t and our assumption that this bond price is a function $F(t, r(t), T)$, that is, a function that is dependent on the current point in time t , the current value of the short rate $r(t)$, and, of course, the bond’s maturity T .

If the short-rate $r(t)$ were constant, that is, equal to a fixed constant r for all t (and if this were already known at time t), then we would have $p(t, T) = F(t, r(t), T) = e^{-r \cdot (T-t)}$. Thus, also for arbitrary stochastic short rates $r(t)$, it would be an obvious

hypothesis that $p(t, T)$ could be of a similar form, giving us, for example,

$$p(t, T) = F(t, r(t), T) = e^{A(t, T) - B(t, T) \cdot r(t)}$$

with some functions A and B that depend on t and T .

If this is indeed the case, then we say bond prices have an **affine model structure**.

As we will see, it is indeed true of many short-rate models (including, in particular, the Vasicek model and the Hull-White model) that bond prices actually have an affine model structure. And to show that, we will now prove the following theorem:

Theorem 3.6 *If the risk-neutral short-rate $r(t)$ has the modelling*

$$dr(t) = \mu(t, r(t)) \cdot dt + \sigma(t, r(t)) \cdot dW(t)$$

where $\mu(t, r(t))$ and $\sigma(t, r(t))^2$ are linear functions in $r(t)$, thus

$$\mu(t, r(t)) = \alpha(t) \cdot r(t) + \beta(t) \text{ und } \sigma(t, r(t))^2 = \gamma(t) \cdot r(t) + \delta(t),$$

then the associated fair bond prices $p(t, T)$ have an affine model structure $e^{A(t, T) - B(t, T) \cdot r(t)}$, and the functions $A(t, T)$ and $B(t, T)$ appearing in it can be calculated from the following differential equation system (A_t and B_t denote the derivative of A and B with respect to the variable t):

$$\begin{aligned} B_t(t, T) + \alpha(t)B(t, T) - \frac{1}{2} \cdot \gamma(t) \cdot (B(t, T))^2 &= -1 \quad \text{with } B(T, T) = 0 \\ A_t(t, T) = \beta(t)B(t, T) - \frac{1}{2} \cdot \delta(t) \cdot (B(t, T))^2 &\quad \text{with } A(T, T) = 0. \end{aligned}$$

Proof We know from the previous section that the bond price $p(t, T) = F^T(t, r)$ is the solution of the differential equation $F_t^T + \mu \cdot F_r^T + \frac{\sigma^2}{2} \cdot F_{rr}^T - r \cdot F^T = 0$ with $F^T(T, r) = 1$.

Thus, we only need to show that $F^T(t, r) := e^{A(t, T) - B(t, T) \cdot r}$ with the functions $A(t, T)$ and $B(t, T)$ that are given by the differential equation system is indeed the solution of the differential equation.

To this end, we substitute for $F^T(t, r)$ the function $e^{A(t, T) - B(t, T) \cdot r}$ in the differential equation and then get, since

$$F_t^T = e^{A(t, T) - B(t, T) \cdot r} \cdot (A_t(t, T) - B_t(t, T) \cdot r)$$

and

$$F_r^T = -e^{A(t, T) - B(t, T) \cdot r} \cdot B(t, T)$$

(continued)

and

$$F_{rr}^T = e^{A(t, T) - B(t, T) \cdot r} \cdot (B(t, T))^2$$

the following equation:

$$A_t(t, T) - (1 + B_t(t, T)) \cdot r - \mu \cdot B(t, T) + \frac{\sigma^2}{2} \cdot (B(t, T))^2 = 0$$

and

$$A(T, T) = B(T, T) = 0.$$

We now substitute for μ and for σ^2 , arrange the terms of the resulting equation by r , and obtain

$$\begin{aligned} A_t(t, T) - \beta(t) \cdot B(t, T) + \frac{1}{2} \cdot \delta(t) \cdot (B(t, T))^2 - (1 + B_t(t, T) + \alpha(t) \cdot \\ B(t, T) - \frac{1}{2} \cdot \gamma(t) \cdot (B(t, T))^2) \cdot r = 0 \end{aligned}$$

If A and B satisfy the differential equation system given in the theorem, then this last equation is indeed satisfied, and the proof has thus been completed. \square

3.13 Bond Prices in the Vasicek Model and Calibration in the Vasicek Model

In the Vasicek model $dr(t) = (b - a \cdot r(t))dt + \sigma \cdot dW(t)$, both the drift term and the diffusion term are linear in r .

Using the notation of the previous section, $\alpha = -a$, $\beta = b$, $\gamma = 0$, and $\delta = \sigma$. Thus, bond prices have an affine model structure $e^{A(t, T) - B(t, T) \cdot r}$, where $A(t, T)$ and $B(t, T)$ must satisfy the following differential equation system:

$$B_t(t, T) - a \cdot B(t, T) + 1 = 0 \text{ with } B(T, T) = 0$$

$$A_t(t, T) - b \cdot B(t, T) + \frac{\sigma^2}{2} \cdot (B(t, T))^2 = 0 \text{ with } A(T, T) = 0.$$

This differential equation system is very easily solved. First, we solve the first equation, which is an ordinary differential equation only in the variable B and get

$$B(t, T) = \frac{1}{a} \cdot \left(1 - e^{-a \cdot (T-t)}\right).$$

We substitute this B in the second equation and then only need to integrate it with respect to t to get the solution for A , namely,

$$A(t, T) = \frac{1}{a^2} \left(\frac{1}{a} \cdot (1 - e^{-a \cdot (T-t)}) - T + t \right) \cdot \left(ab - \frac{\sigma^2}{2} \right) - \\ - \frac{\sigma^2}{4a} \cdot \left(\frac{1}{a} \cdot (1 - e^{-a \cdot (T-t)}) \right)^2.$$

Calibration

The only question now is how to model the drift term ($b - a \cdot r(t)$) “correctly” in this risk-neutral model for $r(t)$, that is, how to choose the constant parameters a and b . (The diffusion term σ has already been modelled as assumed.)

In principle, the question can be answered as follows:

We already have a formula for the bond prices $p(t, T)$ in this model, namely,

$$p(t, T) = e^{A(t, T) - B(t, T) \cdot r}, \text{ with the above functions } A \text{ and } B.$$

With the formula for $p(t, T)$, we then of course have formulas for all other types of interest rates too, such as for forward spot rates or for swap rates.

Yet for the functions A and B , we still have to estimate the parameters a and b .

On the other hand, we know the actual current (time 0) T bond prices and forward spot rates and swap rates, since the market provides that information.

Therefore, at least right now, at time 0, the formula $p(t, T) = e^{A(t, T) - B(t, T) \cdot r}$ should correspond for any T to the **actual current T bond prices**, which we denote here by $p^*(T)$. So the conditions $p^*(T) = e^{A(0, T) - B(0, T) \cdot r(0)}$ should (essentially) be satisfied for any T . This gives an equation with the unknowns a and b for each T for which a bond price is available at the moment.

If only few bond prices are available, one could apply the same approach to forward rates or swap rates instead of bonds (or in addition to bonds) and compare the current actual forward rates or the current actual swap rates with the theoretical prices given by the formulas.

In the following, however, we will simply stick to the requirement that the real current bond prices be represented as well as possible by the formulas, i.e. that a and b are determined in such a way that the equation $p^*(T) = e^{A(0, T) - B(0, T) \cdot r(0)}$ is at least approximately satisfied for as many T as possible.

As a result, however, we get many conditional equations (one for every T) with only two variables a and b to satisfy. We are therefore dealing with an overdetermined system of equations. In general, there will be no such choice of these two variables a and b as to satisfy all of these equations.

One can therefore only try to satisfy the equations at least approximately. How to do this is a question that is analysed in depth in the case study in Volume III Section 3.4, and we refer readers directly to that case study.

Once a choice has been made in some way for a and b , the risk-neutral model we are looking for is parameterized and can be used, for example, to value derivatives

on the short rate or (as we will see later) to value caplets and floorlets (and thus so-called caps and floors).

3.14 Bond Prices in the Hull-White Model and Calibration in the Hull-White Model

In the Hull-White model $dr(t) = (\theta(t) - a \cdot r(t))dt + \sigma \cdot dW(t)$, the drift term and the diffusion term are also both linear in r . With the notations from the previous section, we have $\alpha = -a$, $\beta = \theta(t)$, $\gamma = 0$, and $\delta = \sigma$. So, the bond prices have again an affine model structure $e^{A(t,T)-B(t,T)\cdot r(t)}$, where $A(t, T)$ and $B(t, T)$ must satisfy the following differential equation system:

$$\begin{aligned} B_t(t, T) - a \cdot B(t, T) + 1 &= 0 \text{ with } B(T, T) = 0 \\ A_t(t, T) - \theta(t) \cdot B(t, T) + \frac{\sigma^2}{2} \cdot (B(t, T))^2 &= 0 \text{ with } A(T, T) = 0. \end{aligned}$$

This differential equation system is again very easily solved. First, we solve the first equation, which is an ordinary differential equation only in the variable B , and obviously get the same solution as in the Vasicek model:

$$B(t, T) = \frac{1}{a} \cdot \left(1 - e^{-a \cdot (T-t)}\right). \quad (3.15)$$

We substitute this B in the second equation and then only need to integrate it with respect to t to get the solution for A :

$$A(t, T) = \int_t^T \left(\frac{\sigma^2}{2} \cdot (B(s, T))^2 - \theta(s) \cdot B(s, T) \right) ds \quad (3.16)$$

(where, in the integrand, we substitute the above solution for $B(s, T)$).

Calibration

Now, we have to calibrate the parameters of the model again. In the Hull-White model, the usual procedure is as follows: We first estimate σ and a like we did in Sect. 3.6, such that we get a realistic interest rate structure.

The parameter $\theta(s)$, on the other hand, is again calibrated to match bond prices on the market at time 0 as closely as possible. This is done as follows:

We take the logarithm of the bond price, then differentiate with respect to T , and obtain—as we saw above—the **forward spot rate** for time T from the perspective of t :

$$\begin{aligned} f(t, T) &= -\frac{\partial \log p(t, T)}{\partial T} = -\frac{\partial}{\partial T} \left(\log \left(e^{A(t, T) - B(t, T) \cdot r(t)} \right) \right) = \\ &= B_T(t, T) \cdot r(t) - A_T(t, T). \end{aligned}$$

The derivative of B with respect to T is obtained by simple differentiation of (3.15): $B_T(t, T) = e^{-a \cdot (T-t)}$.

To differentiate A with respect to T , we use Formula (3.16) and the following rule for differentiating integrals:

$$\frac{\partial}{\partial T} \left(\int_t^T f(s, T) ds \right) = \int_t^T f_T(s, T) ds + f(T, T)$$

The correctness of this differentiation rule can be shown schematically (with “infinitesimally small values” h) as follows:

$$\begin{aligned} \frac{\partial}{\partial T} \left(\int_t^T f(s, T) ds \right) &\approx \frac{\int_t^{T+h} f(s, T+h) ds - \int_t^T f(s, T) ds}{h} = \\ &= \frac{\int_t^{T+h} f(s, T+h) ds - \int_t^{T+h} f(s, T) ds}{h} + \\ &\quad + \frac{\int_t^{T+h} f(s, T) ds - \int_t^T f(s, T) ds}{h} = \\ &= \int_t^{T+h} \frac{(f(s, T+h) - f(s, T))}{h} ds + \\ &\quad + \frac{\int_t^T f(s, T) ds}{h} \approx \\ &\approx \int_t^T f_T(s, T) ds + f(T, T) \end{aligned}$$

In our case (see Formula (3.16)), the integrand $f(s, T)$ is exactly

$$\left(\frac{\sigma^2}{2} \cdot (B(s, T))^2 - \theta(s) \cdot B(s, T) \right).$$

Therefore and because $B(T, T) = 0$, we get $f(T, T) = 0$, and thus

$$\begin{aligned} A_T(t, T) &= \frac{\partial}{\partial T} \left(\int_t^T f(s, T) ds \right) = \int_t^T f_T(s, T) ds = \\ &= \int_t^T \frac{\partial}{\partial T} \left(\frac{\sigma^2}{2} \cdot (B(s, T))^2 \right) ds - \int_t^T \theta(s) \cdot B_T(s, T) ds. \end{aligned}$$

Substituting for $B(s, T) = \frac{1}{a} \cdot (1 - e^{-a \cdot (T-s)})$ and for $B_T(s, T) = e^{-a \cdot (T-s)}$ and calculating the first integral are now an easy exercise, and we get

$$A_T(t, T) = \frac{\sigma^2}{2a^2} \left(1 - e^{-aT} \right)^2 - \int_t^T \theta(s) \cdot e^{-a(T-s)} ds$$

We use these results specifically for the current time $t = 0$ and obtain all in all for the forward spot rate:

$$\begin{aligned} f(0, T) &= \frac{\partial \log p(0, T)}{\partial T} = B_T(0, T) \cdot r(0) - A_T(0, T) = \\ &= e^{-a \cdot (T-t)} \cdot r(0) + \int_t^T \theta(s) \cdot e^{-a(T-s)} ds - \frac{\sigma^2}{2a^2} \left(1 - e^{-aT} \right)^2. \end{aligned}$$

We now assume forward spot rates observed on the market at time 0, that is, $f^*(0, T)$, for any time T in the future.

In order for the theoretical forward spot rates to match the observed forward spot rates, the parameter $\theta(s)$ has to be determined for all (or at least next to all) existing T in such a way that

$$f^*(0, T) = e^{-a \cdot (T-t)} \cdot r(0) + \int_t^T \theta(s) \cdot e^{-a(T-s)} ds - \frac{\sigma^2}{2a^2} \left(1 - e^{-aT} \right)^2 \quad (3.17)$$

is satisfied for all T . If we assume that the values $f^*(0, T)$ are indeed available for all T and have such a structure that the derivative of f^* with respect to T , that is, $f_T^*(0, T)$ exists, then Eq. (3.17) can indeed be solved. The solution is, as can be easily verified,

$$\theta(s) = f_T^*(0, s) + a \cdot f^*(0, s) + \frac{\sigma^2}{a} (1 - e^{-a \cdot s})$$

This choice of $\theta(s)$ in the Hull-White model is then used to create an exact replication of the instantaneous bond structure and the instantaneous forward spot rate structure.

Substituting $\theta(s)$ in $A(t, T)$ and $A(t, T)$ and $B(t, T)$ in $p(t, T) = e^{A(t, T) - B(t, T) \cdot r(t)}$ and rearranging, we then obtain the following result:

Theorem 3.7 *Using the Hull-White model with the parameters a , σ , and $\theta(s)$ as chosen above, the following holds for the T bond prices $p(t, T)$ at time t :*

$$p(t, T) = \frac{p^*(0, T)}{p^*(0, t)} \cdot e^{\left(B(t, T) \cdot f^*(0, t) - \frac{\sigma^2}{4a} \cdot B^2(t, T) \cdot (1 - e^{-2at}) - B(t, T) \cdot r(t)\right)},$$

where $B(t, T) = \frac{1}{a} \cdot (1 - e^{-a \cdot (T-t)})$.

(In particular, $p(0, T) = p^*(0, T)$ for all T .)

With the specific risk-neutral parameterization of the Hull-White model as performed above, we can of course—based on the valuation formula deduced in Sect. 3.10—also value derivatives on the short rate $r(t)$ in the Hull-White model!

As noted above, we now have the tools to value derivatives on the short rate, but not yet the tools to value derivatives on bond prices or other interest rates (that are directly related to bond prices, such as Libor rates).

Here, we can only present—without proof—the procedure for valuing call or put options on bond prices (and related interest rates), which is what we will do in the following section. We will then subsequently use that procedure to value interest rate caps, interest rate floors, and interest rate collars.

3.15 Valuation and Put-Call Parity of Call and Put Options on Bond Prices

As announced, we give here without proof the formulas for the price of a call option and the price of a put option expiring in T on S -bond prices $p(t, S)$, where the bond's time to maturity S is obviously greater than the option's time to expiration T . We assume that the short-rate $r(t)$ moves according to a Vasicek or a Hull-White model.

Theorem 3.8 *Let the short-rate $r(t)$ follow a Vasicek model or a Hull-White model. For a call option with price $C(t)$, strike K , and expiration in T on an S -bond's price $p(t, S)$ that is consistent with the dynamics of the short rate, we have*

$$C(t) = p(t, S) \cdot \mathcal{N}(d) - p(t, T) \cdot K \cdot \mathcal{N}\left(d - \sum\right)$$

where

$$d = \frac{1}{\sum} \cdot \log \left\{ \frac{p(t, S)}{p(t, T) \cdot K} \right\} + \frac{1}{2} \cdot \sum$$

and

$$\sum = \frac{1}{a} \cdot \left\{ 1 - e^{-a \cdot (S-T)} \right\} \cdot \sqrt{\frac{\sigma^2}{2a} \cdot \left\{ 1 - e^{-2a \cdot (T-t)} \right\}}.$$

The formula for the price $P(t)$ of an analogous put option follows directly from the following version of the **put-call parity equation**:

Holding a long position in one of the above put options and a short position in one of the above call options and one unit of an S -bond yields the following payoff at time T (the date when the options expire): $\max(0, K - p(T, S)) - \max(0, p(T, S) - K) + p(T, S) = K$.

The value of this combination at time t before expiration T is therefore $K \cdot p(t, T)$. And therefore, for any $t < T$

$$P(t) - C(t) + p(t, S) = K \cdot p(t, T)$$

Consequently,

$$\begin{aligned} P(t) &= K \cdot p(t, T) - p(t, S) + C(t) = \\ &= K \cdot p(t, T) - p(t, S) + p(t, S) \cdot \mathcal{N}(d) - p(t, T) \cdot K \cdot \mathcal{N}\left(d - \sum\right) = \\ &= K \cdot p(t, T) \cdot \left(1 - \mathcal{N}\left(d - \sum\right)\right) + p(t, S) \cdot (\mathcal{N}(d) - 1) = \\ &= K \cdot p(t, T) \cdot \mathcal{N}(-d) - p(t, S) \cdot \mathcal{N}(-d) \end{aligned}$$

3.16 Valuation of Caplets and Floorlets (as Well as Interest Rate Caps and Interest Rate Floors)

Probably the most commonly traded interest rate derivatives are interest rate caps and interest rate floors (as well as interest rate collars).

These are (collections of) call or put options, yet not on bond prices but usually on a Libor spot rate.

The **usual form** of an **interest rate cap** is as follows:

An interest rate cap is typically added to a loan with variable interest rates and serves to ensure that the interest rates on that loan, although variable, cannot exceed a certain limit K (the strike of that interest rate cap).

If we assume (for simplicity) that the loan amount is 1 and that the agreed loan term is from 0 to T and that the interest payments at times $T_1 < T_2 < T_3 < \dots < T_n = T$ are to be made, as is usually the case, at the time-weighted Libor rate $(T_i - T_{i-1}) \cdot L(T_{i-1}, T_i)$ at time T_i , then a corresponding interest rate cap would be as follows:

The interest rate cap has the same term as the loan, from 0 to T .

We denote the strike price of the interest rate cap by K .

At every point in time T_i , the holder of the interest rate cap receives a payment in the amount of

$$(T_i - T_{i-1}) \cdot \max(0, L(T_{i-1}, T_i) - K).$$

The interest rate cap thus guarantees an interest rate at each payment date of no more than the maximum limit K .

So the total interest rate cap consists of n units of individual components $\text{Cap}_1, \text{Cap}_2, \dots, \text{Cap}_n$. These individual components are referred to as **caplets**.

The fair price of the cap Cap is equal to the sum of the prices of all of these caplets.

However, the underlying asset of such a caplet is not a bond price but a Libor spot rate, and the caplet is also not a call option in the conventional sense, for the following reason: The derivative's payment at expiration T_i is not based on the underlying asset's value at time T_i but on its value at an earlier time T_{i-1} .

We will therefore have to reformulate and reinterpret the payoff function $(T_i - T_{i-1}) \cdot \max(0, L(T_{i-1}, T_i) - K)$ at time T_i somewhat, before we can use the call price formula or the put price formula for bond options from the previous section for the valuation of one such caplet.

For this purpose, we denote the time interval $T_i - T_{i-1}$ by δ and recall that

$$L(T_{i-1}, T_i) = \frac{1}{T_i - T_{i-1}} \cdot \left(\frac{1}{p(T_{i-1}, T_i)} - 1 \right) = \frac{1}{\delta} \cdot \left(\frac{1}{p} - 1 \right),$$

if, for brevity, we denote $p(T_{i-1}, T_i)$ by p .

The payment amount at time T_i is then

$$\begin{aligned} \delta \cdot \max(0, L(T_{i-1}, T_i) - K) &= \delta \cdot \max\left(0, \frac{1}{\delta} \cdot \left(\frac{1}{p} - 1 \right) - K\right) = \\ &= \max\left(0, \frac{1}{p} - 1 - \delta \cdot K\right) = \frac{1 + \delta \cdot K}{p} \cdot \max\left(0, \frac{1}{1 + \delta \cdot K} - p\right) \\ &= \frac{R}{p} \cdot \max\left(0, \frac{1}{R} - p\right), \end{aligned}$$

if we denote $1 + \delta \cdot K$ by R .

However, from the perspective of time T_{i-1} , a payment at time T_i in the amount of $\frac{R}{p} \cdot \max\left(0, \frac{1}{R} - p\right) = \frac{R}{p(T_{i-1}, T_i)} \cdot \max\left(0, \frac{1}{R} - p(T_{i-1}, T_i)\right)$, which of course is already known in T_{i-1} , has the value $R \cdot \max\left(0, \frac{1}{R} - p(T_{i-1}, T_i)\right)$.

Thus, the caplet defined above is equivalent to $R = 1 + \delta \cdot K$ units of a put option with strike $\frac{1}{R}$ and expiration T_{i-1} on the T_i bond price.

The put price formula given in the last section can therefore be applied directly to the valuation of this caplet and thus of the entire cap.

We summarize this in a result:

Theorem 3.9 *Let the short rate $r(t)$ follow a Vasicek model or a Hull-White model. We consider Libor spot rates consistent with this short rate.*

For the price $Capl(t)$ of the caplet defined above with strike K and expiration T_i , we have

$$Capl(t) = p(t, T_{i-1}) \cdot \mathcal{N} \left(\sum -d \right) - (1 + \delta \cdot K) \cdot p(t, T_i) \cdot \mathcal{N}(-d)$$

where

$$d = \frac{1}{\sum} \cdot \log \left\{ \frac{p(t, T_i) \cdot (1 + \delta \cdot K)}{p(t, T_{i-1})} \right\} + \frac{1}{2} \cdot \sum$$

and

$$\sum = \frac{1}{a} \cdot \{ 1 - e^{-a \cdot \delta} \} \cdot \sqrt{\frac{\sigma^2}{2a} \cdot \{ 1 - e^{-2a \cdot (T_{i-1} - t)} \}}.$$

The **usual form** of an **interest rate floor** is as follows:

An interest rate floor is typically added to a savings account or a similar type of investment with variable interest rates and serves to ensure that the interest rates on that investment, although variable, cannot fall below a certain limit K (the strike of that interest rate floor).

If we again assume an investment in the amount of 1 and that the agreed term for this investment is from 0 to T and that the interest payments at times $T_1 < T_2 < T_3 < \dots < T_n = T$ are to be made, as is usually the case, at the time-weighted Libor rate $(T_i - T_{i-1}) \cdot L(T_{i-1}, T_i)$ at time T_i , then the interest rate floor would be as follows:

The interest rate floor has the same term as the loan, from 0 to T .

We denote the strike price of the interest rate floor by K .

At every point in time T_i , the holder of the interest rate floor receives a payment in the amount of

$$(T_i - T_{i-1}) \cdot \max(0, K - L(T_{i-1}, T_i)).$$

The interest rate floor thus guarantees an interest rate at each payment date of at least K .

The total interest rate floor thus consists of n units of individual components $Floorl_1, Floorl_2, \dots, Floorl_n$. These individual components are referred to as **floorlets**.

The fair price of the floor is equal to the sum of the prices of all of these floorlets.

The procedure for valuing one of these floorlets expiring in T_i is analogous to the one we used for the caplets, yet noting that this floorlet is equivalent to $1 + \delta \cdot K$ units of call options with expiration T_{i-1} and strike $\frac{1}{1+\delta \cdot K}$ on the T_i bond.

The call price formula given in the last section can thus be applied directly to the valuation of this floorlet and thus of the entire floor.

We summarize this in a result:

Theorem 3.10 *Let the short rate $r(t)$ follow a Vasicek model or a Hull-White model. We consider Libor spot rates that are consistent with this short rate. For the price $Floorl(t)$ of the floorlet defined above with strike K and expiration T_i , we have*

$$Floorl(t) = (1 + \delta \cdot K) \cdot p(t, T_i) \cdot \mathcal{N}(d) - p(t, T_{i-1}) \cdot \mathcal{N}\left(d - \sum\right)$$

where

$$d = \frac{1}{\sum} \cdot \log \left\{ \frac{p(t, T_i) \cdot (1 + \delta \cdot K)}{p(t, T_{i-1})} \right\} + \frac{1}{2} \cdot \sum$$

and

$$\sum = \frac{1}{a} \cdot \{1 - e^{-a \cdot \delta}\} \cdot \sqrt{\frac{\sigma^2}{2a} \cdot \{1 - e^{-2a \cdot (T_{i-1} - t)}\}}.$$

An **interest rate collar** is a product that guarantees that the interest rate on a loan or the interest rate on an investment, as the case may be, will always be between two values K_1 and K_2 ($K_1 < K_2$) at any payment date.

A borrower will buy a cap with strike K_2 and sell a floor with strike K_1 for this purpose: “buyer of the collar”:

Someone holding some type of investment will take the exact opposite position to ensure a minimum interest limit: “seller of the collar”:

Often, the strikes K_1 and K_2 of an interest rate collar are set precisely so that the price of the collar is 0 (meaning that the price of the cap is equal to the price of the floor).

The price of an interest rate collar can be calculated directly with the above formulas.

From the outset of this chapter, we have been pointing out that the correct modelling and valuation of interest rate derivatives is a complex task and that, for the time being, we can only impart a minimum of basic knowledge on this subject here.

Before we conclude this chapter, we want to provide an **important addendum** drawing on the basic knowledge of stochastic analysis acquired in this chapter so far:

3.17 The Black-Scholes Differential Equation

With the same nonchalance that we applied in Sect. 3.10, where we arrived at an approach to determining fair bond prices without concerning ourselves with technical details and difficulties, we are now going to use our rudimentary and heuristic knowledge of stochastic analysis to give a “quick” alternative proof (rather than a rigorous full proof) of the general (i.e. of an even more general version than the earlier formulations of the) Black-Scholes formula, based on stochastic analysis.

In addition, we will also expand our stochastic toolbox somewhat by adding the notion of the stochastic (Ito) integral. Here again, we will just give a heuristic introduction to this notion and its basic properties rather than a rigorous derivation.

In Volume I Section 4, we gave an elementary proof of the Black-Scholes formula by approximating the Wiener model with a binomial model. We will now proceed as follows: First, by analogy with the calculations performed in 3.10, we are going to derive the Black-Scholes differential equation.

The solution of this differential equation would then be the general Black-Scholes formula that we have already discussed.

To solve this Black-Scholes differential equation, however, we will need another tool of stochastic analysis: the Feynman-Kac formula. In order to derive that formula (and also for later purposes), we need another basic concept of stochastic analysis, namely, the stochastic (Ito) integral mentioned above. Thus, in Sect. 3.18, we will heuristically derive the stochastic integral and its basic properties; in Sect. 3.19, we will discuss the concepts of “conditional expectation” and “martingale”; and in Sect. 3.20, we will derive the Feynman-Kac formula.

Finally, in Sect. 3.21, we will reap the rewards, by solving the Black-Scholes equation and thus obtaining the Black-Scholes formula.

Now Let's Derive the Black-Scholes Differential Equation

The starting situation is

- We have an underlying asset S whose price $S(t)$ in the time range $[0, T]$ follows a stochastic differential equation $dS(t) = \mu(t, S(t)) \cdot S(t)dt + \sigma(t, S(t)) \cdot S(t)dW(t)$ with given initial value $S(0)$ and with deterministic functions μ and σ .
(In our earlier derivation over the binomial model, we limited ourselves to the special case $dS(t) = \mu \cdot S(t) \cdot dt + \sigma \cdot S(t) \cdot dW(t)$ with constant parameters μ and σ !)
- Again, we assume a fixed interest rate r for the entire time interval $[0, T]$.
- We have a risk-free investment opportunity, that is, a bond B with price $B(t)$ which then follows the differential equation $.$
- Furthermore, we have a European plain vanilla derivative D on the underlying asset S expiring in T , which is defined by the payoff function Φ . The payoff from the derivative at time T is thus given by $\Phi(S(T))$.

- Our interest is to find the fair price $F(t, S(t))$ of the derivative at time t under the assumption that the price of the underlying at time t is $S(t)$.

What is clear upfront is the requirement that $F(T, S) = \Phi(S)$.

We now proceed in basically the same way as in Sect. 3.10: We are going to run a trading strategy V in which we dynamically trade the two products S and D . The trading strategy will be “self-financing”. We will construct this trading strategy in such a way that its performance is deterministic (i.e. no longer dependent on any random components). Then we can again conclude that the performance of this deterministic trading strategy V has to be the same as that of the (deterministic) bond B , since otherwise there would be arbitrage opportunities. From this conclusion then follows the Black-Scholes equation.

In the following, we will often just write μ or σ for short instead of the functions $\mu(t, S(t))$ or $\sigma(t, S(t))$.

The Ito formula can be applied to the price function $F(t, S(t))$ as a function of the Ito process $S(t)$ (assuming that F is sufficiently differentiable) and thus obtain the dynamics of this price function (F_t and F_s denote the derivative of F with respect to the first and the second variable, respectively):

$$dF = \left(F_t + \mu \cdot s F_s + \frac{(\sigma s)^2}{2} F_{ss} \right) dt + \sigma \cdot s \cdot F_s \cdot dW(t)$$

Rearranging and renaming this representation somewhat, we get

$$dF = F \cdot \alpha \cdot dt + F \cdot \beta \cdot dW(t) \quad (3.18)$$

where

$$\alpha = \frac{F_t + \mu \cdot s \cdot F_s + \frac{(\sigma \cdot s)^2}{2} \cdot F_{ss}}{F} \quad \text{and} \quad \beta = \frac{\sigma \cdot s \cdot F_s}{F}. \quad (3.19)$$

We already know the dynamics of S , namely (in shorthand notation),

$$dS = \mu \cdot s \cdot dt + \sigma \cdot s \cdot dW(t) \quad (3.20)$$

Now, the following argument will look very familiar to you—it is an almost verbatim restatement of the one given in Sect. 3.10:

We create a dynamic portfolio V from the underlying asset S and from the derivative D .

Let this portfolio be defined such that at any point in time t we denote our current total assets in the portfolio by $V(t)$ and invest $u(t) \cdot V(t)$ of these assets in the underlying and invest the remainder, hence $(1 - u(t)) \cdot V(t)$, in the derivative D .

Meaning, at any time t , we hold (in shorthand notation) exactly $\frac{u \cdot V}{S}$ units of the underlying asset S and $\frac{(1-u) \cdot V}{F}$ units of the derivative D .

So the trading strategy is, in principle, such that we always invest only as much money in the underlying asset and the derivative as there is value V in the portfolio. The strategy is thus self-financing. No additional money is added to or withdrawn from the portfolio at any time during the strategy's life.

We will determine the specific value for $u(t)$, that is, the exact definition of the dynamic portfolio, at a later time.

First, we ask ourselves: How does the value $V(t)$ of the portfolio change in an infinitesimally small time interval from t to $t + dt$? In other words, what does $dV(t)$ look like? The answer is simple:

The proportion invested in the underlying asset S changes by $\frac{u \cdot V}{S}$ times dS , and the proportion invested in the derivative D changes by $\frac{(1-u) \cdot V}{F}$ times dF . So we have

$$dV = \frac{u \cdot V}{S} \cdot dS + \frac{(1-u) \cdot V}{F} \cdot dF = V \cdot \left(u \cdot \frac{dS}{S} + (1-u) \cdot \frac{dF}{F} \right) \quad (3.21)$$

From Formula (3.18) and (3.20), we know that

$$dF = F \cdot \alpha \cdot dt + F \cdot \beta \cdot dW(t) \text{ and } dS = \mu \cdot S \cdot dt + \sigma \cdot S \cdot dW(t)$$

Substituting this in Formula (3.21), we get

$$\begin{aligned} dV &= V \cdot (u(\mu \cdot dt + \sigma \cdot dW(t)) + (1-u) \cdot (\alpha \cdot dt + \beta \cdot dW(t))) \\ &= V \cdot ((u\mu + (1-u)\alpha) \cdot dt + (u\sigma + (1-u)\beta) \cdot dW(t)). \end{aligned} \quad (3.22)$$

And now we make an explicit choice for the particular form of our dynamic portfolio. That is, we choose u explicitly. And we choose u in such a way that the random component in the portfolio's dynamics (i.e. the part that is governed by the Brownian motion) is eliminated. Thus

$$u \cdot \sigma + (1-u) \cdot \beta = 0 \text{ and therefore } u = \frac{\beta}{\beta - \sigma} \text{ und } 1-u = \frac{-\sigma}{\beta - \sigma}.$$

Substituting this choice for u in Formula (3.22) for dV , we get

$$dV = V \cdot \left(\frac{\beta\mu - \sigma\alpha}{\beta - \sigma} \right) \cdot dt.$$

Now, we use a fact again that is based on the no-arbitrage principle, as explained earlier: Whenever there are two possibilities P and Q for a deterministic investment with dynamics of the form $dP(t) := p(t) \cdot P(t) \cdot dt$ resp. $dQ(t) := q(t) \cdot Q(t) \cdot dt$, then we must have $p(t) = q(t)$!

Since we now have the bond B with the dynamics $dB(t) = r \cdot B(t) \cdot dt$ and the trading strategy V with the dynamics $dV = V \cdot \left(\frac{\beta\mu - \sigma\alpha}{\beta - \sigma} \right) \cdot dt$, we consequently get

$$r = \frac{\beta\mu - \sigma\alpha}{\beta - \sigma}$$

Here, we substitute $\alpha = \frac{F_t + \mu \cdot S \cdot F_s + \frac{(\sigma \cdot S)^2}{2} \cdot F_{ss}}{F}$ and $\beta = \frac{\sigma \cdot S \cdot F_s}{F}$ (see Formula (3.19)), solve the fractions, and get

$$F_t + r \cdot S \cdot F_s + \frac{(\sigma \cdot S)^2}{2} \cdot F_{ss} = r \cdot F$$

Together with the boundary condition $F(T, S) = \Phi(S)$, this is exactly the Black-Scholes equation. To summarize

Theorem 3.11 (Black-Scholes Equation) *Let S be an underlying asset whose price $S(t)$ in the time range $[0, T]$ follows a stochastic differential equation $dS(t) = \mu(t, S(t)) \cdot S(t)dt + \sigma(t, S(t)) \cdot S(t)dW(t)$ with given initial value $S(0)$ and with deterministic functions μ and σ (with no other payments or costs associated with S .)*

Let D be a European plain vanilla derivative D on the underlying asset S expiring in T , defined by the payoff function Φ .

Let $F(t, s)$ be the function that describes the fair price process of the derivative D for $t \in [0, T]$ and $s > 0$. F then satisfies the partial differential equation:

$$F_t(t, s) + r \cdot s \cdot F_s(t, s) + \frac{(\sigma(t, s) \cdot s)^2}{2} \cdot F_{ss}(t, s) = r \cdot F(t, s)$$

with boundary condition $F(T, s) = \Phi(s)$. Here, r denotes the risk-free interest rate on $[0, T]$.

To solve this partial differential equation, we now need the Feynman-Kac formula. And to derive that formula, we need the concept of the stochastic integral.

3.18 The Stochastic Ito Integral: Heuristic Explanation and Basic Properties

Imagine that following a certain trading strategy, you would like to invest in a stock S for a certain period of time, say, from now (time 0) until time T .

For the stock S , we know that it moves according to a Wiener model, that is, it follows a stochastic differential equation (SDE) of the form $dS(t) = \mu \cdot S(t) \cdot dt + \sigma \cdot S(t) \cdot dW(t)$.

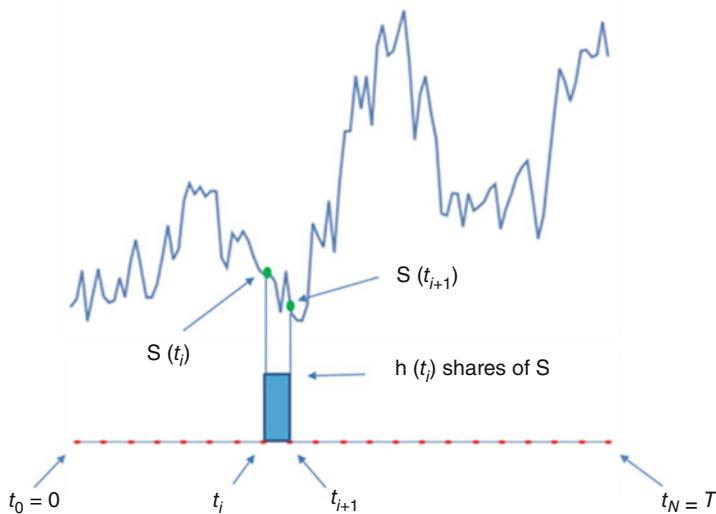


Fig. 3.18 Illustration of the trading strategy

We want our trading strategy (we call it H) to be quite dynamic. Meaning, we adjust our portfolio at very short time intervals. More precisely, we divide the time interval $[0, T]$ into N very short time intervals of length Δt and then adjust our portfolio at the beginning and at the end of each of these time intervals.

For $i = 0, 1, 2, \dots, N$, we denote by t_i the time $i \cdot \Delta t$.

Trades are made at times t_i with $i = 0, 1, 2, \dots, N-1$ and always in such a way that in every t_i within the time range $[t_i, t_{i+1}]$, we always hold exactly $h(t_i)$ shares of the stock. The trading strategy H is thus defined by this function h .

This situation is illustrated in Fig. 3.18.

Now, what is the profit (or loss) we make over the course of this strategy?

Well, the profit in the time range from t_i to t_{i+1} is obviously $S(t_{i+1}) - S(t_i)$ per share of the stock, and since we hold $h(t_i)$ shares, the profit is $h(t_i) \cdot (S(t_{i+1}) - S(t_i))$. The total profit G of the strategy is thus given by

$$G = \sum_{i=0}^{N-1} h(t_i) \cdot (S(t_{i+1}) - S(t_i)).$$

We now recall the dynamics that $S(t)$ follows, namely, $dS(t) = \mu \cdot S(t) \cdot dt + \sigma \cdot S(t) \cdot dW(t)$. Therefore, since the distance Δt between t_i and t_{i+1} is very small,

$$S(t_{i+1}) - S(t_i) \approx \mu \cdot S(t_i) \cdot (t_{i+1} - t_i) + \sigma \cdot S(t_i) \cdot (W(t_{i+1}) - W(t_i))$$

Substituting this in the profit function formula, we get

$$G = \sum_{i=0}^{N-1} h(t_i) \cdot \mu \cdot S(t_i) \cdot (t_{i+1} - t_i) + \sum_{i=0}^{N-1} h(t_i) \cdot \sigma \cdot S(t_i) \cdot (W(t_{i+1}) - W(t_i)).$$

Now, we let the distances Δt between the trading times converge to 0, that is, we transition to a continuous trading strategy. And here we assume some knowledge of basic analysis:

The first of the two sums in the last formula is an ordinary Riemann sum. As Δt goes to 0, this sum converges to the ordinary Riemann integral

$$\int_0^T h(t) \cdot \mu \cdot S(t) dt.$$

We now take a closer look at the second sum $\sum_{i=0}^{N-1} h(t_i) \cdot \sigma \cdot S(t_i) \cdot (W(t_{i+1}) - W(t_i))$. To get a better overview, we first aggregate the expression $h(t_i) \cdot \sigma \cdot S(t_i)$ into the value of a function f at t_i , that is, $f(t_i) := h(t_i) \cdot \sigma \cdot S(t_i)$. The second sum then has the form

$$\sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i)).$$

This sum also bears a certain similarity to a Riemann sum, except that now instead of $(t_{i+1} - t_i)$ we have the difference $(W(t_{i+1}) - W(t_i))$ of a Brownian motion W at two very close points in time.

It is in fact the case that if certain conditions as to the function f (which we will discuss a bit later) are satisfied, this sum too converges to a specific value. We denote this value by $\int_0^T f(t) dW(t)$ and it is called the **“Ito integral”** of f .

Thus, for the profit function of the continuous trading strategy, we have

$$G = \int_0^T h(t) \cdot \mu \cdot S(t) dt + \int_0^T h(t) \cdot \sigma \cdot S(t) dW(t)$$

Some remarks on this derivation and on the notion of the Ito integral are in order:

- We emphasize again that this motivation and sudden installation of the concept of the Ito integral has happened in a negligently superficial manner and is not meant to be more than really just an aid to give readers a rough idea of the meaning of this concept, enabling them to work with this concept in an elementary way. The following additional explanations, some of which are again too simplistic, are also to be seen in this light.
- What kind of mathematical object is this function h that defines the trading strategy H ? Typically, at time 0, we do not yet know which specific strategy we

are going to execute at a future time t . In general, the actual transactions defined by $h(t)$ will rather depend on how the underlying asset's price has performed up to this point and on how successful the trading strategy has been up to the point in time t .

$h(t)$ is thus a random variable, and the entire trading strategy H , which is given by $(h(t))_{t \in [0, T]}$, is a stochastic process!

- For any point in time $t \in [0, T]$, $h(t)$ is likely to depend on the performance of $S(u)$ for $u \in [0, t]$. $h(t)$ cannot depend on the values of $S(u)$ for $u > t$, since at the time t we do not yet know the values $S(u)$ for $u > t$.

Thus, if $S(t)$ follows an Ito process of the form $dS(u) = \alpha(u, S(u)) \cdot du + \beta(u, S(u)) \cdot dW(u)$ on $[0, T]$, then $h(t)$ depends on the evolution of the Brownian motion on $[0, t]$.

If $h(t)$ is indeed dependent for each t on the Brownian motion on $[0, t]$ only, that is, if it is a deterministic function of time t and of $W(u)$ for $u \in [0, t]$, then we say **h is adapted to the Brownian motion on $[0, T]$** .

Incidentally, in this terminology, $h(t)$ may well depend on $h(u)$ for $u < t$ (and thus on the strategy's performance so far).

- The **Ito integral** $\int_0^T f(t) \cdot dW(t)$ is obviously always a **random variable**, even if f is a deterministic function. The actual value of this integral depends on how the Brownian motion $W(t)$ has evolved on $[0, T]$.
- If f is a deterministic function, then the Ito integral $\int_0^T f(t)dW(t)$ exists under the condition that for any $t > 0$, there exists the ordinary Riemann integral $\int_0^T f^2(t)dt$ and has a finite value.
- If f is a stochastic process, then the following two conditions guarantee the existence of the Ito integral $\int_0^T f(t)dW(t)$:
 - f is adapted to the Brownian motion.
 - For any $t > 0$, there exists the ordinary Riemann integral $\int_0^T E[f^2(s)]ds$, and it has a finite value. "E" here denotes the expected value.
- The Ito integral and the conventional Riemann integral both share the usual addition and linearity properties. That is,
 - For $a < b < c$, we have

$$\int_a^c f(s)dW(s) = \int_a^b f(s)dW(s) + \int_b^c f(s)dW(s).$$

- For real α and β and for stochastic processes f and g we have

$$\int_0^t (\alpha \cdot f(s) + \beta \cdot g(s))dW(s) = \alpha \cdot \int_0^t f(s)dW(s) + \beta \cdot \int_0^t g(s)dW(s).$$

- So the Ito integral $\int_0^T f(t)dW(t)$ is a random variable. The first two questions to ask when faced with a random variable are as follows: What is the expected value and what is the variance of this random variable?

We will answer these two questions in the next two items below by taking a step back again and recalling that the Ito integral (if we reverse its derivation) can be approximated arbitrarily well by sums of the form $\sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i))$.

Before calculating the expectation and variance of the Ito integral, we recall two basic properties of the Brownian motion W , which we are going to need below:

- (i) For any two time points $0 < t < s$, we have that $W(s) - W(t)$ is independent of all $W(u)$ with $u \leq t$.
- (ii) $W(s) - W(t)$ is always an $N(0, s - t)$ -distributed random variable. Therefore, $E(W(s) - W(t)) = 0$ und $V(W(s) - W(t)) = E((W(s) - W(t))^2) = s - t$.

- Regarding the expected value of the Ito integral:

$$\begin{aligned} E\left(\int_0^T f(t)dW(t)\right) &\approx E\left(\sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i))\right) \\ &= \sum_{i=0}^{N-1} E(f(t_i) \cdot (W(t_{i+1}) - W(t_i))). \end{aligned}$$

f is now adapted to the Brownian motion (which is a condition for the existence of the Ito integral). So the value $f(t_i)$ depends only on the values $W(u)$ of the Brownian motion with $u \leq t_i$. But of these values—as we recalled above in item (i)— $(W(t_{i+1}) - W(t_i))$ is independent.

Therefore, $(W(t_{i+1}) - W(t_i))$ is independent of $f(t_i)$ and therefore

$$E(f(t_i) \cdot (W(t_{i+1}) - W(t_i))) = E(f(t_i)) \cdot E(W(t_{i+1}) - W(t_i)).$$

As we recalled in item (ii) above, $E(W(t_{i+1}) - W(t_i)) = 0$ and therefore

$$E\left(\int_0^T f(t)dW(t)\right) = 0.$$

This means that—this is very important and will be used often in the following—the **expected value of an Ito integral is always equal to 0!**

- Regarding the variance of the Ito integral:

Since the expected value of the Ito integral is equal to 0, we have

$$\begin{aligned}
 V\left(\int_0^T f(t)dW(t)\right) &= E\left(\left(\int_0^T f(t)dW(t)\right)^2\right) \approx \\
 &\approx E\left(\left(\sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i))\right)^2\right) = \\
 &= E\left(\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f(t_i) \cdot f(t_j) \cdot (W(t_{i+1}) - W(t_i)) \cdot \right. \\
 &\quad \left.\cdot (W(t_{j+1}) - W(t_j))\right) = \\
 &= \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} E\left(f(t_i) \cdot f(t_j) \cdot (W(t_{i+1}) - W(t_i)) \cdot \right. \\
 &\quad \left.\cdot (W(t_{j+1}) - W(t_j))\right).
 \end{aligned}$$

If $i < j$, then $f(t_i)$, $f(t_j)$, and $W(t_{i+1}) - W(t_i)$ are all independent of $W(t_{j+1}) - W(t_j)$ and therefore

$$\begin{aligned}
 E\left(f(t_i) \cdot f(t_j) \cdot (W(t_{i+1}) - W(t_i)) \cdot (W(t_{j+1}) - W(t_j))\right) &= \\
 E\left(f(t_i) \cdot f(t_j) \cdot (W(t_{i+1}) - W(t_i))\right) \cdot E\left(W(t_{j+1}) - W(t_j)\right) &= 0
 \end{aligned}$$

The same is true of course if $j < i$. Therefore,

$$\begin{aligned}
 V\left(\int_0^T f(t)dW(t)\right) &= \sum_{i=0}^{N-1} E\left((f(t_i))^2\right) \cdot E\left((W(t_{i+1}) - W(t_i))^2\right) \\
 &= \sum_{i=0}^{N-1} E\left((f(t_i))^2\right) \cdot (t_{i+1} - t_i) \\
 &\approx \int_0^T E\left[f^2(t)\right] dt.
 \end{aligned}$$

(continued)

We have thus derived the so-called **Ito isometry**: As regards the **variance of an Ito integral**, we always have

$$V \left(\int_0^T f(t) dW(t) \right) = \int_0^T E[f^2(t)] dt$$

(Note that, according to the premises, this variance is always finite.)

- The next question to ask with regard to a random variable whose expected value and variance have already been determined is the following: How is it distributed? We will not prove it here, but illustrate below **that an Ito integral over a deterministic function f always follows a normal distribution**.
- To illustrate that an Ito integral over a deterministic function f is always a normally distributed random variable but also for other purposes (especially in connection with Monte Carlo simulation), we consider below how one could simulate the possible process of an Ito integral. The answer is fairly simple: We approximate the integral again by a sum

$$\int_0^T f(t) dW(t) \approx \sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i)).$$

Then we simulate paths of the Brownian motion step by step at the support points t_0, t_1, \dots, t_N using the relation $W(t_{i+1}) = W(t_i) + \sqrt{\Delta t} \cdot w_i$ with independent $\mathcal{N}(0, 1)$ distributed random variables w_i .

For each path created in this way, we calculate a possible value of $\sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i))$ and thus obtain approximations for realizations of $\int_0^T f(t) dW(t)$. Note that if f is a ($W(t)$ adapted) random variable, then, to compute the values $f(t_i)$, we may have to resort to values $W(u)$ with $u \leq t_i$, which in turn may have to be approximated using the simulated values $W(t_1), W(t_2), \dots, W(t_i)$.

- Based on this, we want to simulate the Ito integral $\int_0^1 W(t) dW(t)$ as an example. The integral's expected value is equal to 0 of course. We also calculate its variance upfront, using Ito isometry:

$$\begin{aligned} V \left(\int_0^1 W(t) dW(t) \right) &= \int_0^1 E((W(t))^2) dt = \\ &= \int_0^1 V(W(t)) dt = \int_0^1 t dt = \frac{1}{2}. \end{aligned}$$

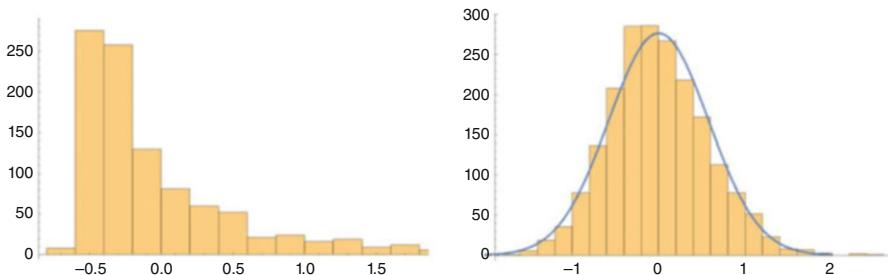


Fig. 3.19 Histograms of realizations of $\int_0^T f(t)dW(t)$ for $f(t) = W(t)$ (left image) and for $f(t) = t$ (right image)

And finally, we use Mathematica to simulate 1000 paths of the standard Brownian motion for the support points $\frac{i}{100}$ with $i = 0, 1, 2, \dots, 100$ and calculate for each path $\sum_{i=0}^{99} W\left(\frac{i}{100}\right) \cdot \left(W\left(\frac{i+1}{100}\right) - W\left(\frac{i}{100}\right)\right)$.

This gives us 1000 approximations for possible realizations of the random variable $\int_0^1 W(t)dW(t)$. We illustrate these values by means of a histogram, recognizing that there is obviously **no** similarity between the shape of the histogram and the shape of the density of an $N\left(0, \frac{1}{2}\right)$ distribution (see Fig. 3.19, left).

If we perform the same simulation procedure for the integral $\int_0^1 t dW(t)$, i.e. for the deterministic function $f(t) = t$ (due to Ito isometry, the variance of this integral is $\int_0^1 t^2 dt = \frac{1}{3}$), and compare the resulting histogram with the density (normalized to the histogram) of the $N\left(0, \frac{1}{3}\right)$ distribution, we see that density and histogram do indeed correspond (see Fig. 3.19, right).

- Let us revisit once more the definition of the Ito integral $\int_0^T f(t)dW(t)$ by approximating sums of the form

$$\sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i)).$$

In comparison, the so-called Riemann sums of the form $\sum_{i=0}^{N-1} f(t_i) \cdot (t_{i+1} - t_i)$ approximate the ordinary Riemann integral $\int_0^T f(t)dt$.

In Riemann integral approximation by Riemann sums, the value t_i at which the function f is evaluated in each case, can be replaced by any value ξ_i between t_i and t_{i+1} . In this case, the Riemann sums $\sum_{i=0}^{N-1} f(\xi_i) \cdot (t_{i+1} - t_i)$ will nevertheless always converge to the same value $\int_0^T f(t)dt$.

On the other hand, when approximating the Ito integral by sums of the form $\sum_{i=0}^{N-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i))$, it is important to take into account at which values the function f is evaluated. For example, if $f(t_i)$ in these sums is replaced

by $f(t_{i+1})$ or by $f\left(\frac{t_i+t_{i+1}}{2}\right)$, one will generally obtain a different value than the Ito integral when taking the limit of the sum to the integral. If we consistently set $f\left(\frac{t_i+t_{i+1}}{2}\right)$ instead of $f(t_i)$, for example, then under certain conditions, there will again be convergence of the sum to a fixed value (which will generally be different from the Ito integral). This value is then called the Stratonovich integral. In some applications, this type of integral is the adequate one. In the case of applications in financial mathematics, however—as we saw in the motivation for the profit function of a trading strategy—the Ito integral is exactly the right concept.

To give interested readers a rough idea of why, in the case of convergence of sums of the form $\sum_{i=0}^{N-1} f(\xi_i) \cdot (W(t_{i+1}) - W(t_i))$ (which, for brevity, we are going to refer to as “Ito sums” in the following) as opposed to sums of the form $\sum_{i=0}^{N-1} f(\xi_i) \cdot (t_{i+1} - t_i)$, the position of the evaluation point ξ_i is essential for the function f , let us consider the following:

In the Riemann sums, the “increments” $(t_{i+1} - t_i)$ are always positive and the sum of all of their lengths is exactly T .

It follows immediately, for example, that—if only the function f is bounded, say, by a value L —even the sum of the absolute values of the Riemann sums will always have a fixed upper limit at $L \cdot T$.

In the case of Ito sums, the increments $(W(t_{i+1}) - W(t_i))$ can be both positive and negative. The increments are random variables with variance $(t_{i+1} - t_i)$, hence with standard deviation $\sqrt{t_{i+1} - t_i}$.

Thus, we can say that the increments $(W(t_{i+1}) - W(t_i))$ have an average length of $\sqrt{t_{i+1} - t_i}$.

$\sqrt{t_{i+1} - t_i}$ is, when the distance between t_i and t_{i+1} goes to 0, substantially greater than $(t_{i+1} - t_i)$.

What about the approximate total length of the increments

$(W(t_{i+1}) - W(t_i))$? If we assume again that an increment has an average length of $\sqrt{t_{i+1} - t_i}$ and that the t_i divide the interval $[0, T]$ into N equal parts of length $\Delta t = \frac{T}{N}$, then the average total length of the increments is $N \cdot \sqrt{\Delta t} = N \cdot \sqrt{\frac{T}{N}} = \sqrt{N} \cdot \sqrt{T}$. But if we refine the subdivision of the interval $[0, T]$ further and further, that is, let N become larger and larger, then this expression will converge to infinity. Therefore, the sum of the absolute values of the summands of an Ito sum will generally not be bounded above. So for uniform convergence to occur in the case of Ito sums, the positive and negative summands in an Ito sum must more or less neutralize each other in some kind of “ingenious” way, and for that to happen, it is quite essential how exactly the increments $(W(t_{i+1}) - W(t_i))$ are weighted by the $f(\xi_i)$.

- If we let the upper bound in a stochastic integral $\int_0^t f(u)dW(u)$ vary from 0 to T , for example, then we will yet again obtain a stochastic process $\left(\int_0^t f(u)dW(u)\right)_{t \in [0, T]}$.

An example of a process of such a form is, for instance, the above profit process G , where we do not only take account of the value at the final time point T but follow the entire profit process $(G(t))_{t \in [0, T]}$ over the period $[0, T]$. This process has the form

$$\left(\int_0^t h(u) \cdot \mu \cdot S(u)du + \int_0^t h(u) \cdot \sigma \cdot S(u)dW(u) \right)_{t \in [0, T]}.$$

- In Sect. 3.2, we motivated and introduced the differential notation for stochastic processes, in particular for Ito processes $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$. In doing so, we argued that this representation is a shorthand or alternative representation for the relation $S(t + \Delta t) - S(t) = \alpha(t, S(t)) \cdot \Delta t + \beta(t, S(t)) \cdot (W(t + \Delta t) - W(t))$ for “infinitesimally small” Δt . In modern mathematics, “infinitesimally small sizes” is not an exact phrasing of course. In such cases, we do not work with infinitely small quantities, but consider taking possible limits $\Delta t \rightarrow 0$, perform them if possible, and analyse them. In our case here, this would require dividing by Δt (otherwise—if we let $\Delta t \rightarrow 0$ right away—we would only get a trivial result $0 = 0$), so $\frac{S(t+\Delta t)-S(t)}{\Delta t} = \alpha(t, S(t)) + \beta(t, S(t)) \cdot \frac{W(t+\Delta t)-W(t)}{\Delta t}$.

If we now let Δt go to 0 and assumed differentiability of S and W , then we would get $\frac{dS(t)}{dt} = \alpha(t, S(t)) + \beta(t, S(t)) \cdot \frac{dW(t)}{dt}$, and this representation would correspond exactly to the Ito representation $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$ “divided by dt ”. Yet the Brownian motion $W(t)$, as we know, is nowhere differentiable! And by the same token, an Ito process S is generally not differentiable either. Therefore, the Ito representation $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$ cannot be an alternative notation for the relation $\frac{dS(t)}{dt} = \alpha(t, S(t)) + \beta(t, S(t)) \cdot \frac{dW(t)}{dt}$.

Must we conclude then that $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$ has only a heuristic, intuitive meaning and no mathematically exact meaning?

The answer is we look at the Ito representation “one level higher up”, that is, we first replace the parameter t by the new parameter u and integrate both sides formally from 0 to t and get

$$S(t) - S(0) = \int_0^t \alpha(u, S(u))du + \int_0^t \beta(u, S(u)) \cdot dW(u).$$

Now, this expression contains only mathematically exact variables (provided, of course, that β is a process for which the Ito integral exists), and it is indeed the correct exact interpretation of the shorthand notation $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$.

To summarize,

The Ito notation $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$ for an Ito process $S(t)$ is a shorthand notation for the correct and exact integral representation:

$$S(t) = S(0) + \int_0^t \alpha(u, S(u)) du + \int_0^t \beta(u, S(u)) \cdot dW(u).$$

The representation $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$ is a short and compact and also highly intuitive representation, but bear in mind when using it that the exact version is always the integral version!

- How can we calculate the “value” of an Ito integral?

First of all, remember that the “value” of an Ito integral is not a real number, but a random variable! In many cases, you won’t need any explicit representation of this resulting random variable. Indeed, in the following, we will very rarely need to actually “calculate” an Ito integral.

In some cases, it is possible to calculate an Ito integral using the Ito formula (see Sect. 3.4) and applying a suitable approach. Here is one example to illustrate this: We calculate $\int_0^t (W(u))^2 dW(u)$:

If we integrated this rather “naively”, just as one would integrate the function x^2 , we might expect a result of the form $\frac{(W(u))^3}{3}$. Based on that, we would then apply Ito formula to the process $S(t) = W(t)$, that is, $dS(t) = 0 \cdot dt + 1 \cdot dW(t)$, and to the function $g(t, x) := \frac{x^3}{3}$.

(Recall the Ito formula:

Let S be an Ito process of the form $dS(t) = \alpha(t, S(t)) \cdot dt + \beta(t, S(t)) \cdot dW(t)$ with fixed given initial value $S(0)$. Let

$$\begin{aligned} g : \mathbb{R} \times \mathbb{R} &\rightarrow \mathbb{R} \\ (t, x) &\rightarrow g(t, x) \end{aligned}$$

be a function which is once continuously differentiable with respect to t and twice continuously differentiable with respect to x . Then the process $Y(t) := g(t, S(t))$ is again an Ito process, and we get

$$dY = \left(g_t(t, S(t)) + g_x(t, S(t)) \cdot \alpha + g_{xx}(t, S(t)) \cdot \frac{\beta^2}{2} \right) \cdot dt + g_x(t, S(t)) \cdot \beta \cdot dW.$$

In our case, $\alpha = 0$, $\beta = 1$, $g_t = 0$, $g_x = x^2$, $g_{xx} = 2x$, and $Y = \frac{(W(t))^3}{3}$, and based on Ito formula, it follows therefore that $dY = W(t)dt + (W(t))^2dW$. Writing this in the correct integral notation, we get

$$\frac{(W(t))^3}{3} = \int_0^t W(u)du + \int_0^t (W(u))^2 dW(u)$$

and therefore

$$\int_0^t (W(t))^2 dW(t) = \frac{(W(t))^3}{3} - \int_0^t W(t) dt.$$

Thus, the result has some similarity to the result that we naively assumed, namely, $\frac{(W(t))^3}{3}$, yet with one additional term (in this case $-\int_0^t W(t) dt$).

3.19 Conditional Expectations and Martingales

In the context of stochastic processes, there are two more concepts that you will come to appreciate as very helpful and enlightening, and we will introduce them intuitively and heuristically in the following. These two concepts are those of the conditional expectation and the martingale.

We start again with a stochastic process $(S(t))_{t \in [0, T]}$, and we assume that we are at time 0 and focus on two future time points v and w . Let $0 < v < w \leq T$.

Of course, we are interested in finding the expected value $E(S(w))$ of the process S at time w . So, the question is

“What is the expectation $E(S(w))$ of $S(w)$? ”

A similar, yet more nuanced, interest in the expected value of S at the time w might be:

“What is the expected value of $S(w)$ as a function of $S(v)$, that is, of the value of S at time v ? ”

We denote this expected value by $E(S(w)|S(v))$.

If we are currently at time v , then we see the actual value of $S(v)$ and can calculate the expected value $E(S(w)|S(v))$ from the perspective of v . If we are at time 0, however, we do not yet know where $S(v)$ is going to be. The value $E(S(w)|S(v))$ from the perspective of time 0 thus depends on the as yet unknown random value $S(v)$. So, from the perspective of time 0, $E(S(w)|S(v))$ is a random variable whose value depends on the value of $S(v)$. We refer to $E(S(w)|S(v))$ as the **conditional expectation of $S(w)$ under $S(v)$** .

Here is an **example**:

We have a Brownian motion $W(t)$ on $[0, T]$. From the perspective of time 0, the expectation for any w is $E(W(w)) = 0$. However, if we knew at time 0 that W would have the value $W(v) = x$ at a time v (with $v < w$), then the following would hold from the perspective of time 0: $W(w) = x + (W(w) - W(v))$ and $W(w) - W(v)$ being an $N(0, w - v)$ -distributed random variable. Thus, we would get $E(W(w)|W(v)) = E(x + (W(w) - W(v))) = x + E(W(w) - W(v)) = x = W(v)$.

So, in terms of the Brownian motion, for any time w in the future, we have $E(W(w)) = 0 = W(0)$, but for any $v < w$, we have $E(W(w)|W(v)) = W(v)$.

From any point in time v and value $W(v)$, the expectation of $W(w)$ in the future is $W(v)$. We have attempted to illustrate this situation in Fig. 3.20.

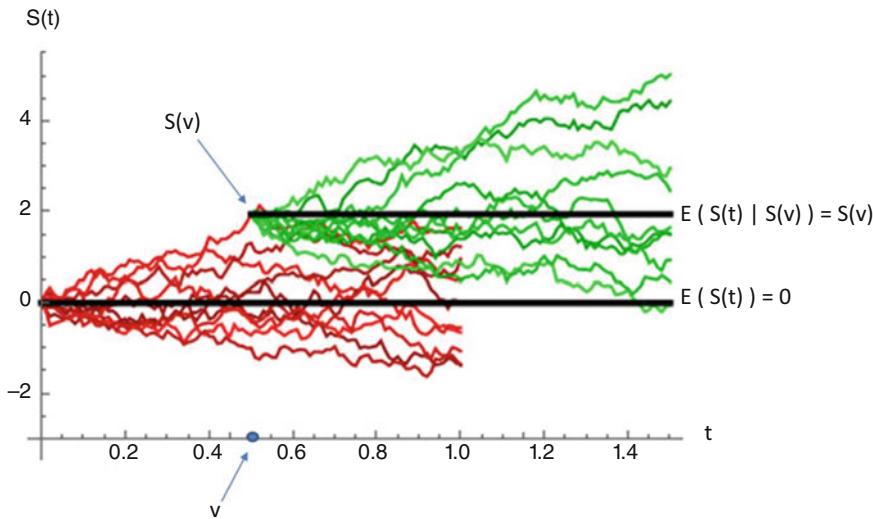


Fig. 3.20 Illustration of conditional expectation and martingale

A stochastic process $(S(t))_{t \in [0, T]}$ where for all arbitrary time points v and w with $0 < v < w \leq T$ the conditional expectation satisfies $E(S(w) | S(v)) = S(v)$ is called a **martingale**.

By the way, whenever the profit process $(G(t))$ refers to a fair random game (e.g. a coin toss), that profit process is assumed to be a martingale: When starting the fair game, my realistic profit expectation for all future points in time is ± 0 .

Now, if I have already played some rounds of that game up to time v and have accumulated a profit of $G(v) = X$ EUR so far, then my average future (!) profit expectation as from now (from v) continues to be ± 0 for the future rounds too, of course. However, since I have already won X EUR, the average total profit expectation is therefore $X = G(u)$.

We established above that the Brownian motion is a martingale.

Here is **another example**:

We are now going to “prove” that **any stochastic process $(S(t))_{t \in [0, T]}$ defined by an Ito integral, that is, where $S(t) = \int_0^t f(s)dW(s)$ for all t , is a martingale**.

For purposes of that proof, we will again use an Ito sum that approximates the integral arbitrarily well, and we will use some obvious properties of the conditional expectation that we did not explicitly note above (such as its unconditional additivity and its multiplicativity when factors are independent of one another). For arbitrary time points v and w with $0 \leq v < w$ and a subdivision of the form $v = t_p < t_{p+1} < \dots < t_{q-1} < t_q = w$ of the interval $[v, w]$ and due to the independence of $f(t_i)$

and $(W(t_{i+1}) - W(t_i))$, we then get

$$\begin{aligned}
& E \left(\int_0^w f(s) dW(s) \middle| \int_0^v f(s) dW(s) = x \right) = \\
& E \left(\int_0^v f(s) dW(s) + \int_v^w f(s) dW(s) \middle| \int_0^v f(s) dW(s) = x \right) = \\
& E \left(\int_0^v f(s) dW(s) \middle| \int_0^v f(s) dW(s) = x \right) + \\
& + E \left(\int_v^w f(s) dW(s) \middle| \int_0^v f(s) dW(s) = x \right) = \\
& x + E \left(\int_v^w f(s) dW(s) \middle| \int_0^v f(s) dW(s) = x \right) \approx \\
& x + E \left(\sum_{i=p}^{q-1} f(t_i) \cdot (W(t_{i+1}) - W(t_i)) \middle| \int_0^v f(s) dW(s) = x \right) = \\
& x + \sum_{i=p}^{q-1} E \left(f(t_i) \middle| \int_0^v f(s) dW(s) = x \right) \cdot \\
& \cdot E \left(W(t_{i+1}) - W(t_i) \middle| \int_0^v f(s) dW(s) = x \right)
\end{aligned}$$

Let us now take a closer look at $E(W(t_{i+1}) - W(t_i) | \int_0^v f(s) dW(s) = x)$.

The random component of the integral $\int_0^v f(s) dW(s)$ depends exclusively on the evolution of the Brownian motion $W(s)$ in the interval $[0, v]$. The increment $W(t_{i+1}) - W(t_i)$ is therefore independent of the integral $\int_0^v f(s) dW(s)$ (since $v \leq t_i < t_{i+1}$). The expected value of the increment $W(t_{i+1}) - W(t_i)$ is thus completely independent of the value x of the integral $\int_0^v f(s) dW(s)$. Hence

$$E \left(W(t_{i+1}) - W(t_i) \middle| \int_0^v f(s) dW(s) = x \right) = E(W(t_{i+1}) - W(t_i)) = 0.$$

And so $E(\int_0^w f(s) dW(s) | \int_0^v f(s) dW(s) = x) = x$. Thus

$$E \left(\int_0^w f(s) dW(s) \middle| \int_0^v f(s) dW(s) \right) = \int_0^v f(s) dW(s),$$

and $\left(\int_0^t f(s)dW(s)\right)_{t \in [0, T]}$ is therefore a martingale.

Important Note on Notations

So for an Ito process $(S(t))_{t \in [0, T]}$, we now know how to handle the concept of the conditional expectation $E(S(w)|S(v))$. We want to determine the expectation of $S(w)$ on the assumption that $S(v)$ is known. Yet $S(v)$ is uniquely determined by the Brownian motion W from 0 to v . Therefore, instead of the notation $E(S(w)|S(v))$, one could, equivalently, use the notation $E(S(w)|(W(t))_{t \in [0, v]})$. The latter offers some advantages (e.g. if $S(w)$ is a very complex process—possibly a combination of several other processes). In shorthand notation (based on a theoretical background that we cannot go into here), it has become customary to use $E(S(w)|F_v)$ instead of $E(S(w)|(W(t))_{t \in [0, v]})$. F_v here denotes the σ -algebra at time v of the filtration of the Brownian motion. In the following, we will denote conditional expectations either by $E(S(w)|F_v)$ or by the original notation, depending on which one happens to be more practical or intuitive in each case.

3.20 The Feynman-Kac Formula

We will arrive at this astonishing formula—which connects two seemingly completely unrelated fields of mathematics—by just playing around a bit with the Ito formula at first.

We start with an Ito process $(X(t))_{t \in [0, T]}$ of the form

$$dX(t) = \mu(t, X(t))dt + \sigma(t, X(t))dW(t), \text{ of any function}$$

$$\begin{aligned} F : [0, T] \times \mathbb{R} &\rightarrow \mathbb{R} \\ (t, x) &\rightarrow F(t, x) \end{aligned}$$

with the only assumption that F is continuously differentiable at least once in t and at least twice in x , and we then apply Ito formula to the function $g(t, x) := e^{rt} \cdot F(t, x)$, i.e. to the stochastic process $Y(t) := g(t, X_t)$. r is an arbitrary real number. We have

$$g_t = r \cdot e^{rt} \cdot F(t, x) + e^{rt} \cdot F_t(t, x)$$

$$g_x = e^{rt} \cdot F_x(t, x) \text{ and } g_{xx} = e^{rt} \cdot F_{xx}(t, x)$$

and therefore by means of Ito formula

$$\begin{aligned} dY(t) &= \left(r \cdot e^{rt} \cdot F(t, X(t)) + e^{rt} \cdot F_t(t, X(t)) + \right. \\ &\quad \left. + \mu(t, X(t)) \cdot e^{rt} \cdot F_x(t, X(t)) + \frac{\sigma(t, X(t))^2}{2} \cdot \right. \\ &\quad \left. \cdot e^{rt} \cdot F_{xx}(t, X(t)) \right) dt + \sigma(t, X(t)) \cdot e^{rt} \cdot F_x(t, X(t)) dW(t). \end{aligned}$$

We now formulate this shorthand differential notation in the correct integral version, but for once, we do not integrate from 0 to T but from any arbitrary point in time t to T . The integral form is then

$$\begin{aligned} Y(T) = Y(t) + \int_t^T & \left(r \cdot e^{ru} \cdot F(u, X(u)) + e^{ru} \cdot F_t(u, X(u)) + \mu(u, X(u)) \right. \\ & \cdot e^{ru} \cdot F_x(u, X(u)) + \frac{\sigma(u, X(u))^2}{2} \cdot e^{ru} \cdot F_{xx}(u, X(u)) \Big) du + \\ & + \int_t^T \sigma(u, X(u)) \cdot e^{ru} \cdot F_x(u, X(u)) dW(u). \end{aligned}$$

Assuming that F is a function that coincidentally “happens to” satisfy the following, for all $u \in [0, T]$ and all $x \in R$, we have

$$r \cdot F(u, x) + F_t(u, x) + \mu(u, x) \cdot F_x(u, x) + \frac{\sigma(u, x)^2}{2} \cdot F_{xx}(u, x) = 0$$

and $F(T, x) = \Phi(x)$, where Φ is a fixed given function. This would give us $Y(T) = Y(t) + \int_t^T \sigma(u, X(u)) \cdot e^{ru} \cdot F_x(u, X(u)) dW(u)$, so

$$e^{rT} \cdot F(T, X(T)) = e^{rt} \cdot F(t, X(t)) + \int_t^T \sigma(u, X(u)) \cdot e^{ru} \cdot F_x(u, X(u)) dW(u).$$

In this last equation, we now take the conditional expectation with respect to \mathcal{F}_t from both sides, noting that

$$E(F(t, X(t)) | \mathcal{F}_t) = F(t, X(t)),$$

and also

$$E(F(T, X(T)) | \mathcal{F}_t) = E(\Phi(X(T)) | \mathcal{F}_t)$$

and (since stochastic integrals are martingales)

$$\begin{aligned} E \left(\int_t^T \sigma(u, X(u)) \cdot e^{ru} \cdot F_x(u, X(u)) dW(u) \middle| \mathcal{F}_t \right) = \\ = \int_t^t \sigma(u, X(u)) \cdot e^{ru} \cdot F_x(u, X(u)) dW(u) = 0. \end{aligned}$$

It follows therefore that

$$F(t, X(t)) = e^{r(T-t)} \cdot E(\Phi(X(T)) | \mathcal{F}_t)$$

or in equivalent notation

$$F(t, x) = e^{r(T-t)} \cdot E(\Phi(X(T)) | X(t) = x).$$

We summarize this in a suitable manner:

Theorem 3.12 (Feynman-Kac Formula) *Let*

$$\begin{aligned} F : [0, T] \times \mathbb{R} &\rightarrow \mathbb{R} \\ (t, x) &\mapsto F(t, x) \end{aligned}$$

be a function which is continuously differentiable at least once in t and at least twice in x and which satisfies the following partial differential equation:

$$r \cdot F(t, x) + F_t(t, x) + \mu(t, x) \cdot F_x(t, x) + \frac{\sigma(t, x)^2}{2} \cdot F_{xx}(t, x) = 0$$

with the boundary condition $F(T, x) = \Phi(x)$. Then

$$F(t, x) = e^{r(T-t)} \cdot E(\Phi(X(T)) | X(t) = x),$$

where $X(t)$ is an Ito process with dynamics

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

This Feynman-Kac formula thus creates a connection between solving a partial differential equation and calculating the conditional expectation of a stochastic process. An amazing connection!

Just a **simple example** to illustrate this:

We want to solve the partial differential equation (PDE)

$$F(t, x) + F_t(t, x) + F_{xx}(t, x) = 0$$

with boundary condition $F(T, x) = x^2$

on the time range $[0, T] \times \mathbb{R}$.

The equation fits exactly into the Feynman-Kac formula with the parameters

$$r = 1$$

$$\mu(t, X(t)) = 0$$

$$\sigma(t, X(t)) = \sqrt{2}$$

$$\Phi(x) = x^2$$

The stochastic process X that we need for solving this thus has the form

$$dX(t) = 0 \cdot dt + \sqrt{2} \cdot dW(t) \text{ i.e. } X(t) = \sqrt{2} \cdot W(t).$$

Applying Feynman-Kac, we obtain

$$F(t, x) = e^{r(T-t)} \cdot E\left((X(T))^2 \mid X(t) = x\right) = 2 \cdot e^{r(T-t)} \cdot E\left((W(T))^2 \mid W(t) = \frac{x}{\sqrt{2}}\right).$$

To determine $E\left((W(T))^2 \mid W(t) = \frac{x}{\sqrt{2}}\right)$, we apply a frequently used “trick”:

$$\begin{aligned} & E\left((W(T))^2 \mid W(t) = \frac{x}{\sqrt{2}}\right) = \\ &= E\left((W(T) - W(t))^2 + 2W(t) \cdot W(T) - (W(t))^2 \mid W(t) = \frac{x}{\sqrt{2}}\right) = \\ &= E\left((W(T) - W(t))^2 \mid W(t) = \frac{x}{\sqrt{2}}\right) + \\ &+ E\left(2W(t) \cdot W(T) \mid W(t) = \frac{x}{\sqrt{2}}\right) - E\left((W(t))^2 \mid W(t) = \frac{x}{\sqrt{2}}\right) = \\ &= E\left((W(T) - W(t))^2\right) + \frac{2x}{\sqrt{2}} \cdot E\left(W(T) \mid W(t) = \frac{x}{\sqrt{2}}\right) - \frac{x^2}{2} = \\ &= (T-t) + \frac{2x}{\sqrt{2}} \cdot \frac{x}{\sqrt{2}} - \frac{x^2}{2} = (T-t) + \frac{x^2}{2}. \end{aligned}$$

So the solution of the partial differential equation is given by

$$F(t, x) = 2 \cdot e^{T-t} \cdot (T-t) + x^2 \cdot e^{T-t}.$$

(The reader should carefully think through the legitimacy of each single step in the last lines!)

Let us also check the solution: For this resulting $F(t, x)$, we have

$$\begin{aligned} F_t(t, x) &= -2e^{T-t} \cdot (T-t) - 2e^{T-t} - x^2 \cdot e^{T-t} \\ F_x(t, x) &= 2x \cdot e^{T-t} \text{ and } F_{xx}(t, x) = 2e^{T-t} \end{aligned}$$

Substituting in the PDE, we get

$$2e^{T-t} \cdot (T-t) + x^2 \cdot e^{T-t} - 2e^{T-t} \cdot (T-t) - 2e^{T-t} - x^2 \cdot e^{T-t} + 2e^{T-t} = 0$$

and the boundary condition $F(T, x) = 2e^{T-T} \cdot (T - T) + x^2 \cdot e^{T-T} = x^2$ is also satisfied.

3.21 The Black-Scholes Formula

Now, we are also able to solve the Black-Scholes equation derived in Sect. 3.17 using Feynman-Kac.

We recall the Black-Scholes equation:

Let S be an underlying asset whose price $S(t)$ in the time range $[0, T]$ follows a stochastic differential equation $dS(t) = \mu(t, S(t)) \cdot S(t)dt + \sigma(t, S(t)) \cdot S(t)dW(t)$ with given initial value $S(0)$ and with deterministic functions μ and σ .

Let D be a European plain vanilla derivative D on the underlying asset S expiring in T , defined by the payoff function Φ . Let $F(t, s)$ be the function that describes the fair price process of the derivative D for $t \in [0, T]$ and $s > 0$. F then satisfies the partial differential equation

$$F_t(t, s) + r \cdot s \cdot F_s(t, s) + \frac{(\sigma(t, s) \cdot s)^2}{2} \cdot F_{ss}(t, s) = r \cdot F(t, s)$$

with boundary condition $F(T, s) = \Phi(s)$. Here, r denotes the risk-free interest rate on $[0, T]$.

Rearranging this PDE somewhat we get

$$-rF(t, s) + F_t(t, s) + r \cdot s \cdot F_s(T, s) + \frac{(\sigma \cdot s)^2}{2} \cdot F_{ss}(t, s) = 0$$

with $F(T, s) = \Phi(s)$.

So here, we are (almost) in a situation where Feynman-Kac can provide a solution. In any case, what we have to do now is set the following:

$$\mu(t, s) = r \cdot s$$

$$\sigma(t, s) = \sigma \cdot s$$

in the Feynman-Kac formula. We just have to make sure to replace the “r” in Feynman-Kac by the negative interest rate “-r” in Black-Scholes terminology.

This gives us the following **solution of the Black-Scholes equation**:

$$F(t, s) = e^{-r(T-t)} \cdot E \left(\Phi(\tilde{S}(T)) \mid \tilde{S}(t) = s \right)$$

where \tilde{S} is a stochastic process with dynamics

$$d\tilde{S}(t) = r \cdot \tilde{S}(t)dt + \sigma(t, s) \cdot \tilde{S}(t) \cdot dW(t)$$

We have thus found a second, “direct” proof of the Black-Scholes formula without taking the “detour” via the binomial model. Nevertheless, in the next section, we will give another alternative proof of the Black-Scholes formula, again based on stochastic analysis.

3.22 The Black-Scholes Model as a Complete Market and Hedging of Derivatives

There are two reasons why we are now going to prove the Black-Scholes formula one more time:

1. In the above proof of the Black-Scholes equation (and consequently of the Black-Scholes formula), we proceeded in the following way: We created a self-financing dynamic portfolio from an underlying asset S and a derivative D such that it delivered a risk-free performance. Comparing that with the performance of the risk-free bond B then led to the Black-Scholes equation. However, some derivatives D may not necessarily be tradable at all times during their term. Particularly derivatives that are not traded on an exchange but are agreed OTC between two contracting parties are then often no longer tradable. In this case, the above proof is no longer feasible (we cannot generate a dynamic portfolio with a non-tradable derivative)! We would even have to ask ourselves whether the fact that a derivative cannot be traded throughout wouldn't reduce the value to the buyer of that derivative, that is, whether the fair price of a derivative that is not permanently tradable might be lower than if it were permanently tradable. In any case, we need an alternative proof to answer this question. The above proof is not applicable here.
2. The above proof does not provide information on how to optimally hedge the derivative D .

The proof that now follows will be applicable even to derivatives that are not tradable throughout their life, and it will additionally give us the perfect hedging strategy for a derivative D .

The financial market which we have been operating in for some time and will continue to operate in for the following consists of the two basic products B (bond) and S (underlying asset), where B follows the dynamics $dB(t) = r \cdot B(t)dt$ and S follows the dynamics $dS(t) = \mu(t, S(t)) \cdot S(t)dt + \sigma(t, X(t)) \cdot S(t)dW(t)$ with deterministic functions μ and σ . Furthermore, there are (European) derivatives D in this financial market which are given by their payoff function Φ at their expiration date T .

This financial market is called **one-dimensional Black-Scholes model**.

We say, a **derivative D is replicable in this financial market** if there is a self-financing trading strategy (or a self-financing dynamic portfolio) consisting of the bond B and the underlying asset S such that for the performance V of this strategy we have $V(S(T)) = \Phi(S(T))$. In other words, regardless of how the underlying

asset S changes in value from time 0 to time T , the value of our portfolio at time T is exactly the same as the payoff that the derivative D yields at time T . The portfolio that we use to replicate the derivative D is referred to as the **replicating portfolio** of D .

Of course, we need an initial investment to start the self-financing trading strategy, that is, to buy the initial portfolio. This initial investment amount is equal to the value of the strategy at the beginning of the trading process, i.e. equal to $V(0)$.

Now, it follows directly from the no-arbitrage principle that the **fair price $F(0, S(0))$ of the derivative D must be equal to $V(0)$ at time 0**.

That's because if $V(0) < F(0, S(0))$, then we would be able to short the derivative D and execute the trading strategy. At time T , the payoff of the derivative and the value of the dynamic portfolio cancel each other out, and we would have realized a risk-free profit of $F(0, S(0)) - V(0)$ without any capital investment. If $V(0) > F(0, S(0))$, then we would choose the exact opposite procedure and also pocket a risk-free profit without having invested any capital.

In more general terms, this means at any time t at which the derivative is tradable, its price $F(t, S(t))$ must be equal to the value $V(t)$ of the replicating portfolio.

Therefore, whenever a derivative D is replicable in a financial market and we know its replicating strategy, we also know the fair price of that derivative. In this case—when this procedure is used for determining the fair price—the derivative doesn't have to be tradable throughout! Furthermore, the replicating strategy is a perfect hedging strategy for the derivative D .

So by finding the replicating strategy, we obtain both the fair price of the derivative and the perfect hedging procedure!

A financial market in which every derivative is replicable is called a complete market.

We show below:

Theorem 3.13 *The one-dimensional Black-Scholes model is a complete market.*

Proof The obvious way to do the proof is to try to explicitly construct the replicating portfolio for any derivative D with payoff function Φ .

So we are looking for a self-financing strategy with a value process V such that the equation $V(T, S(T)) = \Phi(S(T))$ always holds. For this purpose, we will invest, at each point in time t , a portion in the amount of $u(t) \cdot V(t, S(t))$ of our portfolio's current total assets $V(t, S(t))$ in the underlying asset and the remainder, that is, $(1 - u(t)) \cdot V(t, S(t))$, in the bond B .

Meaning, in shorthand notation, at any time t , we hold exactly $\frac{u \cdot V}{S}$ units of the underlying asset S and $\frac{(1-u) \cdot V}{B}$ units of the bond B .

The trading strategy is, in principle, again such that we always invest only as much money in the underlying asset and the derivative as there is value V in the portfolio. The strategy is thus self-financing. No additional money is added to or withdrawn from the portfolio at any time during the strategy's life. We will

determine the specific value for $u(t)$, that is, the exact definition of the dynamic portfolio, at a later point.

For the dynamics of the value process V , we again have, on the one hand, the relation

$$\begin{aligned} dV &= \frac{u \cdot V}{S} \cdot dS + \frac{(1-u) \cdot V}{B} \cdot dB = V \cdot \left(u \cdot \frac{dS}{S} + (1-u) \cdot \frac{dB}{B} \right) = \\ &= V \cdot (u \cdot (\mu \cdot dt + \sigma \cdot dW(t)) + (1-u) \cdot r \cdot dt) = \\ &= V \cdot ((u \cdot \mu + (1-u) \cdot r)dt + u \cdot \sigma \cdot dW(t)), \end{aligned}$$

thus

$$dV = V \cdot ((u \cdot \mu + (1-u) \cdot r) dt + V \cdot u \cdot \sigma \cdot dW(t)).$$

On the other hand, applying the Ito formula, we get for the value process $V(t, S(t))$:

$$dV = \left(V_t + \mu \cdot s V_s + \frac{(\sigma s)^2}{2} V_{ss} \right) dt + \sigma \cdot s \cdot V_s \cdot dW(t).$$

This gives us the equations

$$V \cdot ((u \cdot \mu + (1-u) \cdot r) = \left(V_t + \mu \cdot s V_s + \frac{(\sigma s)^2}{2} V_{ss} \right)$$

and

$$V \cdot u \cdot \sigma = \sigma \cdot s \cdot V_s.$$

From the second equation, we get $u = \frac{s \cdot V_s}{V}$. Substituting this value for u in the first equation and rearranging, we obtain the following equation:

$$V_t(t, s) + r \cdot s \cdot V_s(t, s) + \frac{(\sigma(t, s) \cdot s)^2}{2} \cdot V_{ss}(t, s) = r \cdot V(t, s).$$

This relationship must hold for the value process of any such dynamic trading strategy V . That begs the question, however, whether we can create V such that it can actually be used to replicate the derivative D , that is, such that the equation $V(T, s) = \Phi(s)$ holds.

Now, Feynman-Kac guarantees that the above PDE with auxiliary condition $V(T, s) = \Phi(s)$ can be solved. So a replicating trading strategy V for the derivative D does indeed exist! The proof of the theorem is thus completed.

The price process F of the derivative must have the same value V at any time that the derivative is tradable. Therefore, F must also satisfy the PDE

$$F_t(t, s) + r \cdot s \cdot F_s(t, s) + \frac{(\sigma(t, s) \cdot s)^2}{2} \cdot F_{ss}(t, s) = r \cdot F(t, s)$$

with the auxiliary condition $F(T, r) = \Phi(s)$.

Thus, in this way, we have once again derived the Black-Scholes equation. It therefore applies even if the derivative D is not permanently tradable. **The fair price of a derivative D is not affected by the fact whether D is permanently tradable or not!**

Now, what does the replicating strategy, i.e. the perfect hedging strategy for the derivative D , look like specifically?

We start with an investment in the amount of $F(0, S(0))$. At each point in time t , we hold $\frac{u(t) \cdot V(t, S(t))}{S(t)} = \frac{S(t) \cdot V_s(t, S(t))}{V(t, S(t))} \cdot \frac{V(t, S(t))}{S(t)} = V_s(t, S(t)) = F_s(t, S(t))$ units of the underlying asset S in the portfolio. The remaining investment amount is held in the bond B . So, in this way, we have once again demonstrated the **optimality of delta hedging**. \square

3.23 The Multidimensional Black-Scholes Model and Its Completeness

In Sect. 2.30, we already dealt with a multidimensional Black-Scholes model. Our underlying assets there were a bond B with dynamics $dB(t) = r \cdot B(t)dt$ (with constant interest rate r) and d risk-bearing and quite possibly mutually dependent assets S_1, S_2, \dots, S_d .

The dependence between these assets was modelled based on their dynamics being governed to a greater or lesser extent by the same d independent Brownian motions W_1, W_2, \dots, W_d . In Sect. 2.30, we represented the assets S_i explicitly in the form of geometric Brownian motions. Now, we choose the corresponding stochastic differential representation:

So every S_i has dynamics of the form

$$dS_i(t) = \mu_i \cdot S_i(t)dt + \sum_{k=1}^d \sigma_{i,k} \cdot S_i(t)dW_k(t) \quad \text{for } i = 1, 2, \dots, d$$

It is assumed here that the covariance matrix $\Sigma := (\sigma_{i,k})_{i,k=1,2,\dots,d}$ is a regular, that is, an invertible matrix.

In this financial market model (the d -dimensional Black-Scholes model), we can again consider derivatives whose payoff at expiration T depends on the values of the d underlying assets S_1, S_2, \dots, S_d at time T . The payoff is thus given by $\Phi(S_1(T), S_2(T), \dots, S_d(T))$.

We can then show the following:

Theorem 3.14 *Under the above assumptions, the d -dimensional Black-Scholes model is complete, and the pricing process $F(t, S_1(t), S_2(t), \dots, S_d(t))$ satisfies the following Black-Scholes equation (in the following we write s for short to denote*

the vector (s_1, s_2, \dots, s_d)):

$$F_t(t, s) = \sum_{i=1}^d r \cdot s_i \cdot F_{s_i}(t, s) + \frac{1}{2} \cdot \sum_{i,j=1}^d s_i \cdot s_j \cdot F_{s_i, s_j}(t, s) \cdot c_{i,j} - r \cdot F(t, s) = 0$$

with boundary condition $F(T, s) = \Phi(s)$.

Here, the values $c_{i,j}$ are given by $C := (c_{i,j})_{i,j=1,2,\dots,d} := \sum \cdot \sum^T$.

The solution of this partial differential equation is given by

$$F(t, s) = e^{-r(T-t)} \cdot E\left(\Phi\left(\tilde{S}_1(T), \tilde{S}_2(T), \dots, \tilde{S}_d(T)\right) \mid \tilde{S}_1(t) = s_1, \tilde{S}_2(t) = s_2, \dots, \tilde{S}_d(t) = s_d\right)$$

where $\tilde{S}_i(t)$ for each i follows the dynamics $d\tilde{S}_i(t) = r \cdot \tilde{S}_i(t)dt + \sum_{k=1}^d \sigma_{i,k} \cdot \tilde{S}_i(t)dW_k(T)$.

The proof of this theorem is in principle analogous to the proof in the one-dimensional case. However, we need a multidimensional version of the Ito formula and the Feynman-Kac formula and have to put in more technical effort. We will not give the proof of the theorem here. We only note that the given setting in this form is necessary for the theorem to retain its validity. Put simply, the following applies:

If the number of Brownian motions W_i is smaller than the number of products S_i , then a risk-free product can be generated by a linear combination of the products S_i . And if that product does not match bond B , arbitrage would be possible. But if it matches the bond, then one of the S_i can be represented as a linear combination of the bond B and the other products S_j , making it superfluous in the system. The number of products can thus be reduced (until the number of products S_i equals the number of Brownian motions). The same applies if the covariance matrix Σ is not invertible.

If the number of Brownian motions is greater than the number of products S_i , then the Black-Scholes market is generally no longer complete. Intuitively this is the case because there are too few products B and S_1, \dots, S_d to ensure that a (dynamic) combination of these products will eliminate all random sources W_i and that the value $\Phi(S_1(T), S_2(T), \dots, S_d(T))$ can definitely be attained at time T . (In a certain sense, the situation is then comparable to the situation in a trinomial model, which we will discuss in the next section.)

3.24 Incomplete Markets (e.g. the Trinomial Model)

In our first proof, we derived the Black-Scholes equation based on the assumption of continuous tradability of the underlying asset and the derivative. In the second proof,

we no longer needed the assumption of continuous tradability of the derivative to derive the Black-Scholes formula.

The question now is the following: What if the underlying asset is also not permanently tradable during the derivative's life?

It is quite obvious that in this case, it will most certainly not be possible to replicate every derivative by a trading strategy with an underlying and a bond, which means that such a market can therefore no longer be complete.

However, a market can be incomplete for other reasons too.

We ask ourselves the question: In a market that is not complete, what, if anything, can then be said about fair prices of derivatives (that cannot be replicated in this market)?

The answers will be:

1. In many cases, it will no longer be possible to state a derivative's unique **exact** fair price. But it will be possible to specify a **range** $[a, b]$ that the price must lie in; otherwise (if the price were less than a or greater than b), arbitrage would be possible.
2. In some cases, it is possible to use further market data in order to be able to determine unique fair prices even for derivatives that cannot be replicated by the underlying instruments.

We will give an example for both the first case and the second case (in the next section) below.

The first example of an incomplete market is a **one-step trinomial model**. We present this model here not because we consider it to be of great practical relevance, but because it provides a very good illustration of how (and why) one arrives at (only) a range for the fair price of a derivative in incomplete markets, rather than an exact fair price.

For this reason, we will not discuss trinomial models and the valuation of derivatives in such models in general terms here, but will calculate a concrete numerical example.

The structure of a one-step trinomial model is analogous to that of a binomial one-step model:

- We have one single time step. In our numerical example, this time step extends from time 0 to time 1.
- For simplicity in our example, let the interest rate r for risk-free investments during this time step be $r = 0$. This way we don't need to consider interest calculations in the model.
- In this model, we consider a stock price that changes from value $S(0)$ to value $S(1)$.
- Let $S(0) = 90$ and let $S(1)$ have the possible values 150, 100, or 50.
- Let the value 150 be assumed with probability p , the value 100 with probability q , and the value 50 with probability $1 - p - q$ (with $p + q < 1$). We will see that the actual values of p and q will again be irrelevant to our valuations.

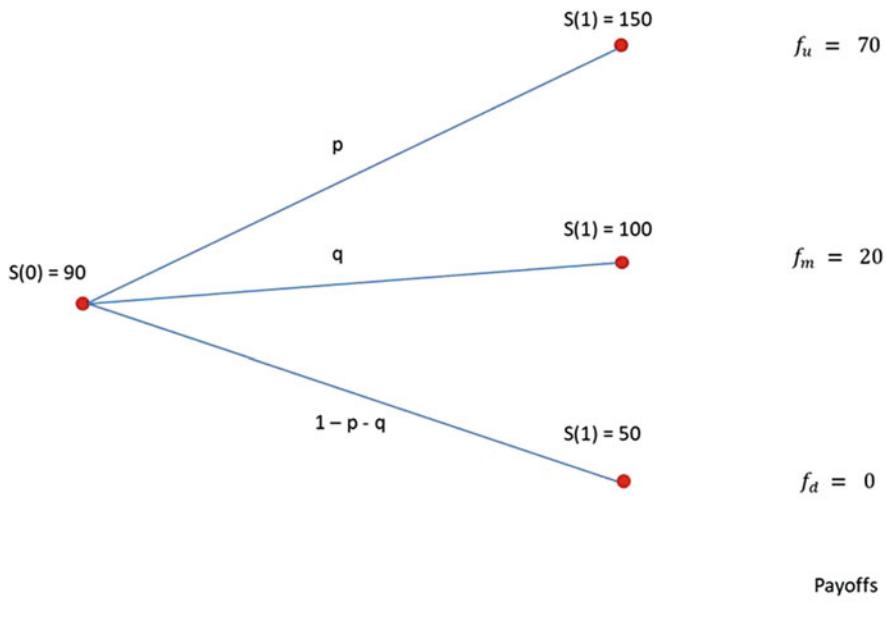


Fig. 3.21 Example of a one-step trinomial model with payoffs of a call option with strike 80

In this trinomial model, we now want to find or at least narrow down the fair price of a call option C on the underlying asset S with expiration at time 1 and with strike price $K = 80$.

The situation is shown in Fig. 3.21. The payoffs of the call option with strike $K = 80$ are denoted by f_u , f_m , and f_d .

In pricing derivatives in a one-step binomial model, we proceeded in the same way as in the proof of completeness of the Black-Scholes model in the previous section. We replicated the payoff of the derivative using a trading strategy based on the underlying asset and the risk-free interest rate. Replicating was possible and nothing special, in that the replicating portfolio could only be a static portfolio anyway: It was created at time 0 and had to replicate the payoff already at the next time point 1, so no further adjustment was possible or necessary.

If we now try to replicate the payoff of the derivative in the trinomial model:

- Then, at time 0, we create a portfolio consisting of x units of the underlying asset and y EUR in cash.
- We denote the value of the portfolio at time 0 by $V(0)$ and $V(0) = 90x + y$.
- The portfolio's value $V(1)$ at time 1 can be
 - $50x + y$ if $S(1)$ is 50.
 - $100x + y$ if $S(1)$ is 100.

- $150x + y$ if $S(1)$ is 150.
- If the portfolio were to replicate the option's payoff, then
 $50x + y = 0$,
 $100x + y = 20$ and
 $150x + y = 70$
would have to be satisfied.

It is easy to see that this system of equations

$$\begin{aligned} 50x + y &= 0 \\ 100x + y &= 20 \\ 150x + y &= 70 \end{aligned}$$

of three equations in two unknowns is not solvable. While it is possible to find a unique solution x, y for the first two equations, namely, $x = \frac{2}{5}$ and $y = -20$, substituting these two values in the third equation yields the value 40 and not 70 as required.

Thus, the call option cannot be replicated in this model.

Still, is there anything that can be said about the fair price (or a fair price range) of this option in this model? At what price $F(0)$ of the derivative at time 0 would arbitrage be possible? Obviously only if we either

- Succeed in constructing, at the same price $F(0)$ at time 0, a portfolio of underlying asset and cash whose value at time 1 is definitely always greater than or equal to the derivative's payoff and is truly greater than the derivative's payoff in at least one case
- Conversely succeed in constructing, at the same price $F(0)$ at time 0, a portfolio of underlying asset and cash whose value at time 1 is definitely always smaller or equal to the derivative's payoff and is truly smaller than the derivative's payoff in at least one case.

We therefore seek to find the smallest possible value b for the construction of a portfolio as in (i) for $F(0) = b$ and find the largest possible value for the construction of a portfolio as in (ii) for $F(0) = a$.

Then arbitrage is possible for all values of $F(0)$ with $F(0) \leq a$ or $F(0) \geq b$, whereas no arbitrage is possible for the values $F(0)$ with $a < F(0) < b$. We have thus found an arbitrage-free range (a, b) for $F(0)$.

Now, how do we find a and b ?

This is a simple problem from the field of linear programming. We can illustrate the solution geometrically for our simple numerical example. First, we calculate a : For this purpose, we draw the range in the (x, y) plane in which

$50x + y \leq 0$ (see Fig. 3.22, top left, blue shaded area),

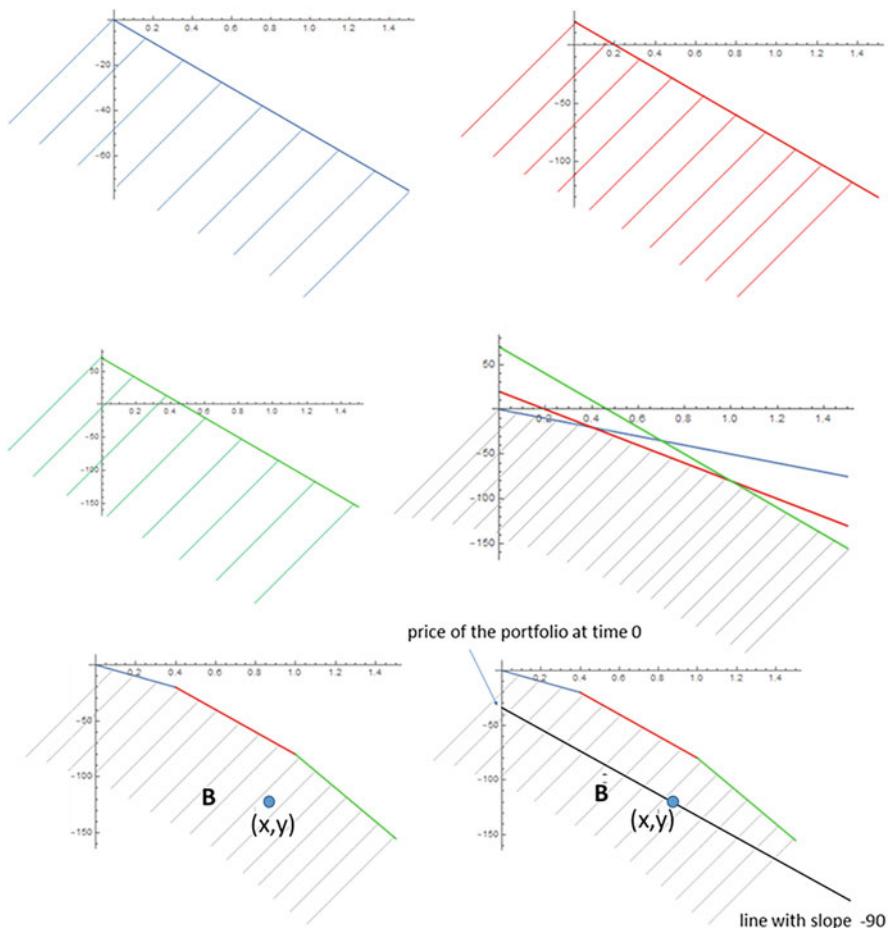


Fig. 3.22 Illustration of the solution to the linear programming problem for determining an arbitrage-free range for the price of a call option in a trinomial model

in which

$$100x + y \leq 20 \quad (\text{see Fig. 3.22, top right, red shaded area})$$

and in which

$$150x + y \leq 70 \quad (\text{see Fig. 3.22, second row left, green shaded area}).$$

Finally, we take the intersection of these three areas. This gives us the area B which satisfies $50x + y \leq 0$ and $100x + y \leq 20$ and $150x + y \leq 70$ (see Fig. 3.22, second row right). Below left, we reproduced the relevant part of the last figure once more.

The following holds for each point (x, y) in the interior of this area B : If a portfolio is created at time 0 from x units of the underlying asset S and y EUR

in cash, then this portfolio will have a value at time 1 that is certainly smaller than the payoff from the option.

The purchase cost of the portfolio $x \cdot S + y$ at time 0 is exactly $90x + y$ and can be gathered directly from the graph as follows: Draw a straight line with slope -90 through the point (x, y) . The intersection of this line with the y-axis gives exactly the value $90x + y$, that is, the purchase cost of the corresponding portfolio (see Fig. 3.22, bottom right).

Within that range B , we now look for the point (x, y) at which the portfolio $x \cdot S + y$ would require the greatest purchase cost $90x + y$ at time 0 and determine its purchase cost. This then is our value a !

If the derivative had a price less than a , then we could find a point (x, y) in the range B where the cost of buying the portfolio $x \cdot S + y$ at time 0 would be greater than the cost of the derivative but where the value at time 1 would certainly be less than the payoff from the derivative. Going short on this portfolio and buying the derivative would then lead to arbitrage.

But how do we find that point (x, y) in the range B for which the portfolio $x \cdot S + y$ would require the greatest purchase cost $90x + y$ at time 0?

Geometrically, the answer is quite simple: We move the black line in Fig. 3.22 upward (parallel to itself) until only one point of the range B still sits on that line. See Fig. 3.23. The portfolio point (x, y) that we were seeking to find is exactly where the blue line segment intersects the red line segment.

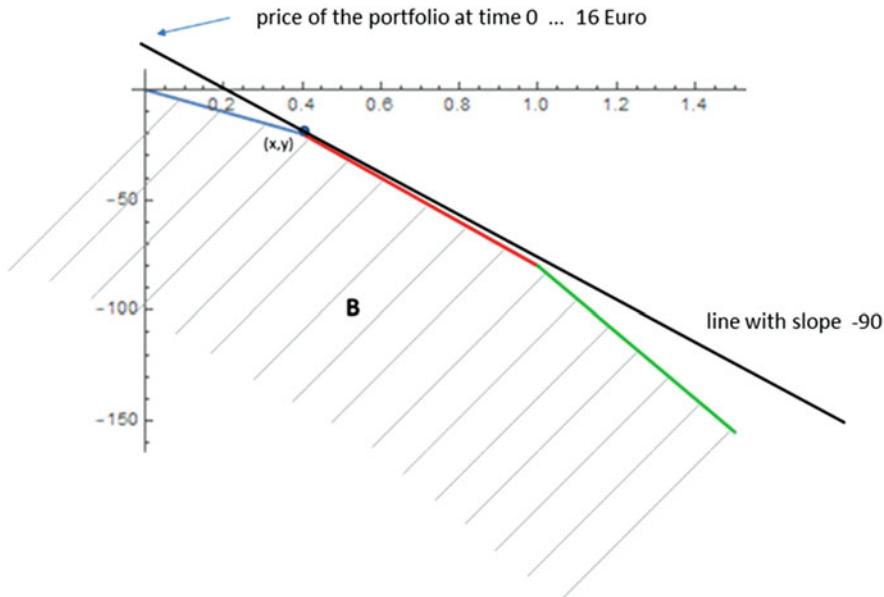


Fig. 3.23 Determining the portfolio in B with maximum purchase cost

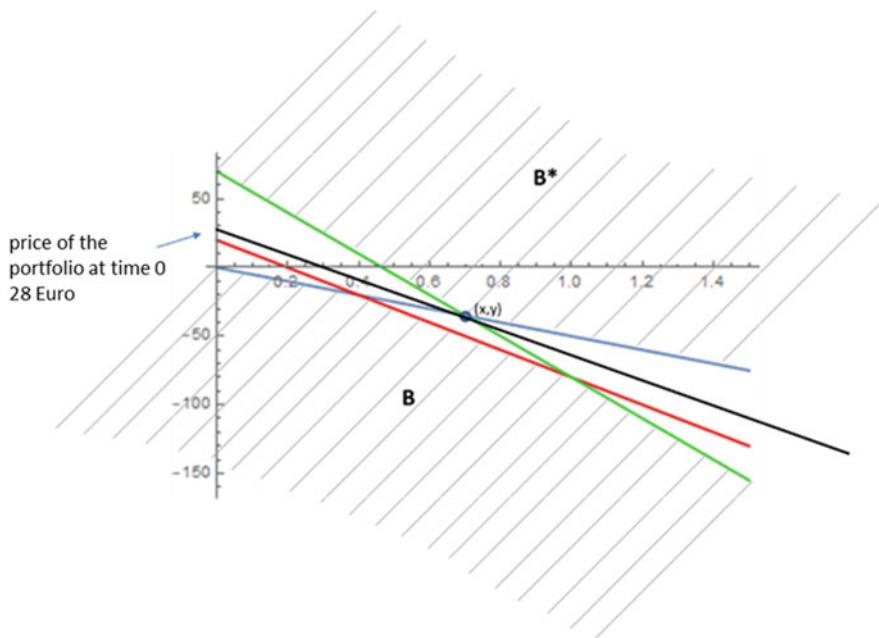


Fig. 3.24 Determining the range B^* and the portfolio in B^* with minimum purchase cost

The blue line has the equation $50x + y = 0$, the red line has the equation $100x + y = 20$, and the intersection (x, y) of the two lines therefore has the coordinates $(0.4, -20)$. The purchase cost for this portfolio and thus the value a that we were seeking is $90 \cdot 0.4 - 20 = 16$ EUR.

We thus obtain the value 16 as the lower bound for the call option's price.

To determine the upper bound b , we proceed precisely the other way around (see Fig. 3.24 for illustration of the procedure): We determine the area B^* that satisfies $50x + y \geq 0$ and $100x + y \geq 20$ and $150x + y \geq 70$). Then we move the black line downward until only one point of B^* still sits on that line. This “last remaining point (x, y) on the line” is exactly the intersection between the blue and the green line. We easily calculate that this point has the coordinates $(x, y) = (0.7, -35)$. This portfolio has the smallest possible purchase cost in B^* , namely, $0.7 \cdot 90 - 35 = 28$ EUR.

Using the analogous reasoning as above, it follows that the upper bound b of the no-arbitrage range for the call option must equal 28.

Thus, for the price of the call option in the trinomial model, we have found the maximum range $[16, 28]$ in which no-arbitrage opportunities exist. Arbitrage would be possible at any price of the option outside this interval.

The one-step trinomial model is only one financial market among many that is not complete.

We have already discussed on various occasions that the Wiener model, while useful, has certain disadvantages. In particular, the probability of the occurrence of extreme events in financial prices is often underestimated with the Wiener model.

Of course, there are mathematical models that model the fat tails of stock price dynamics significantly better than the Wiener model. These are special Levy processes, in particular the so-called hyperbolic distributions. But these models have a major disadvantage: Most of them are not complete! See, for example, [4].

3.25 Incomplete Markets (e.g. Non-tradable Underlying Asset)

Examples of derivatives on non-tradable underlying assets include weather derivatives or, in a certain sense, even sports bets. To some extent, the so-called CAT bonds (where repayment or interest depends on the occurrence of precisely defined events (often catastrophes, hence the name)) can also be considered to be derivatives on non-tradable underlying assets.

The underlying assets of weather derivatives are clearly defined weather conditions such as temperatures or rainfall at specific locations.

Most weather derivatives (and CAT bonds) are entered into OTC between investors (companies) and a bank or an insurance company and obviously serve investors primarily as an insurance product. The most common areas where such derivatives are used as insurance products are in risk management of (large-scale) farming, in the energy sector, and in tourism.

Weather derivatives are increasingly also exchange-traded. An overview of weather derivatives traded on the CME including their terms and conditions and what they are based on can be found at <https://www.cmegroup.com/trading/weather/>

For example, the CME defines certain temperature indices that some of the CME-traded weather derivatives are based on (e.g. the HDD and CDD indices).

The situation we now face in pricing derivatives on non-tradable underlying assets is as follows:

The financial market where we are now operating again consists of the two basic products B (bond, with constant interest rate r) and S (underlying), where B follows the dynamics $dB(t) = r \cdot B(t)dt$ and S follows the dynamics $dS(t) = \mu(t, S(t)) \cdot S(t)dt + \sigma(t, X(t)) \cdot S(t)dW(t)$ with deterministic functions μ and σ . Furthermore, there are (European) derivatives D in this financial market which are given by their payoff function Φ at their expiration date T .

However, we now assume that **the underlying S is not tradable** and that, conversely, the derivatives are continuously tradable throughout their lifetimes.

This would be the case, for instance, if we analysed weather derivatives traded on the CME (and assumed that the underlying instrument, such as the CDD temperature index, actually moves according to an Ito process S of the form $dS(t) = \mu(t, S(t)) \cdot S(t)dt + \sigma(t, X(t)) \cdot S(t)dW(t)$ with deterministic functions μ and σ).

To value a derivative D with expiration in T , payoff function Φ , and price process $F(t, S(t))$, it is necessary that another derivative E with the same expiration T and a price process G is available.

If this is the case, then the following will remind you of the procedure we used for deriving the formula of bond prices in Sect. 3.10:

F and G are functions of t and of the value $S(t)$ of the Ito process, and because of the Ito formula, they are themselves Ito processes. The representation of F and G as Ito processes is also obtained by applying the Ito formula. That's what we are going to do now. First, we recall again the Ito representation of $S(t)$, which we will use in the following in its short representation $dS(t) = \mu \cdot S \cdot dt + \sigma \cdot S \cdot dW(t)$.

Using the Ito formula, we obtain

$$dF = \left(F_t + \mu \cdot S \cdot F_s + \frac{(\sigma \cdot S)^2}{2} F_{ss} \right) dt + \sigma \cdot S \cdot F_s \cdot dW(t)$$

and

$$dG = \left(G_t + \mu \cdot S \cdot G_s + \frac{(\sigma \cdot S)^2}{2} G_{ss} \right) dt + \sigma \cdot S \cdot G_s \cdot dW(t).$$

We rearrange and rename these two representations somewhat and get

$$dF = F \cdot \alpha_F \cdot dt + F \cdot \beta_F \cdot dW(t) \quad \text{and} \quad dG = G \cdot \alpha_G \cdot dt + G \cdot \beta_G \cdot dW(t), \quad (3.23)$$

where

$$\begin{aligned} \alpha_F &= \frac{F_t + \mu \cdot S \cdot F_s + \frac{(\sigma \cdot S)^2}{2} \cdot F_{ss}}{F} \quad \text{resp.} \\ \alpha_G &= \frac{G_t + \mu \cdot S \cdot G_s + \frac{(\sigma \cdot S)^2}{2} \cdot G_{ss}}{G} \quad \text{and} \end{aligned} \quad (3.24)$$

$$\beta_F = \frac{\sigma \cdot S \cdot F_s}{F} \quad \text{resp.} \quad \beta_G = \frac{\sigma \cdot S \cdot G_s}{G}. \quad (3.25)$$

Now, we create a dynamic portfolio V from the derivatives D and E . Let this portfolio be defined such that at any point in time t we denote our current total assets in the portfolio by $V(t)$ and we invest $u(t) \cdot V(t)$ of these assets in the derivative D and the remainder, hence $(1 - u(t)) \cdot V(t)$ in the derivative E .

This means:

At any time t we hold (in shorthand notation) exactly $\frac{u \cdot V}{F}$ units of the derivative D and $\frac{(1-u) \cdot V}{G}$ units of the derivative E .

So the trading strategy for creating this portfolio is, in principle, such that we always invest only as much money in the strategy, that is, in the two derivatives, as there is value V in the portfolio. The strategy is thus **self-financing**. No additional money is added to or withdrawn from the portfolio at any time during the strategy's life.

For the value process V of the dynamic portfolio, we again have

$$dV = \frac{u \cdot V}{F} \cdot dF + \frac{(1-u) \cdot V}{G} \cdot dG = V \cdot \left(u \cdot \frac{dF}{F} + (1-u) \cdot \frac{dG}{G} \right) \quad (3.26)$$

From Formula (3.23), we know that

$$\frac{dF}{F} = \alpha_F \cdot dt + \sigma_F \cdot dW(t) \text{ and } \frac{dG}{G} = \alpha_G \cdot dt + \sigma_G \cdot dW(t).$$

Substituting this in Formula (3.26), we get

$$\begin{aligned} dV &= V \cdot (u \cdot (\alpha_F \cdot dt + \beta_F \cdot dW(t)) + (1-u) \cdot (\alpha_G \cdot dt + \beta_G \cdot dW(t))) = \\ &= V \cdot ((u \cdot \alpha_F + (1-u) \cdot \alpha_G) \cdot dt + \\ &\quad + (u \cdot \beta_F + (1-u) \cdot \beta_G) \cdot dW(t)) \end{aligned} \quad (3.27)$$

And now we make an explicit choice for the particular form of our dynamic portfolio. That is, we choose u explicitly. And we choose u in such a way that the random component in the portfolio's dynamics (i.e. the part that is governed by the Brownian motion) is eliminated. Thus

$$u \cdot \beta_F + (1-u) \cdot \beta_G = 0 \text{ and thus } u = \frac{\beta_G}{\beta_G - \beta_F} \text{ and } 1-u = \frac{-\beta_F}{\beta_G - \beta_F}.$$

Substituting this choice for u in the Formula (3.27) for dV , we get

$$dV = V \cdot \left(\frac{\beta_G}{\beta_G - \beta_F} \cdot \alpha_F - \frac{\beta_F}{\beta_G - \beta_F} \cdot \alpha_G \right) dt.$$

The performance $V(t)$ of this portfolio V is thus deterministic, and we conclude

$$\frac{\beta_G}{\beta_G - \beta_F} \cdot \alpha_F - \frac{\beta_F}{\beta_G - \beta_F} \cdot \alpha_G = r$$

We rearrange this equation somewhat and get

$$\Leftrightarrow \frac{\alpha_F(t) - r}{\beta_F(t)} = \frac{\alpha_G(t) - r}{\beta_G(t)}$$

Now, this is a highly remarkable equation: Both sides of the equation have variables that are strikingly reminiscent of a Sharpe ratio (trend term minus interest rate divided by volatility). The left side shows exactly this expression for the derivative D , and the right side shows exactly this expression for the derivative E . We call this expression (trend term minus interest rate divided by volatility) the respective derivative's "**market price of risk**" and denote it by $\lambda_F(t)$ resp. $\lambda_G(t)$.

The essential insight yielded by that last formula is again the following: Regardless of how D or E was chosen, the market price of risk at time t will always have the same value, irrespective of the value of D and of E ! So this market price of

risk is a function of the entire market and will always have the same value regardless of which derivative we are analysing. We can thus write this market price of risk function simply as a function $\lambda(t)$ of time (and irrespective of D or E).

We assume in the following that we know the market price of risk $\lambda(t)$ of this market (for purposes of pricing the derivative D , we could, for example, try to extract it from a liquid derivative E on the same underlying). We would then have

$$\frac{\alpha_F(t) - r}{\beta_F(t)} = \lambda(t)$$

Here, we substitute $\alpha_F(t)$ and for $\beta_F(t)$ by their representation from Formulas (3.24) and (3.25), thus

$$\alpha_F = \frac{F_t + \mu \cdot S \cdot F_s + \frac{(\sigma \cdot S)^2}{2} \cdot F_{ss}}{F} \quad \text{and} \quad \beta_F = \frac{\sigma \cdot S \cdot F_s}{F}$$

and rearranging, we obtain the following equation for the fair value F of the derivative D

$$F_t + (\mu - \lambda \sigma) \cdot S \cdot F_s + \frac{(\sigma \cdot S)^2}{2} \cdot F_{ss} - r \cdot F = 0$$

with the obvious boundary condition $F(T, s) = \Phi(s)$.

Thus, for the fair price of the derivative D , we have again derived the Black-Scholes formula but with a term dependent on the market price of risk ($\mu - \lambda \cdot \sigma$). This partial differential equation can again be solved using the Feynman-Kac formula, and the solution we obtain is again a representation in a form we are already familiar with. We have

$$F(t, s) = e^{-r(T-t)} \cdot E \left(\Phi \left(\tilde{S}(T) \mid \tilde{S}(t) = s \right) \right),$$

where $\tilde{S}(t)$ follows the dynamics $d\tilde{S}(t) = (\mu - \lambda(t) \cdot \sigma) \cdot \tilde{S}(t) dt + \sigma \cdot \tilde{S}(t) dW(t)$.

In conclusion, we note that this version of the Black-Scholes formula also includes the original Black-Scholes formula. That is because, if S is in fact tradable, then we can simply use S itself instead of the derivative E in the above argument. And the market price of risk calculated from S is (since $dS(t) = \mu \cdot S \cdot dt + \sigma \cdot S \cdot dW(t)$) exactly $\lambda(t) = \frac{\mu - r}{\sigma}$. The term $\mu - \lambda \cdot \sigma$ then becomes r , and thus the dynamics of $\tilde{S}(t)$ becomes $d\tilde{S}(t) = r \cdot \tilde{S}(t) dt + \sigma \cdot \tilde{S}(t) dW(s)$. And with that, we are back at the original Black-Scholes formula.

References

1. Damiano Brigo and Mercurio Fabio. *Interest Rate Models—Theory and Practice*. Springer Finance, 2007.
2. Tomas Björk. *Arbitrage Theory in Continuous Time*. Oxford Finance Series, 2009.
3. Michael J Steele. *Stochastic Calculus and Financial Applications*. Springer, 2000.
4. Gerhard Larcher, Martin Predota, and Robert Tichy. “Arithmetic average options in the hyperbolic model”. In: *Monte Carlo Methods and Appl.* 9.3 (2003), pp. 227–239.