Christian Ekstran

Financia Derivativ

Financial Derivatives Modeling

Christian Ekstrand

Financial Derivatives Modeling

Springer

Christian Ekstrand
Stockholm
Sweden
christian.ekstrand@seb.se

ISBN 978-3-642-22154-5 e-ISBN 978-3-642-22155-2
DOI 10.1007/978-3-642-22155-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011936378

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The purpose of this book is to give a comprehensive introduction to the modeling of financial derivatives, covering the major asset classes and stretching from Black and Scholes' lognormal modeling to current-day research on skew and smile models. The intended reader has a solid mathematical background and works, or plans to work, at a financial institution such as an investment bank or a hedge fund. The aim of the book is to equip the reader with modeling tools that can be used in the (future) work involving derivatives pricing, trading, or risk management.

The field of derivatives modeling is extensive and to keep the book within a reasonable size, certain sacrifices have been made. For instance, the implementation of models is not discussed as this can be viewed as an art rather than science and is therefore an ungrateful subject for a text book. Minor asset classes, such as inflation products, and asset classes that require specific mathematical tools, e.g., credit and mortgage products, have been left out. Furthermore, the financial basics are covered at a faster pace than in other introductory books to the area. For example, the martingale theory is summarized in a compact appendix, and the introduction to the Black–Scholes model is done by working directly in continuous space-time, in contrast to the pedagogical approach of initially reviewing the binomial model. This enables us to quickly go beyond the Black–Scholes framework and thereby focus on skew and smile models and on derivatives in specific asset classes.

The book is divided into four parts. The first part consists of Chaps. 1–4 and contains the general framework of derivatives pricing. This part is essential for the understanding of the rest of the book. An exception is Chap. 4 which a novice reader might find too abstract and is advised to skip and come back to later when the necessary financial maturity has been reached. The rest of the book consists of chapters that can be read independently. Chapters 5–8 cover skew and smile modeling. The pricing of exotic derivatives is the subject of the third part, Chaps. 9–10. The concluding fourth part comprises Chaps. 11–14 and applies the pricing methods to specific asset classes.

Stockholm *Christian Ekstrand*

Contents

Part I
Derivatives Pricing Basics

Chapter 1
Pricing by Replication

This chapter provides an introduction to the theory of derivatives pricing. We start by defining the fundamental objects – the underlyings – that the theory depends upon. We state the conditions that the underlyings are assumed to satisfy and explain how the theory can be applied. The presentation here has an abstract character while the remainder of the book contains specific examples of models based on this theory.

1.1 Underlyings and Derivatives

The theory of derivatives pricing is based on a set $\{S^i\} = \{S^1, S^2, S^3, \ldots\}$ of predefined financial assets that can be stocks, bonds, etc. The price of an asset S is a real number which we also denote by S, or by S_t when we want to emphasize the time dependence. We assume that today's prices $\{S^i_{t=0}\}$ are given and refer to these assets as the *underlyings* of the theory.

We are interested in pricing contracts V for which the prices at time T are known as expressions of the price S_T of an underlying. As the future values can be derived from the values of the underlyings, these contracts are called *derivatives*. Examples include $V_T = (S_T)^2$ or $V_T = S_T - K$ for a fixed K, but also more general payoff types such as V_T depending on the values of several underlyings at T or on the average value of S_t attained in the time interval $t \in [0, T]$.

We show later that the *present value (PV)* $V_{t=0}$ of a derivative can be computed using only a couple of natural assumptions. We are not, however, interested in computing the present values of the underlyings. The reason is that the underlyings are often too complex to be handled in a universal pricing framework. For instance, the price of an equity stock depends on multiple hard-to-measure factors including the employees' morale, the interaction between the divisions of the company, the management's decisions and the state of the world economy. Instead, our philosophy is that the underlyings are correctly priced through the supply and demand by market participants.

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_1,

1.2 Assumptions

To set up a theoretical framework for derivatives pricing, it is necessary to impose certain conditions on the underlyings. We assume that the underlyings are liquidly traded meaning that they can be bought and sold at any instance in time with equal bid and offer prices. We allow S_t to be equal to any real value (or positive real value), i.e. we let the tick size be zero. When purchasing an asset, it takes a couple of days until the asset and the payment change hands. This time period is called the settlement lag and is set to zero for simplicity.

We allow for assets to be shorted, i.e. a negative number of assets can be held. In practice, assets can be shorted by entering futures or forward contracts. Alternatively, assets can be borrowed (typically from a broker) and then sold whereafter they are bought back and returned at a later time. We assume that the underlyings are non-defaultable and that there are neither any costs associated with holding the underlyings (e.g. storage costs) nor any cash flows generated by them (e.g. dividends).

The markets are assumed to be efficient: there are no dominant market participants and no market manipulation, the markets are unregulated, the decisions made by the market participants are based solely on financial arguments, there is no shortage of cash, etc. Furthermore, all market participants are assumed to have excellent credit rating, which means that they never default.

In financial modeling, the *zero-coupon bond* P_{tT} is a particularly useful instrument. It pays \$1 (1 unit of the currency under consideration) at T for certain and has no other cash flows either before or after T. It can also be viewed as a loan taken at t that together with the interest rate yields a repayment of \$1 at T. P_{tT} is also called the *discount factor* from T as it measures the time t value of \$1 at T. As derivatives pricing involves discounting cash flows, we impose the same conditions on zero-coupon bonds as for the underlyings (indeed, the bond itself acts as an underlying for interest rate derivatives). This means, in particular, that P_{0T} are assumed to be given for all T.

The assumptions are not made because we believe that they are satisfied in real markets but to obtain a theory that is as simple as possible. In practice, it is often necessary to take into consideration the fact that the assumptions are violated. Some of the violations can be taken care of with minor adjustments to the theory. For example, in Sect. 3.11 we describe how a careful discounting accounts for a non-zero settlement lag. Other violations, such as a non-zero bid-offer spread, require more general models.

With only a few exceptions, we choose not to consider such generalized models. One of the reasons for this decision is that the question as to which assumption has the greatest impact on the derivatives price is often hard to answer and depends on both the type of the underlying and the derivative, and can even change during the lifetime of the derivative. Furthermore, there exist no well-established models that take such effects into account.

Generalized models often involve complex and thereby lower performing computer implementations. As performance is of crucial importance for many market participants, a generalized model is sometimes not a viable option. The alternative is to accept the model uncertainty arising from the violations of the assumptions. It means that if we sell a derivative, the customer needs to be charged an extra premium for the model risk we undertake. The size of the premium is delicate as it should be large enough to compensate for the model risk but small enough for the price to remain competitive. One of the main reasons for employing skilled quantitative analysts (colloquially known as quants) is to reduce the model risk, which leads to a lower premium. The result is a higher competitiveness and more deals won over rival firms.

Parts I–III of the book cover pricing of derivatives with underlyings that satisfy the idealized assumptions above. As we move on to the pricing of real-life derivatives in Part IV, it is necessary to relax some of the constraints.

1.3 The No-Arbitrage Assumption

Let $V = \sum_i b_i S^i$ be a financial portfolio, i.e. a weighted sum of financial assets, where the b_is are real numbers. We allow the portfolio to be rebalanced at each instance in time, whereby some assets are sold while others are bought. The restructuring of the content is called the *strategy* of the portfolio. A *self-financing strategy* has no in- or out-flux of cash, i.e. each purchase is exactly funded by a sale. Unless stated otherwise, we only consider self-financing strategies and therefore refer to them simply as strategies.

A strategy V is said to be an *arbitrage strategy* if it has zero initial value $V_0 = 0$, is always positive $P(V_T \geq 0) = 1$ and strictly positive with a non-zero probability $P(V_T > 0) > 0$ for a given future time T. An arbitrage strategy permits a possible positive future cash flow without any downside risk. Any arbitrage strategy that exists in a market is therefore taken advantage of by traders until supply and demand forces have adjusted the prices so much that the arbitrage disappears. As a result, arbitrage strategies are rare, and when they exist, they only do so for a short time. For this reason, it is possible to base a theory on the assumption that arbitrage strategies do not exist. This assumption has far reaching consequences and is the base of derivatives pricing.

An immediate implication of the *no-arbitrage assumption* is that certain financial strategies must be excluded from the theoretical framework. For instance, consider an investment into an *overnight deposit*, i.e. a loan that starts today and ends tomorrow. When tomorrow comes, we reinvest the proceeds in a new overnight deposit that ends the day after tomorrow. As interest rates are always non-negative (with a few peculiar historical exceptions), repeating this procedure results in a strategy with earnings greater than the holding of the amount in cash. We conclude that the combined strategy of being long the above strategy and short cash is an arbitrage strategy.

To exclude arbitrage strategies of the above type, it is necessary to remove the substrategy with the smallest payoff, which in this instance is the holding of cash. From now on, we assume that all suboptimal strategies are excluded from the theoretical framework. This leaves us with a setting that is completely arbitrage free.

1.4 Replication

Let us now turn to a particular implication of the no-arbitrage assumption that serves as the foundation for derivatives pricing. For this purpose, consider two strategies V and U for which we know for certain that $V_T \geq U_T$ at a given future time T. We claim that $V_0 \geq U_0$. The statement can be proven by showing that $U_0 = V_0 + k$, for $k > 0$, leads to a contradiction. The strategy $X = V + (kP_{0T}^{-1})P_{tT} - U$ of being long V, short U and long kP_{0T}^{-1} bonds maturing at T satisfies $X_0 = 0$ and $X_T > 0$, violating the no-arbitrage assumption. The argument can be generalized to arbitrary times t before T: $P(V_T \geq U_T) = 1 \Rightarrow P(V_t \geq U_t) = 1$. Indeed, if this were not true, there would exist a scenario for which $V_t < U_t$ and $P(V_T \geq U_T) = 1$, violating the no-arbitrage assumption when using the time t as a starting point.

The direction of the inequality can be reversed by interchanging the roles of V and U. Together with the original inequality, we arrive at the following conclusion: if two strategies for certain are equal at a future date, $P(V_T = U_T) = 1$, their values at any earlier time $t \leq T$ must be equal, $P(V_t = U_t) = 1$. In particular, when t is today's date we obtain $V_0 = U_0$ (Fig. 1.1).

To determine today's price V_0 of a complex contract V, the no-arbitrage assumption can be applied in the following way: assume that it is possible to construct a strategy U which is worth as much as V at a future time T (independently of the scenarios followed by the market) and for which the present value U_0 is known. The no-arbitrage principle implies that V_0 equals U_0. In fact, as the portfolios have equal values for all $t < T$, the strategy U is said to replicate V.

To reconnect with Sect. 1.1, we are typically interested in the pricing of a contract V for which the future value V_T is known as an expression of underlying values $\{S_T^i\}_i$. We construct a replicating portfolio U consisting of underlyings. The current value U_0 can be determined from $\{S_0^i\}$, which are assumed to be known. The

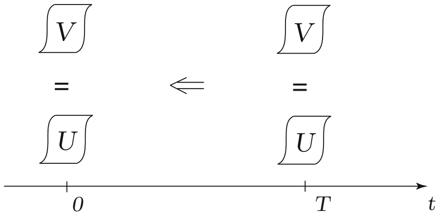

Fig. 1.1 Contracts that for certain have the same future values must have the same present values

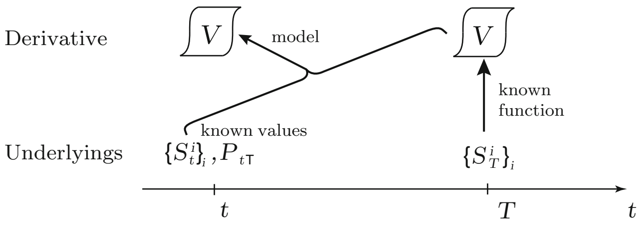

Fig. 1.2 Modeling of a derivative with a single cash flow

no-arbitrage principle then allows us to find the present value of V through $V_0 = U_0$ (Fig. 1.2).

In Chap. 2, we restrict ourselves to *static replication* for which the content of U is set up at $t = 0$ and held until T, without any additional trading between $t = 0$ and T. This is in contrast to *dynamic replication*, covered in Chap. 3, for which the content of U is changed through time in order to replicate V.

It is sometimes not possible to find a practically applicable strategy that replicates V. It can then be useful to construct a strategy U that superreplicates V, i.e. $V_T \leq U_T$, which gives an upper bound on today's price: $V_0 \leq U_0$. In the same manner, a strategy that subreplicates V leads to a lower bound.

Chapter 2
Static Replication

A portfolio is said to be static if it is unmanaged, which means that the content is not changed through time. In this chapter we review some important situations where static replication can be used for pricing or for finding upper and lower bounds on prices.

We start with pricing forward contracts and general fixed-time payments. We then derive constraints on option prices in preparation for the next chapter, where options are priced with dynamic replication. Finally, the method of static replication is applied to more exotic contracts such as early exercisables and barrier options.

2.1 Forward Contracts

Under the specifications of a *forward contract*, the counterparties are obliged to exchange a certain underlying S for a *strike* price K at a given *maturity* T. The contract is therefore worth $S - K$ at T.

The pricing of a forward contract is trivial as we can immediately conclude that the time t value is $S - KP_{tT}$. Indeed, if we own this amount at t, by selling K bonds maturing at T, enough money is generated for a purchase of the underlying S. The strategy is worth $S - K$ at maturity, which is the same amount as that of the forward contract. The no-arbitrage principle implies that the forward contract must be worth $S - KP_{tT}$.

The cash amount K that is used in the initiation of a forward contract is by convention such that the contract *values to par*, i.e. the price equals to zero. This value of the strike is called the *forward*. It is usually denoted by F and is equal to $P_{0T}^{-1}S$.

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_2,

2.2 European Options

A *European call option* gives the contract owner the right to buy the underlying S at time T for a given amount K. In contrast to forward contracts, the owner is not committed to the purchase but only does so if it is profitable. The option value at maturity is therefore $(S_T - K)_+ = \max(S_T - K, 0)$.

An option is said to be *in the money (ITM)* if $S > K$, *out of the money (OTM)* if $S < K$ and *at the money (ATM)* if $S = K$. Sometimes the forward is used in this classification, i.e. the conditions $F > K$, $F < K$ and $F = K$ are used to define whether an option is ITM, OTM or ATM. It is usually clear from the context which definition is used.

A *European put option* gives the owner the right to sell the underlying. The value at maturity is $(K - S_T)_+$. A *digital European call option* pays \$1 if $S_T > K$ and 0 otherwise. Thus, the value at maturity is $\theta(S_T - K)$, where θ is the *Heaviside function*. Similarly, a *digital European put option* is worth $\theta(K - S_T)$ at T.

Options belong to the type of contracts that cannot be priced with static replication. We postpone the pricing of these contracts to the next chapter in which dynamic replication is introduced. For the remainder of this chapter, we assume that option prices are known and use them for the static replication of more complex contracts.

2.3 Non-Linear Payoffs

We determine the present value of a contract V that pays $h(S)$ at T for a fixed, but arbitrary, function h with a well-defined second derivative. We have already covered the special case $h(S) = S - K$, for which the pricing can be done using the present values of the underlying S and the zero-coupon bond maturing at T. In the same manner, any linear payoff $h(S) = \alpha S - K$ can be priced. When $h(S)$ is non-linear, on the other hand, additional information is needed. The present values of European call options maturing at T turn out to provide sufficient information. This statement is made clear by the following computation:

$$
\begin{aligned}
h(S) &= \int_0^\infty h(K)\delta(S-K)dK \\
&= \int_0^\infty \left(-\frac{d}{dK}\left(h(K)\theta(S-K)\right) + h'(K)\theta(S-K)\right)dK \\
&= h(0) + \int_0^\infty h'(K)\theta(S-K)dK \\
&= h(0) + \int_0^\infty \left(-\frac{d}{dK}\left(h'(K)(S-K)_+\right) + h''(K)(S-K)_+\right)dK \\
&= h(0) + h'(0)S + \int_0^\infty h''(K)(S-K)_+ dK
\end{aligned}
$$

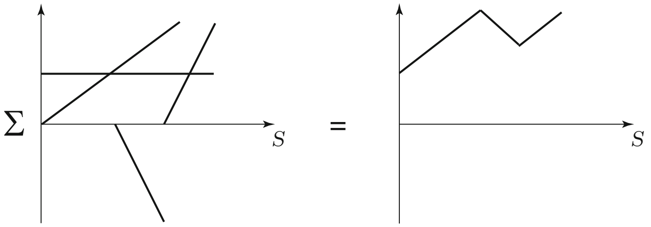

Fig. 2.1 Replication of an arbitrary payoff with zero-coupon bonds, the underlying and call options

Thus, if we at time $t = 0$ buy $h(0)$ number of bonds maturing at T, $h'(0)$ number of underlyings S and $h''(K)$ number of options with strikes in $[K, K + dK]$, for all K, then the contract is worth $h(S)$ at T (Fig. 2.1). The no-arbitrage principle implies that

$$V = h(0)P_{0T} + h'(0)S + \int_0^\infty h''(K)V^{\mathrm{C}}(K)dK$$

where $V^{\mathrm{C}}(K)$ is the present value of a call option with strike K.

Fixed-time payoffs can also be statically replicated with other option types. For example, according to the third line in the calculation above, digital calls options can be used. In this instance, the underlying S is not needed for the replication.

We move on to discuss which option type is preferable in the replication, call options or digital call options. From a theoretical point of view, the question is irrelevant as the two product types can be statically replicated from each other. For instance, using the third line in the above equation for $h(S) = (S - K)_+$ gives

$$(S - K)_+ = \int_0^\infty \theta(K' - K)\theta(S - K')dK' = \int_K^\infty \theta(S - K')dK'$$

Conversely, the relation $\theta(S - K) = -\frac{d}{dK}(S - K)_+$ shows that a digital call option can be approximated by a *call spread*, i.e. the difference between two call options, having the following payoff at T:

$$\frac{1}{\Delta K}\left((S - K + \Delta K)_+ - (S - K)_+\right)$$

We conclude that a digital call option can be approximated by two call options while a large number of digitals are needed to approximate a call option. This suggests that call options should be used in static replication of contracts paying $h(S)$. The main reason for using call options, however, is that they are more liquid market instruments than digital call options.

The replicating formula is also applicable to payoffs with a discontinuous (mathematical) derivative. Consider, for example, a put option paying $h(S) =$

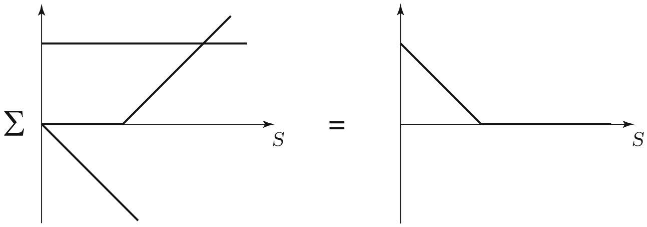

Fig. 2.2 Put-call parity

$(K-S)_+$ at T. Using $h'(S) = -\theta(K-S)$ and $h''(S) = \delta(K-S)$ in the replication formula gives:

$$V^{\mathrm{P}}(K) = KP_{0T} - S + V^{\mathrm{C}}(K)$$

This relation is called *put-call parity* and shows that a call and a put option only differ by a linear payoff (Fig. 2.2). The parity is obvious as the difference in payoff at maturity

$$(S-K)_+ - (K-S)_+ = S - K$$

is equal to the payoff of a forward contract.

Since a put equals a call up to a linear payoff, some of the calls in the replication formula can be replaced with puts. As puts are cheaper than calls when the strike is low, the cost of the options in the replication formula can be reduced by replacing low strike calls with puts. The details can be understood from the computation

$$\begin{aligned}
&\int_0^{\bar{K}} h''(K)(K-S)_+ dK + \int_{\bar{K}}^{\infty} h''(K)(S-K)_+ dK \\
&= h'(\bar{K})(\bar{K}-S)_+ - \int_0^{\bar{K}} h'(K)\theta(K-S)dK \\
&\quad -h'(\bar{K})(S-\bar{K})_+ + \int_{\bar{K}}^{\infty} h'(K)\theta(S-K)dK \\
&= h'(\bar{K})(\bar{K}-S)_+ - h(\bar{K})\theta(\bar{K}-S) + \int_0^{\bar{K}} h(K)\delta(K-S)dK \\
&\quad -h'(\bar{K})(S-\bar{K})_+ - h(\bar{K})\theta(S-\bar{K}) + \int_{\bar{K}}^{\infty} h(K)\delta(S-K)dK \\
&= h'(\bar{K})(\bar{K}-S) - h(\bar{K}) + h(S)
\end{aligned}$$

The no-arbitrage principle implies that

$$\begin{aligned}&\int_0^{\bar{K}} h''(K)V^{\mathrm{P}}(K)dK + \int_{\bar{K}}^{\infty} h''(K)V^{\mathrm{C}}(K)dK\\&= \big(h'(\bar{K})\bar{K} - h(\bar{K})\big)\,P_{0T} - h'(\bar{K})S + V\end{aligned}$$

which shows how a fixed-time payoff can be replicated with low strike puts and high strike calls. Denoting the right-hand side with $g(\bar{K})$, we obtain

$$\begin{aligned}g'(\bar{K}) &= h''(\bar{K})(\bar{K}P_{0T} - S)\\g''(\bar{K}) &= h'''(\bar{K})(\bar{K}P_{0T} - S) + h''(\bar{K})P_{0T}\end{aligned}$$

Assume for a moment that the second derivative of h is positive. The only extreme point of $g(\bar{K})$ is then a minimum located at the forward $\bar{K} = P_{0T}^{-1}S = F$. We conclude that

$$V = \big(h(F) - h'(F)F\big)\,P_{0T} + h'(F)S + \int_0^F h''(K)V^{\mathrm{P}}(K)dK + \int_F^{\infty} h''(K)V^{\mathrm{C}}(K)dK$$

is the replication of V that has the cheapest option content. The same result is obtained if h has a negative second derivative.

Liquid European options are found in the market only for a finite set of strikes. It means that the static replication strategy for non-linear payoff is not directly applicable in practice. Instead, European option prices for arbitrary strikes are typically inferred from the liquid market quotes by mathematical interpolation. Once this has been done, static replication can be used. As the outcome depends on the interpolation scheme, different market participants arrive at different conclusions regarding the price. This is particularly apparent when the payoff depends on strikes outside the liquid range, making extrapolation a necessity.

2.4 European Option Price Constraints

Before constructing option pricing models, it is useful to derive the asymptotic limits and the no-arbitrage conditions that a European call option price V has to satisfy. These conditions can be used to exclude inappropriate models. As the corresponding constraints for put options follow from put-call parity, it is sufficient to focus on European call options.

Consider first the asymptotic behavior of V: for very large values of K the call option is worthless, $V = 0$. In the limit of small values of K, the option certainly gets exercised at maturity. The option holder then needs the amount KP_{0T} today to pay the strike price K at T in order to receive S. Today's value of the contract is therefore $S - KP_{0T}$.

We proceed to the no-arbitrage conditions and observe that the value of a contract paying $h(S) \geq 0$ at T is obviously positive. As $h(S) = \int_0^{\infty} h(K)\delta(S-K)dK$, this requirement is equivalent with demanding positivity of a contract paying $\delta(S-K)$. Using

$$\delta(S-K)=\frac{d^2}{dK^2}(S-K)_+=\frac{d^2}{dK^2}V_T$$

which is the $\Delta K\to 0$ limit of $(\Delta K)^{-2}\left(V_T(K+\Delta K)-2V_T(K)+V_T(K-\Delta K)\right)$, we obtain the constraint $\frac{d^2}{dK^2}V\geq 0$.

There is also a constraint for option combinations with different maturities: V_2-V_1 is positive for V_1 and V_2 European call options with strike K and maturities $T_1<T_2$. The statement can be verified by proving that a portfolio of one long unit of V_2 and one short unit of V_1 is always positive at T_1. It is sufficient to consider the situation when V_1 is ITM at T_1, giving a portfolio value of V_2-S+K. From the conditions $V(K\to\infty)\to 0$, $V(K\to 0)\to S-KP_{0T}$ and $V''(K)\geq 0$ it follows that $V(K)\geq S-KP_{0T}$, which implies that $V_2(t=T_1)\geq S-KP_{T_1T_2}\geq S-K$, proving the statement. Observe that $V_2\geq V_1$ is equivalent with the infinitesimal condition $\frac{dV}{dT}\geq 0$.

In summary, the following constraints must be satisfied by the European call option price (Fig. 2.3):

- $V(K\to 0)\to S-KP_{0T}$
- $V(K\to\infty)\to 0$
- $V(T\to 0)\to(S-K)_+$
- $\dfrac{d^2}{dK^2}V\geq 0$
- $\dfrac{d}{dT}V\geq 0$

From these fundamental constraints, it is possible to derive other interesting conditions on the option price. For instance, from conditions 1, 2 and 4, we obtain upper and lower bounds on the European call option price and its (mathematical) derivative with respect to the strike (which is the digital option price).

$$\begin{aligned}(S-KP_{0T})_+ &\leq V \leq S\\ -P_{0T}=\left.\frac{dV}{dK}\right|_{K=0} &\leq \frac{dV}{dK}\leq \left.\frac{dV}{dK}\right|_{K=\infty}=0\end{aligned}$$

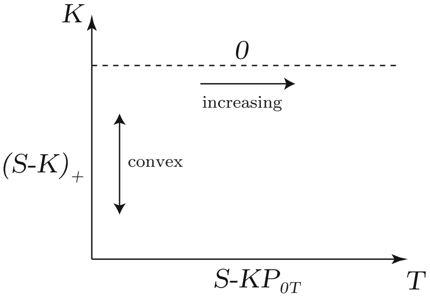

Fig. 2.3 No-arbitrage conditions and asymptotics for European call option prices

2.5 American and Bermudan Options

American options can be exercised any time up to the maturity T. For example, an American call option exercised at $t < T$ gives a payment $S - K$ at t. This is in contrast to European options which can only be exercised at maturity T. *Bermudan options* are something in between (just as Bermuda lies somewhere between Europe and America): they can only be exercised at certain prespecified dates. The extra optionality makes an American option more valuable than a Bermudan option, which in turn is worth more than a European option. There are, however, many instances where this extra optionality is worthless and all option types have equal value.

We saw in the previous section that $(S - KP_{0T})_+$ is a lower bound for the European call option price. As this amount is greater than the exercise value $S - K$, a call option should never be exercised early. We conclude that American, Bermudan and European call options have equal prices. Observe that if we permit the underlying to have cash flows such as dividend payments, there can be situations for which it is optimal to exercise early in order to obtain these cash flows.

For European put options, the lower bound $(KP_{0T} - S)_+$ is below $K - S$ for $S < K$ which means that there are instances when an early exercise is preferable. American and Bermudan put options are therefore worth strictly more that their European counterpart. To find an upper price bound, observe that, because of the possibility to exercise early, American and Bermudan put options with a time-dependent strike KP_{tT} must be worth more than a European put option with strike K. However, as $(KP_{tT} - S)_+$ is a lower bound for the European price, an early exercise is not feasible as it yields $KP_{tT} - S$. We conclude that the American, Bermudan and European put options have equal prices in this instance of a time-dependent strike. As put option prices increase with the strike, American and Bermudan put options with strike KP_{0T} are worth less than the corresponding options with strike P_{tT}, which in turn is worth as much as a European put option with strike K. Replacing K with KP_{0T}^{-1}, we conclude that American and Bermudan put options with strike K are bounded from below by the European put option with strike K and from above by the European put option with strike KP_{0T}^{-1}.

The argument leading to put-call parity for European options do not carry through to American and Bermudan options. Instead, the parity relation can be replaced with an upper and lower bound on the put option price when formulated in terms of the call option, or vice versa. Indeed, using put-call parity on the European put options in the above bounds together with the fact that American, Bermudan and European call options are worth equally much, we obtain

$$KP_{0T} - S + V^{\mathrm{C}} \leq V^{\mathrm{P}} \leq K - S + V^{\mathrm{C}}$$

where we have used the common notation V for both American and Bermudan options. The right-hand side option has strike KP_{0T}^{-1} which because of decaying call prices with strike values can be replaced with strike K.

Early exercise decisions for call and put options with zero strike value are particularly simple to analyze. Consider first the trivial situation of a call option on an underlying that is restricted to being positive. As there is no cost in exercising the option, we definitely do that at some point in time and it does therefore not matter when the exercise is made. When the underlying can be negative as well as positive, it is suboptimal to exercise early in the zero strike case. Indeed, exercising early and holding the underlying to maturity is associated with the risk of the underlying becoming negative, which can be avoided by postponing the exercise till maturity. A swaption, i.e. an option on a swap, is an example of a zero strike option on an underlying that can be negative.

2.6 Barrier Options

We consider *barrier options* that are of call type, i.e. they pay $(S - K)_+$ at T if the underlying S breached (or did not breach) a barrier level B sometime between $t = 0$ and T. Options of put type can be treated in a parallel way. A barrier option is said to be of knock-out type if the payment occurs conditional on that the barrier was not breached. If the barrier needs to be breached for the payment to occur, the option is said to be of knock-in type. For instance, the payoff for a knock-out call option with a lower barrier can be written as $\mathbb{1}_{\min\{S_{t\in[0,T]}\}>B}(S_T - K)_+$.

As a barrier is either breached or not, the sum of a knock-out option and a knock-in option is equal to a standard option. This is known as the parity relation for barrier options. Assuming that we know how to price standard options, we can focus on the pricing of one of the option types. We choose to focus on knock-outs.

Knock-out call options can be classified into four different types depending on whether the barrier is above or below the strike and on whether the barrier is an upper or lower barrier. For an upper barrier that is below the strike, the call option is worthless, which means that there are only three non-trivial types of knock-outs.

Let us start with a lower barrier that is below the strike. The barrier lies in the out-of-the-money region and has relatively little effect on the option. Under certain modeling assumptions, we compute its price $V^{\mathrm{C}}_{B,K}$ in the next section. The corresponding digital option, obtained from the K derivative of $V^{\mathrm{C}}_{B,K}$, is denoted by $\bar{V}^{\mathrm{C}}_{B,K}$. The corresponding put options are denoted by $V^{\mathrm{P}}_{B,K}$ and $\bar{V}^{\mathrm{P}}_{B,K}$ and have an upper barrier that lies above the strike. For now, we assume that these prices are given and use them to price the other two types of knock-out options.

When the lower barrier lies above the strike, we see in Fig. 2.4 that the price of the knock-out option equals

$$V^{\mathrm{C}}_{B,B} + (B - K)\bar{V}^{\mathrm{C}}_{B,B}$$

In the instance of an upper barrier above the strike, we see in Fig. 2.5 that this contract can be written as a sum of a spread put and a digital put, all with an upper

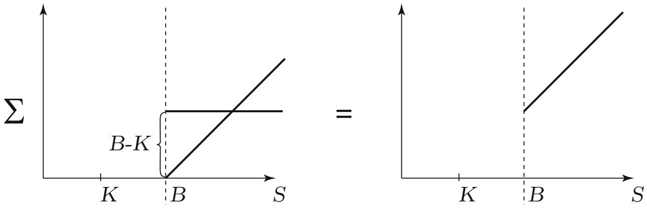

Fig. 2.4 Replication of a barrier option with strike < barrier < spot

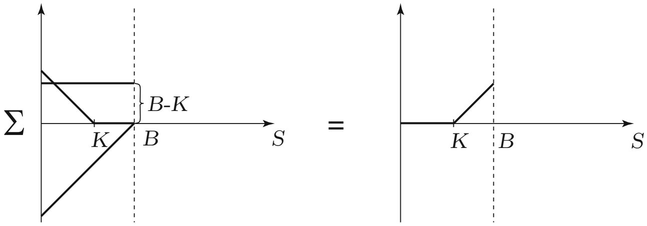

Fig. 2.5 Replication of a barrier option with spot, strike < barrier

knock-out barrier. The price is equal to

$$V^{\mathrm{P}}_{B,K} - V^{\mathrm{P}}_{B,B} + (B - K)\bar{V}^{\mathrm{P}}_{B,B}$$

2.7 Model-Dependent Pricing

The relations derived so far have been independent of the process followed by the underlying and can be viewed as model independent results. We now show how static replication can be applied in more complex situations by using modeling assumptions. We illustrate the technique by showing how barrier options can be statically replicated by European options.

We price a knock-in option that matures at T and has strike K and barrier $B < K$. The corresponding knock-out option can be priced using the parity relation for barrier options. An alternative pricing method for barrier options is presented in Sect. 9.1.

We initially assume that the underlying value S equals the barrier value which means that the option has knocked in. It is clear graphically from Fig. 2.6 that if K is large enough, there exists a put option with strike $K' < B$ that is worth as much as the call option. We temporarily assume that the call and the put are equal at all times as long as $S = B$, i.e. along the dotted line in the figure. Under this assumption the knock-in call option is worth as much as the put option. Indeed, their prices are equal if the barrier is touched while they are both worthless otherwise. Observe that the method is model dependent as a model must be used to find the strike for the put option.

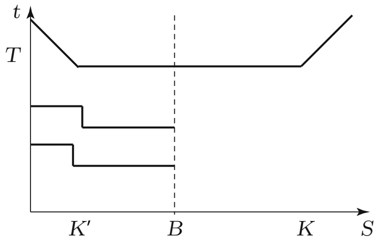

Fig. 2.6 Relication of a knock-in call by a put option and a strip of digital put options

It is possible to relax the assumption that the two options should have equal prices along the whole barrier. To prove this statement, assume first that their prices are equal along the barrier when close to maturity. The further away we come from the maturity, the more the prices start to deviate. When far enough away from the maturity, the prices will differ more than a specified tolerance level. The reason is that a different put strike should have been used for the two options to have equal values at the barrier. The incorrectness in the payoff profile used at maturity is therefore equal to the difference between two put option, with similar strikes, which can be approximated by a digital put option. The price difference at the current time can therefore approximately be adjusted by adding a digital put option. When going further back in time along the barrier the price difference will increase again, which can be periodically reset by adding more digital put options, see Fig. 2.6.

The reflection in the barrier can be done more efficiently by not using a single put options but several put options with the same strike K'. The result is a payoff profile with a steeper slope. This reflects the payoff of the call for longer time periods away from T, reducing the number of digital put options along the boundary. An example of when more than one put option is necessary is within the lognormal model that will be discussed thoroughly in the book.

The replication becomes particularly simple when the strike equals the barrier and interest rates are assumed to be zero. The price of a knock-out call option is then $S_0 - K$. The payoff can be replicated by holding this amount of cash and by entering at zero cost a forward contract to purchase the underlying for S_0 at T. If the barrier is not touched, the replicating strategy is worth $(S_T - S_0) + S_0 - K = S_T - K$ which is as much as the barrier option value. On the other hand, the forward position can be liquidated should the option be knocked out, yielding $(B - S_0) + S_0 - K = 0$ and showing that the replication is successful.

Chapter 3
Dynamic Replication

When pricing certain derivatives, the content of the replicating portfolio needs to be rebalanced through time. The derivative is then said to be dynamically replicated. For the replication to succeed, some knowledge about the propagation of the underlying is needed. As the future is unknown, the best that can be done is an educated guess on the future distributions. The various market participants base their guesses on their individual beliefs and do therefore not necessarily agree completely on the fair price. Thus, dynamic replication is model dependent in theory as well as in practice. This is in contrast to static replication that can be model dependent in practice (see Sect. 2.3) but often not in theory. For this reason, dynamic replication should only be used to price contracts for which static replication is not viable. An example of such a contract is the European call option which we discuss in detail.

We start by describing a naive dynamic replication strategy and explain why it cannot be applied in practice. Motivated by the failure of this strategy, we introduce a more sophisticated framework based on stochastic calculus. We show how stochastic calculus can be used to price fixed-time payoffs and in particular European call options. The resulting Black–Scholes formula is analyzed and the concept of implied volatility is introduced. We then view the pricing on a more abstract level and introduce the fundamental theorem of asset pricing and consider the relation between PDEs and SDEs. The chapter ends with discussions of convexity adjustments and dynamic replication of futures contracts.

3.1 Naive Replication of European Options

A naive attempt to replicate call options involves a portfolio that is empty if $S \leq KP_{tT}$ and consists of one underlying S and short K zero-coupon bonds maturing at T if $S > KP_{tT}$. When the underlying crosses the level $S = KP_{tT}$ from below, the underlying S can be bought by selling K bonds. If it crosses the level from above, the K bonds can be bought back by selling the underlying, resulting in an empty

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_3,

portfolio. The strategy is worth $(S - K)_+$ at maturity, so the replication appears to be successful.

The replicating portfolio is worth $(S - KP_{tT})_+$ at time t and, in particular, it is equal to 0 if $S < KP_{tT}$. Clearly, this cannot be true as there is a chance of S exceeding K at maturity, implying a non-zero option price. To understand the flaw in our argument, we analyze the t-dependence of S.

A study of historical time series of prices on financial products as equity stocks or FX rates shows an erratic behavior of S as a function of t. In particular, it appears that $S(t)$ does not have a well-defined (mathematical) derivative: $(S(t + \Delta t) - S(t))/\Delta t$ is Δt-dependent even for small values of Δt. Actually, this behavior makes sense intuitively because if $S(t)$ had a well-defined derivative, the relation

$$S(t + \Delta t) \approx S(t) + S'(t)\Delta t$$

would enable us to approximately predict the value of $S(t + \Delta t)$ given only information at the earlier time t. Compare with the situation where a cash sum $S(t)$ is invested in zero-coupon bonds. The value $S(t)P^{-1}_{t,t+\Delta t}$ would then have been obtained at $t + \Delta t$. Absence of arbitrage therefore implies that $S(t + \Delta t) \approx S(t)P^{-1}_{t,t+\Delta t}$. Thus, instruments S not behaving like zero-coupon bonds can only be included in our theoretical framework if we allow a t-dependence such that the derivative $S'(t)$ is ill-defined.

The fact that $S(t)$ does not have a well-defined derivative means, for example, that when S crosses the level $S = KP_{tT}$ from below, we do not succeed in buying exactly at the level, but only at a bit higher value. Similarly, we only succeed in selling the underlying for a little bit too low value when it crosses from above.

Assume that the underlying is bought at $KP_{tT} + \delta$ and sold at $KP_{tT} - \delta$. If the underlying first moves up and then down, or vice versa, a loss of 2δ is made. These losses accumulate during the lifetime of the option. Letting $\delta \to 0$ to better replicate the option does not save us from the losses because of the stochastic nature of the underlying. For example, if modeling S with a Brownian motion, only an infinitesimal amount is lost each time the level is crossed. However, it is well known that if a Brownian motion crosses a level once, it does so an infinite number of times in any open time interval containing the first crossing. The infinite number of crossings of the barrier combined with an infinitesimal loss at each crossing sums up to a finite loss. It means that our replication strategy is not possible either in practice or in theory.

The price of a call option can be written as $(S - KP_{tT})_+ + g(K)$, where $g(K)$ is a measure of the cost of buying and selling when the level KP_{tT} is hit. If K is very large or very small, the average number of hits is small and so is $g(K)$. On the other hand, for K close to $P^{-1}_{tT}S$, we expect many hits and $g(K)$ must be large. We conclude that the option price must be given by a bell-shaped positive function $g(K)$ centered somewhere around $P^{-1}_{tT}S$ and added to $(S - KP_{tT})_+$. In the following, we develop more advanced models to quantitatively determine the option price.

3.2 Dynamic Strategies

Consider a strategy with initial amount $V(t = 0)$ and such that for each time t, $\Delta(t, S)$ number of underlyings S is held and the rest of the amount is invested in zero-coupon bonds P_{tT}. To follow the evaluation through time, we first assume that the portfolio is only restructured at times $0 = T_0 < T_1 < \ldots < T_{n-1} < T_n = T$. We let the portfolio contain Δ_i number of underlyings S in $[T_i, T_{i+1})$ and use the notation P_{in} for the T_i value of the bond maturing at T_n. At T_0, the portfolio consists of Δ_0 underlyings and therefore $V(T_0) - \Delta_0 S_0$ worth of bonds, i.e. $(V(T_0) - \Delta_0 S_0)P_{0n}^{-1}$ number of bonds:

$$\begin{aligned} V(T_0) &= \Delta_0 S_0 + \left((V(T_0) - \Delta_0 S_0)P_{0n}^{-1}\right) P_{0n} \\ &= \Delta_0 S_0 + \left((V(T_0)P_{0n}^{-1} - \Delta_0 F_0)\right) P_{0n} \end{aligned}$$

The value of the portfolio at times T_1, T_2 and T_k is equal to

$$\begin{aligned} V(T_1) &= \Delta_0 S_1 + \left((V(T_0)P_{0n}^{-1} - \Delta_0 F_0)\right) P_{1n} \\ &= \Delta_1 S_1 + \left((\Delta_0 - \Delta_1)F_1 + (V(T_0)P_{0n}^{-1} - \Delta_0 F_0)\right) P_{1n} \\ V(T_2) &= \Delta_1 S_2 + \left((\Delta_0 - \Delta_1)F_1 + (V(T_0)P_{0n}^{-1} - \Delta_0 F_0)\right) P_{2n} \\ &= \Delta_2 S_2 + ((\Delta_1 - \Delta_2)F_2 + (\Delta_0 - \Delta_1)F_1 \\ &\quad +(V(T_0)P_{0n}^{-1} - \Delta_0 F_0)\big) P_{2n} \\ V(T_k) &= (\Delta_k F_k + (\Delta_{k-1} - \Delta_k)F_k + (\Delta_{k-2} - \Delta_{k-1})F_{k-1} + \ldots \\ &\quad +(\Delta_0 - \Delta_1)F_1 + (V(T_0)P_{0n}^{-1} - \Delta_0 F_0)\big) P_{kn} \\ &= (\Delta_{k-1}(F_k - F_{k-1}) + \Delta_{k-2}(F_{k-1} - F_{k-2}) + \ldots \\ &\quad +\Delta_0(F_1 - F_0) + V(T_0)P_{0n}^{-1}\big) P_{kn} \\ \Leftrightarrow V(T_k)/P_{kn} &= \sum_{i=0}^{k-1} \Delta_i (F_{i+1} - F_i) + V(T_0)/P_{0n} \end{aligned}$$

In the continuous-time limit, we obtain

$$V(t)/P_{tT} = \int_0^t \Delta dF + V(t = 0)/P_{0T}$$

which gives the following value for $t = T$:

$$V(T) = \int_0^T \Delta dF + V(t = 0)/P_{0T}$$

The above equations indicate that it is more natural to work with the quotient $U(t) = V(t)/P_{tT}$ than with $V(t)$. Thus, instead of quoting portfolio prices in dollar terms, we quote them relative to a tradable asset, in this case the bond maturing at T. The asset, in terms of which the prices are quoted, is called the *numeraire*. U is called the forward value of the contract V just as $F_t = S_t/P_{tT}$ is the forward value of S. In terms of the forward values, the value of U at T can be obtained from the trading strategy $(\Delta, U(t = 0))$ according to

$$U(T) = \int_0^T \Delta dF + U(t = 0)$$

The portfolio value generally depends on the path followed by F. We now restrict ourselves to strategies that give path-independent values of the portfolio, i.e. $U(t)$ depends only on F_t and t, but not on the values assumed by F before t. To analyze such strategies, it is necessary to postulate the propagation of the forward F. We make the assumption that F satisfies a *stochastic differential equation (SDE)*:

$$dF_t = \mu(t, F_t)dt + \sigma(t, F_t)dW_t$$

It means that during a small time step dt, the value of F_t changes by $\mu(t, F_t)dt$ plus a part that is proportional to a change dW_t of a Brownian motion. The term $\mu(t, F_t)$ is called the *drift* while $\sigma(t, F_t)$ is called the *volatility*. This chapter only considers continuous processes as the SDE above; the generalization to non-continuous processes is the subject of Chap. 8.

The products of stochastic differentials are special because they are not zero as for ordinary differentials. For example,

$$E\left[(dW_t)^2\right] = E\left[(W_{t+dt} - W_t)^2\right] = dt$$

holds since $W_{t_2} - W_{t_1}$ is normally distributed with mean 0 and variance $t_2 - t_1$. The products $(dt)^2$ and $dW_t dt$, and the variance of $(dW_t)^2$ are all of higher orders in dt. This suggests the use of the following *product rule for stochastic differentials*: $dW_t dW_t = dt$, and all other differential products equal to zero. A rigorous derivation of the product rule can be found in the Appendix.

Combining the product rule for stochastic differentials with Taylor expansion gives the following change in a path-independent U during an infinitesimal time step dt:

$$dU(t, F_t) = U_t dt + U_F dF_t + \frac{1}{2}\sigma(t, F_t)^2 U_{FF} dt$$

This chain rule of stochastic differentiation is called *Ito's lemma*. Observe that the subindices on U denote partial derivatives while for F they denote a dependence on t. Inserting the previous result $dU = \Delta dF$ in the above equation, we conclude that a path-independent strategy has to satisfy

$$\begin{cases} \Delta = U_F \\ U_t + \frac{1}{2}\sigma(t, F_t)^2 U_{FF} = 0 \end{cases}$$

The (parabolic) *partial differential equation (PDE)* in the second line is of fundamental importance for derivatives pricing and will be used repeatedly in the book.

For an example of a strategy with a path-independent portfolio value, assume that the volatility has the form $\sigma(t, F_t) = \sigma F_t$ and consider $U = F_t^\lambda g(t)$. $g(t)$ can be solved by inserting this expression into the PDE:

$$g'(t) + \frac{1}{2}\sigma^2\lambda(\lambda - 1)g(t) = 0 \Leftrightarrow g(t) = g(0)\exp\left(-\frac{1}{2}\sigma^2\lambda(\lambda - 1)t\right)$$

We then obtain

$$U(t) = U(0)\frac{F_t^\lambda}{F_0^\lambda}\exp\left(-\frac{1}{2}\sigma^2\lambda(\lambda - 1)t\right)$$

and

$$\Delta = U_F = \lambda\frac{U(t)}{F_t} \Rightarrow \frac{dU}{U} = \lambda\frac{dF}{F}$$

We see that the daily percentage increase of U is λ times the increase in F. At each point in time, the amount $\Delta F_t = \lambda U(t)$ is held in the underlying, i.e. a multiple λ of the total portfolio amount. These *leveraged strategies* are particularly popular for λ equal to –3, –2, –1, 2 and 3.

To illustrate that the leverage strategies imply path-independent portfolio values for the model with $\sigma(t, F_t) = \sigma(t)F_t$, we simulate the underlying process over a time period of one year. We use 500 paths, 100 time steps and assume that $\lambda = 2$, $\sigma(t) = 20\%$, $\mu(t, F_t) = F_t \cdot 5\%$ and that interest rates are zero. We assume that the initial values of the underlying and the portfolio is 1. Figure 3.1 displays the 1Y value of the portfolio for various end values of the simulated underlying paths. Apart from some numerical noise, the portfolio value only depends on the end value of the simulation and not on the path that was taken. The figure also shows the strategy for

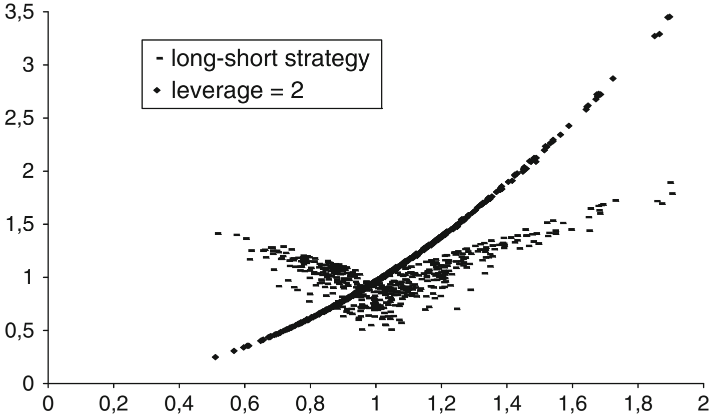

Fig. 3.1 Comparing strategies with path-independent and path-dependent portfolio values

which $\Delta = 1$ if the underlying value is greater than 1, and $\Delta = -1$ if the underlying value is less than 1. This strategy does not satisfy the above PDE and we therefore expect the portfolio value to be path dependent. Indeed, it is clear from the figure that the portfolio value cannot be written as a function of only the underlying value.

Observe that the drift part $\mu(t, F_t)$ does not enter the PDE. It means that the question of whether a strategy leads to a path-independent portfolio value only depends on the model via the volatility and not on the form of the drift. This important fact is discussed in more detail in Sect. 3.8.

3.3 Replication of Fixed-Time Payoffs

We saw in the previous section how a future portfolio value $U(T)$ can be computed from a given initial value $U(t = 0)$ and strategy Δ. We showed that if the strategy is chosen so that U satisfies a certain PDE, then $U(t)$ depends only on t and F_t, and not on the path followed by F. We now attack the reverse problem: how to construct the trading strategy $(\Delta, U(t = 0))$ that reproduces a given future payoff $U(T) = h(F_T) = h(S_T)$.

As $U(T) = h(F_T)$, the strategy is path independent and the PDE of the previous section must be fulfilled. We can therefore use $U(T) = h(F_T)$ as the final condition for the PDE. As the PDE is of first order in time, this is sufficient information for computing $U(t, F_t)$ for all $t \in [0, T]$. Setting $t = 0$ gives $U(t = 0)$ and taking the F_t derivative gives Δ.

The findings in this section can be summarized as:

If F satisfies the SDE

$$dF_t = \mu(t, F_t)dt + \sigma(t, F_t)dW_t$$

then a payoff $V(T) = h(F_T)$ can be attained by solving the PDE

$$\begin{cases} U_t + \frac{1}{2}\sigma(t, F_t)^2 U_{FF} = 0 \\ U(t = T, F) \qquad = h(F) \end{cases}$$

and using the strategy $V(t = 0) = P_{0T}U(t = 0)$ and $\Delta = U_F = \frac{\partial U}{\partial F}$.

3.4 The Black–Scholes Formula

We assume that any scaling λF, $\lambda > 0$, satisfies the same SDE as the forward itself:

$$d(\lambda F_t) = \mu(t, \lambda F_t)dt + \sigma(t, \lambda F_t)dW_t$$

Together with $d(\lambda F) = \lambda dF$, this scale invariance implies that $\sigma(t, \lambda F) = \lambda\sigma(t, F)$ or $\sigma(t, F) = \sigma(t)F$ for some F-independent function $\sigma(t)$, and the corresponding result for the drift.

Should the forward value change from F to λF, the scale invariance states that a basket of $1/\lambda$ underlyings propagates as the original forward value. This is not, however, how financial assets behave. For instance, a substantial fall in the price of an equity stock is often a sign of a weakness in the issuing company, which makes investors nervous and leads to a higher trading activity and a higher volatility. This behavior can be expressed mathematically as

$$\sigma(t, \lambda F) > \lambda\sigma(t, F)$$

for $0 < \lambda < 1$. In Chap. 5, we analyze SDEs that satisfy this, and related, inequalities. For now, we restrict ourselves to the corresponding equality as a first non-trivial attempt of derivatives pricing. We then have

$$dF \sim \sigma F dW_t$$

and F is then said to follow a *geometric Brownian motion* or a *lognormal process*. We have chosen to omit the drift as it does not impact the pricing.

We assume, for simplicity, that the *lognormal volatility* σ is independent of t. As we show in later chapters, the time-dependent generalization is straightforward. We need to solve the following problem for European call options:

$$\begin{cases} U_t + \frac{1}{2}\sigma^2 F^2 U_{FF} = 0 \\ U(t = T, F) \qquad = (F - K)_+ \end{cases}$$

The PDE is called the *Black–Scholes equation* and was originally derived in Black and Scholes (1973). Using the transformation

$$U(t, F) = K\Psi(\tau, x)$$

with

$$\begin{cases} \tau = \sigma^2(T - t) \\ x = F/K \end{cases}$$

leads to the dimensionless problem

$$\begin{cases} \Psi_\tau - \frac{1}{2}x^2\Psi_{xx} = 0 \\ \Psi(\tau = 0, x) \quad = (x - 1)_+ \end{cases}$$

A transformation to a PDE with constant coefficients is possible by using $\Omega(\tau, z) = \Psi(\tau, x)$, for $z = \ln x$:

$$\begin{cases} \Omega_\tau + \frac{1}{2}\Omega_z - \frac{1}{2}\Omega_{zz} = 0 \\ \Omega(\tau = 0, z) \qquad\quad = (e^z - 1)_+ \end{cases}$$

Through the substitution

$$\Omega = e^{z/2-\tau/8}\Phi$$

we obtain the heat equation:

$$\begin{cases} \Phi_\tau - \frac{1}{2}\Phi_{zz} \ = 0 \\ \Phi(\tau = 0, z) = \left(e^{z/2} - e^{-z/2}\right)_+ \end{cases}$$

The above problem can be solved by first focusing on the corresponding equations for the Green's function:

$$\begin{cases} p_\tau - \frac{1}{2}p_{zz} \qquad = 0 \\ p(\tau = 0, z, z') = \delta(z - z') \end{cases}$$

It is then clear that

$$\Phi(\tau, z) = \int_{-\infty}^{\infty} p(\tau, z, z') \left(e^{z'/2} - e^{-z'/2}\right)_+ dz'$$

satisfies the correct PDE and initial condition. The Green's function problem can be solved by using the Fourier transform of p:

$$\begin{cases} p(\tau, z, z') \ = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{p}(\tau, k, z')e^{-ikz}dk \\ \hat{p}(\tau, k, z') = \int_{-\infty}^{\infty} p(\tau, z, z')e^{ikz}dz \end{cases}$$

As $\partial_z^2 e^{-ikz} = -k^2 e^{-ikz}$, the transformed PDE reads

$$\hat{p}_\tau = -\frac{1}{2}k^2\hat{p} \Leftrightarrow \hat{p}(\tau, k, z') = e^{-k^2\tau/2}\hat{p}(\tau = 0, k, z')$$

with initial condition

$$\hat{p}(\tau = 0, k, z') = \int_{-\infty}^{\infty} \delta(z - z')e^{ikz}dz = e^{ikz'}$$

It gives us

$$p(\tau, z, z') = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-k^2\tau/2}e^{-ik(z-z')}dk = \frac{1}{\sqrt{2\pi\tau}}e^{-(z-z')^2/2\tau}$$

where the last equality can be verified by using calculus of residues or through tables of Fourier transforms, see Gradshteyn and Rhyzik (2007).

Collecting the results, we obtain

$$\begin{aligned}\Phi(\tau,z) &= \frac{1}{\sqrt{2\pi\tau}}\int_0^\infty e^{-(z-z')^2/2\tau}\left(e^{z'/2}-e^{-z'/2}\right)dz' \\ &= \frac{1}{\sqrt{2\pi\tau}}e^{-z^2/2\tau}\int_0^\infty \Big(e^{-(z'-(z+\tau/2))^2/2\tau}e^{(z+\tau/2)^2/2\tau} \\ &\quad -e^{-(z'-(z-\tau/2))^2/2\tau}e^{(z-\tau/2)^2/2\tau}\Big)dz' \\ &= \frac{1}{\sqrt{2\pi}}e^{-z^2/2\tau}\Big(e^{(z+\tau/2)^2/2\tau}\int_{-(z+\tau/2)/\sqrt{\tau}}^\infty e^{-z'^2/2}dz' \\ &\quad -e^{(z-\tau/2)^2/2\tau}\int_{-(z-\tau/2)/\sqrt{\tau}}^\infty e^{-z'^2/2}dz'\Big) \\ &= e^{\tau/8}\left(e^{z/2}N\left(\frac{z}{\sqrt{\tau}}+\frac{1}{2}\sqrt{\tau}\right)-e^{-z/2}N\left(\frac{z}{\sqrt{\tau}}-\frac{1}{2}\sqrt{\tau}\right)\right)\end{aligned}$$

where the *cumulative normal function* is defined by

$$N(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-y^2/2}dy$$

We finally arrive at

$$\begin{aligned}\Omega(\tau,z) &= e^{z/2-\tau/8}\Phi(\tau,z)=e^zN\left(\frac{z}{\sqrt{\tau}}+\frac{1}{2}\sqrt{\tau}\right)-N\left(\frac{z}{\sqrt{\tau}}-\frac{1}{2}\sqrt{\tau}\right) \\ \Rightarrow U(t,F) &= K\Omega(\sigma^2(T-t),\ln(F/K)) \\ &= FN(d_+)-KN(d_-)\,,\quad d_\pm=\frac{\ln(F/K)}{\sigma\sqrt{T-t}}\pm\frac{1}{2}\sigma\sqrt{T-t}\end{aligned}$$

which is the celebrated *Black–Scholes formula*. With spot values, the formula takes the form

$$V(t,S)=SN(d_+)-P_{tT}KN(d_-)\,,\quad d_\pm=\frac{\ln(S/(P_{tT}K))}{\sigma\sqrt{T-t}}\pm\frac{1}{2}\sigma\sqrt{T-t}$$

In the same way, with the terminal condition $U(t=T,F)=(K-F)_+$, we obtain the Black–Scholes formula for put options:

$$\begin{aligned}U(t,F) &= KN(-d_-)-FN(-d_+) \\ V(t,S) &= P_{tT}KN(-d_-)-SN(-d_+)\end{aligned}$$

3.5 Analysis of the Black–Scholes Formula

To analyze the Black–Scholes formula, we first note that, due to put-call parity, it is possible to only consider call options. We use $\tilde{U}(\omega, x) = U(F, K, t, T, \sigma)/K$ with

$$\begin{cases} \omega = \sigma\sqrt{T-t} \\ x = F/K \end{cases}$$

to obtain a dimensionless formula:

$$\tilde{U}(\omega, x) = xN\left(d_+\right) - N\left(d_-\right), \quad d_\pm = \frac{\ln x}{\omega} \pm \frac{1}{2}\omega$$

where $x, \omega \in (0, \infty)$. This expression is easy to analyze as it depends on only two variables: ω and x, instead of the traditional six: S, K, t, T, σ and P_{tT}. The dimensional reduction is possible as the option price has a similar dependence on many of its variables. For instance, an increasing maturity has the same effect as an increasing volatility.

$\tilde{U}$ has the asymptotic limits

$$\begin{aligned} \tilde{U}(x \to 0) &\to 0 \\ \tilde{U}(x \to \infty) &\to x - 1 \\ \tilde{U}(\omega \to 0) &\to (x-1)_+ \\ \tilde{U}(\omega \to \infty) &\to x \end{aligned}$$

and by using the relations

$$\begin{aligned} \frac{dN}{dx} &= n(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ \frac{dn}{dx} &= -xn(x) \\ \frac{dd_\pm}{dx} &= 1/x\omega \\ \frac{dd_\pm}{d\omega} &= -d_\mp/\omega \\ n(d_-) &= xn(d_+) \end{aligned}$$

we obtain the lowest-order partial derivatives:

$$\tilde{U}_x = N(d_+) + xn(d_+)\frac{dd_+}{dx} - n(d_-)\frac{dd_-}{dx} = N(d_+) > 0$$

$$\begin{aligned}\tilde{U}_\omega &= xn(d_+)\frac{dd_+}{d\omega} - n(d_-)\frac{dd_-}{d\omega} = n(d_-) > 0\\ \tilde{U}_{xx} &= n(d_+)\frac{dd_+}{dx} = \frac{1}{x\omega}n(d_+) > 0\\ \tilde{U}_{x\omega} &= n(d_+)\frac{dd_+}{d\omega} = -\frac{d_-}{\omega}n(d_+)\\ \tilde{U}_{\omega\omega} &= -d_-n(d_-)\frac{dd_-}{d\omega} = \frac{d_+d_-}{\omega}n(d_-)\end{aligned}$$

From these expressions, we conclude that the Black–Scholes formula has the correct asymptotics and satisfies the necessary constraints derived in Sect. 2.4. The only partial derivatives of order 2 or less that have zeros are

$$\begin{aligned}\tilde{U}_{x\omega} &= 0 \Leftrightarrow d_- = 0 \Leftrightarrow x = e^{\omega^2/2}\\ \tilde{U}_{\omega\omega} &= 0 \Leftrightarrow d_\pm = 0 \Leftrightarrow x = e^{\pm\omega^2/2}\end{aligned}$$

It follows, in particular, that

$$\max_\omega \tilde{U}_\omega(\omega, x) = \frac{1}{\sqrt{2\pi}}\max(x, 1)$$

To visualize the Black–Scholes formula, consider the graphs in Fig. 3.2 where one of the parameters is fixed while the other is varied. Observe that for increasing ω, the price grows from the intrinsic payoff $(x-1)_+$ to the value of the forward x. Thus, an increasing volatility, or an increasing time to maturity, leads to a higher option value. An increasing x gives a call option price that grows from zero to the value of the forward x. This means that the call option price increases with higher values of the underlying while it decreases with the strike.

A typical volatility of a financial asset is in the order of magnitude of 20% but can be as low as a few percent or as high as 100%. The maturity of an option is typically from a few weeks up to 20 years or so, but is usually around a couple of years or less. Thus, a typical value of ω is in the order of magnitude of 0.2. As we

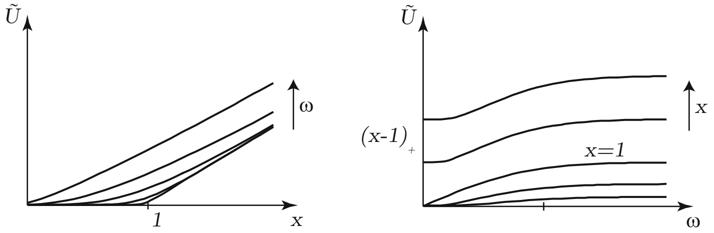

Fig. 3.2 Dependence of the normalized Black–Scholes formula on its variables

see examples of later in the book, such a small value opens up for the possibility to make a perturbative expansion of the option price in terms of ω.

3.6 Implied Volatility

The Black–Scholes formula depends on the contract specific parameters K and T, today's date t, the current value of the underlying S, the volatility σ and the discount factor to maturity. The numerical values of these variables can be found in the contract specification and from the market data. The exception is σ, which can be estimated from, for example, historical data together with a certain belief of the future up to the maturity of the contract. As σ is the only variable that is not directly observable, the Black–Scholes formula can be viewed as a transformation between the volatility and the price. The transformation is invertible as the price increases with the volatility. Thus, given a call option price, the volatility can be determined. Unfortunately, no closed-form formula exists and an approximate function or a root-finding routine is necessary.

Using the one-to-one correspondence between prices and volatilities, options can be quoted in terms of volatilities. Indeed, this is the market convention and the value σ_{imp} of the volatility that gives the market price is called the *implied volatility*. The market convention is to use the actual/365 day-count convention, see Sect. 13.1. In order to differentiate between the implied volatility and the volatility appearing in the SDE, the latter is usually referred to as the *local volatility*. When it is clear from the context which of the two types of volatilities is meant, we refer to them simply as volatilities.

For a more detailed discussion of implied volatilities, consider two call options with different strikes, but with the same maturity, for which we back out the implied volatilities from market prices. These volatilities should be identical according to the Black–Scholes model, but in reality we find that they are different. Indeed, the Black–Scholes model is nothing but a model and the real world does not necessarily behave accordingly. It means that the implied volatility is a function of the strike. The same statement holds when the maturity is varied, which gives $\sigma_{\mathrm{imp}} = \sigma_{\mathrm{imp}}(T, K)$.

At first, it seems as we have gained nothing from the Black–Scholes model: we started with prices $V(T, K)$ that depend on the strike and maturity and we ended up with implied volatilities $\sigma_{\mathrm{imp}}(T, K)$ that depend on the same variables. The usefulness of the Black–Scholes model can be seen by assuming a flat volatility surface: $\sigma_{\mathrm{imp}}(T, K) = \sigma_{\mathrm{imp}}$. We then obtain reasonable option prices with correct asymptotics and without violating the no-arbitrage conditions. If, on the other hand, we had assumed that $V(T, K)$ is independent of the strike and maturity, non-sense results would have been obtained. Furthermore, the price surface $V(T, K)$ has a complex shape while the volatility surface $\sigma_{\mathrm{imp}}(T, K)$ is much flatter. It is therefore often simpler to base more advanced option models on the volatility instead of directly in terms of the price. The Black–Scholes model is also useful because it gives us information about how to risk manage options, see Sect. 4.3.

The fact that the implied volatility is not constant means that it contains market information that is incorrectly modeled or overlooked by the Black–Scholes model. For instance, equity stocks are typically observed to become relatively more volatile after a fall in the value. This increase in volatility has the consequence that call options with low strikes must be priced with a higher volatility than call options with high strikes, i.e. the implied volatility curve for equity options decreases with increasing strikes. An implied volatility curve is said to be *skewed* if it tilts in a region containing the ATM point.

A similar scenario occurs in the FX market: if an FX rate drops substantially, the volatility typically increases. If, on the other hand, the FX rate has a sharp upturn, the inverse FX rate drops, which should again lead to an increase in the volatility. Observe that this kind of symmetry argument does not exist for an equity stock as we expect the currency buying the equity stock to be much more stable than the stock itself. We conclude that implied volatility curves for FX often have a minimum close to ATM. Implied volatility curves with this shape are said to have a *smile*.

Please note that we have only given single reasons why implied volatilities have a skew for equity options and a smile for FX options. The reality is much more complex and there are several explanations for the non-flatness of the implied volatility surface, including the fact that hedging is only done discretely in time, a non-zero correlation between the underlying and the volatility, transaction costs from bid-offer spreads, supply and demand considerations, fat tail probability distributions, etc. There even exist situations for which equity options have implied volatilities smiles and FX options have implied volatilities skews.

Because of put-call parity, adding $KP_{0T} - S$ to the call option price $V(\sigma_{\text{imp}}(T, K))$ gives the put option price with the same strike and maturity. As the Black–Scholes formula fulfills put-call parity when the put volatility equals the call volatility, $\sigma_{\text{imp}}(T, K)$ is the unique volatility in the Black–Scholes formula that gives the correct price for put options. We therefore conclude that the value of the implied volatility $\sigma_{\text{imp}}(T, K)$ is independent of the option type.

Since the Black–Scholes formula can be interpreted as a transformation between the price and the implied volatility, the latter depends on the current time t and the forward F (or the spot S) as well as on K and T. As the option price is very sensitive to changes in F, it is important to model the implied volatility surface to reflect the market behavior when F changes.

A particularly simple class of models, the *sticky-strike models*, assumes that the implied volatility surface is independent of changes in F. The fact that this model type is in disagreement with typical market behavior can be understood from the example of an implied volatility smile with the minimum located at the ATM point. A changing forward then means that the minimum moves away from the ATM point.

An alternative class of models, the *sticky-delta models*, assumes that the implied volatility depends on K and F only through the combination K/F: $\sigma_{\text{imp}}(T, K; t, F) = \sigma_{\text{imp}}(T, K/F, t)$. To understand the behavior of these models when the forward changes, assume that today's volatility curve is given by $g(K)$ for the maturity T. The implied volatility curve that matches these prices is given by $\sigma_{\text{imp}}(T, z, t) = g(Fz)$. After this calibration of the curve, let the time go and revisit the model at a later date, when the forward value equals $\tilde{F}$. A sticky delta model then predicts the volatility curve

$$\sigma_{\text{imp}}(T, K; t, \tilde{F}) = \sigma_{\text{imp}}(T, K/\tilde{F}, t) = g\left(\frac{F}{\tilde{F}}K\right)$$

from which we conclude that the implied volatility curve slides along with the change in the forward.

In reality, the *dynamics* (i.e. the dependence on F) of the implied volatility curve are often somewhere between that of sticky strike and sticky delta. The dynamics can be estimated by analyzing historical data on option prices.

3.7 Relations between PDEs and SDEs

In Sect. 3.2, we started with an SDE for the forward and derived a PDE for the derivatives price. This indicates a close connection between SDEs and PDEs, which we now analyze in more detail.

Let X be a process that satisfies the SDE

$$dX = \mu(t, X)dt + \sigma(t, X)dW_t$$

and denote by $p(T, \chi; t, x)d\chi$ the probability that X_T is in a small interval $[\chi, \chi + d\chi]$ conditional on $X_t = x$. p is called the *Green's function* of the process and was previously encountered in Sect. 3.4. p satisfies the *Chapman-Kolmogorov equation*

$$p(T, \chi; t, x) = \int p(T, \chi; t', x')p(t', x'; t, x)dx'$$

for any t' between t and T. For an arbitrary function $h(x)$, set

$$g(t, x) = E[h(X_T)] = \int h(x_T)p(T, x_T; t, x)dx_T$$

Ito's lemma gives

$$dg(t, X(t)) = (g_t + \mu g_x + \frac{1}{2}\sigma^2 g_{xx})dt + \sigma g_x dW_t$$

As g is the expectation of a function, it cannot have any drift, i.e. the dt term must be zero:

$$g_t + \mu g_x + \frac{1}{2}\sigma^2 g_{xx} = 0$$

h being arbitrary implies that

$$p_t + \mu p_x + \frac{1}{2}\sigma^2 p_{xx} = 0$$

From this PDE in the backward coordinates (t, x), the Chapman-Kolmogorov equation can be used to derive a corresponding PDE in the forward coordinates (T, χ):

$$\begin{aligned}
0 = \frac{\partial}{\partial t} p(T, \chi; t', x') &= \int \frac{\partial}{\partial t} \big(p(T, \chi; t, x) p(t, x; t', x')\big)\, dx \\
&= \int \Bigg(p(t, x; t', x') \left(-\mu \frac{\partial}{\partial x} - \frac{1}{2}\sigma^2 \frac{\partial^2}{\partial x^2} \right) p(T, \chi; t, x) \\
&\quad + p(T, \chi; t, x) \frac{\partial}{\partial t} p(t, x; t', x') \Bigg)\, dx \\
&= \int p(T, \chi; t, x) \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\mu - \frac{1}{2}\frac{\partial^2}{\partial x^2}\sigma^2 \right) p(t, x; t', x') dx
\end{aligned}$$

where the derivatives on the right-hand side acts on everything to the right of them. Let us summarize the results obtained so far:

Let X satisfy the SDE

$$dX = \mu(t, X)dt + \sigma(t, X)dW_t$$

and let $p(T, \chi; t, x)d\chi$ denote the probability that X_T is in a small interval $[\chi, \chi + d\chi]$ conditional on $X_t = x$. Then the following relations must be satisfied:

- *Chapman-Kolmogorov Equation:*

$$p(T, \chi; t, x) = \int p(T, \chi; t', x') p(t', x'; t, x) dx', \quad t' \in (t, T)$$

- *Backward Kolmogorov Equation:*

$$p_t + \mu p_x + \frac{1}{2}\sigma^2 p_{xx} = 0$$

- *Forward Kolmogorov (or Fokker-Planck) Equation:*

$$p_T + \frac{\partial}{\partial \chi}\mu p - \frac{1}{2}\frac{\partial^2}{\partial \chi^2}\sigma^2 p = 0$$

We conclude that the following two problems have the same solution:

1. $\begin{cases} g_t + \mu(t,x)g_x + \dfrac{1}{2}\sigma(t,x)^2 g_{xx} &= 0 \\ g(t=T,x) &= h(x) \end{cases}$
2. $g(t,x) = E\left[h(X(T))|dX = \mu(t',X(t'))dt' + \sigma(t',X(t'))dW_{t'}, X(t) = x\right]$

In the derivation of this statement, $g(t,X(t))$ could have been replaced with $g(t,X(t))\exp(-\int_0^t r(u,X(u))du)$ and $h(x)$ with $h(x)\exp(-\int_0^T r(u,X(u))du)$. The result is a well-known theorem:

Feynman-Kac Theorem: The two problems below have the same solution.

1. $\begin{cases} g_t + \mu(t,x)g_x + \dfrac{1}{2}\sigma(t,x)^2 g_{xx} - r(t,x)g &= 0 \\ g(t=T,x) &= h(x) \end{cases}$
2. $g(t,x) = E\Big[h(X(T))e^{-\int_t^T r(u,X(u))du}|$
   $dX = \mu(t',X(t'))dt' + \sigma(t',X(t'))dW_{t'}, X(t) = x\Big]$

3.8 The Fundamental Theorem of Asset Pricing

In Sect. 3.2, the process

$$dF_t = \mu(t,F_t)dt + \sigma(t,F_t)dW_t$$

lead us to the derivatives-pricing problem

$$\begin{cases} U_t + \dfrac{1}{2}\sigma(t,F_t)^2 U_{FF} &= 0 \\ U(t=T,F) &= h(F) \end{cases}$$

According to the Feynman-Kac theorem, this problem can alternatively be solved by

$$U = E[h(F)]$$

where the expectation is taken under the SDE

$$dF_t = \sigma(t,F_t)dW_t$$

The original SDE for F has thereby been turned into one that lacks drift.

We now consider measures Q that assign different probabilities to events than the real-world measure P. It is well known that the drift, but not the volatility, of an SDE is affected when changing measure from P to an equivalent measure Q, see the Appendix. In fact, for any given SDE, there exists a measure that cancels the drift.

As derivatives pricing can be done by eliminating the drift of the SDE for the forward, the above discussion indicates a connection to measure changes. Indeed, the pricing of derivatives can be interpreted as a transformation from the real-world measure P to the measure Q in which the drift vanishes. The forward of the derivatives price can then be computed as the expectation of its value at maturity. As the underlying forward is driftless, it is also equal to the expectation of its future value: $F_t = E_t^Q[F_{t'}]$, $t' > t$. Processes with the property that they are equal to the expectation of their future values therefore play a fundamental role in derivatives pricing. This type of processes are called *martingales*. We conclude that the tradables of concern to us, i.e. S and V, are martingales in the Q measure when quoted relative to the numeraire P_{tT}.

The above discussion can be generalized to encompass several underlyings and an arbitrary numeraire: for a given set of tradables $\{S^i\}$ and a tradable N, the numeraire, there exists a probability measure Q under which $\{S^i/N\}$ are martingales. It can be proven that the existence of Q is a consequence of the absence of arbitrage in the market. The converse is also true: if there exists a measure under which $\{S^i/N\}$ are martingales, then the market must be arbitrage free. The equivalence between absence of arbitrage and the existence of a martingale probability measure is called the *fundamental theorem of asset pricing*. If limiting ourselves to measures only associated with the stochastic information in $\{S^i\}$ and N, it can be proven that Q is unique.

To explain the intuition behind the fundamental theorem of asset pricing, assume that time is discrete with a single time step from 0 to T and that the market only consists of two assets, S^1 and S^2, where we let the latter be the numeraire, $N = S^2$. The fact that the market is arbitrage free implies that there exists an event with non-zero probability in the real-world measure P such that $S^1/N > 1$ at T. Similarly, there exists a non-zero probability event with $S^1/N < 1$ at T. By assigning different probability weights to the events it is clearly possible to construct a probability measure Q under which S^1/N is a martingale.

For the reverse statement, assume that S^1/N is a martingale in a measure Q equivalent to the real-world measure P. Unless S_1 equals N with probability one, there must exist an event with non-zero probability in P such that $S^1/N > 1$, or equivalently such that $S_1 > S_2$. Similarly, there must exist an event with non-zero probability such that $S_1 < S_2$. We conclude that the market must be arbitrage free. The generalization of the proof of the fundamental theorem of asset pricing to continuous time and several assets is more mathematically challenging but the principle remains the same.

The bond P_{tT} is the most common choice of numeraire when pricing a derivative that has a payment at a single point T in time. The reason for this choice is that the value of the numeraire becomes particularly simple at $t = T$: $P_{TT} = 1$. The corresponding martingale probability measure is called the *forward measure*. Another popular numeraire is the *money market account*, defined as a continuous reinvestment into the short rate. It means that \$1 invested at $t = 0$ has the time T value

$$(1 + r_0\delta t)(1 + r_{\delta t}\delta t)(1 + r_{2\delta t}\delta t)\ldots \to e^{\int_0^T r_u du}$$

The corresponding martingale measure is called the *risk-neutral measure*. This measure is identical to the forward measure when interest rates are deterministic.

The fundamental theorem of asset pricing provides us with a bit faster route to price derivatives compared to what was done in the Sect. 3.2. By working in the martingale measure Q, F has only a diffusion part and we never have to be concerned with the drift. We do not have to set up a hedging portfolio to do the pricing as we know that U is a martingale and the price is equal to the expectation of the time T value. Furthermore, if the volatility needs to be estimated from, for instance, a historical time-series analysis, then this is possible as only the drift is affected by measure changes.

3.9 Expectation of Non-Linear Payoffs

We have seen that derivatives prices can be calculated through expectations. To prepare for calculations later in the book, we collect some useful approximative results regarding the computation of expectations.

The expectation of $E[h(F_T)]$ is trivial if $h(F) = 1$. As the forward F is a martingale, $F_t = E_t[F_T]$, the computation is also trivial if $h(F) = F$. It remains to compute the expectation for non-linear functions h. The fact that the computation is simple for linear functions suggests that we should make use of Taylor expansion:

$$\begin{aligned} E[h(F_T)] &\approx E\left[h(\tilde{F}) + h'(\tilde{F})(F_T - \tilde{F}) + \frac{1}{2}h''(\tilde{F})(F_T - \tilde{F})^2\right] \\ &= h(\tilde{F}) + h'(\tilde{F})(F_t - \tilde{F}) + \frac{1}{2}h''(\tilde{F})E\left[(F_T - \tilde{F})^2\right] \end{aligned}$$

With $\tilde{F} = F_t$, the equation becomes

$$E[h(F_T)] \approx h(F_t) + \frac{1}{2}h''(F_t)E\left[(F_T - F_t)^2\right]$$

Using

$$\mathrm{Var}(F_T) = \mathrm{Var}(F_T - F_t) = E\left[(F_T - F_t)^2\right] - E\left[F_T - F_t\right]^2 = E\left[(F_T - F_t)^2\right]$$

we arrive at

$$E[h(F_T)] \approx h(F_t) + \frac{1}{2}h''(F_t)\mathrm{Var}(F_T)$$

As h is often a convex function, the second term on the right-hand side gives the lowest-order contribution coming from the convexity. This is called the (lowest-order) *convexity adjustment* and we observe that it is proportional to the variance of the underlying. Higher-order contributions to the convexity adjustment can be added by using more terms in the Taylor expansion.

The situation is sometimes the reverse when $E[h(F_T)]$ is known and $E[F_T]$ is sought. This can be handled by using $\tilde{F} = \bar{F}_t$ with

$$h(\bar{F}_t) = E[h(F_T)]$$

leading to

$$E[F_T] \approx \bar{F}_t - \frac{1}{2}\frac{h''(\bar{F}_t)}{h'(\bar{F}_t)}E\left[(F_T - \bar{F}_t)^2\right]$$

Consider then

$$\begin{aligned}\mathrm{Var}(F_T) &= \mathrm{Var}(F_T - \bar{F}_t) = E\left[(F_T - \bar{F}_t)^2\right] - E\left[F_T - \bar{F}_t\right]^2 \\ &\approx E\left[(F_T - \bar{F}_t)^2\right]\end{aligned}$$

where the lowest-order approximation of $E[F_T]$ was used in the last step. We obtain

$$E[F_T] \approx \bar{F}_t - \frac{1}{2}\frac{h''(\bar{F}_t)}{h'(\bar{F}_t)}\mathrm{Var}(F_T)$$

It is well known that by using a number ξ between F_t and F_T, the Taylor expansion can be made exact:

$$h(F_T) = h(F_t) + h'(F_t)(F_T - F_t) + \frac{1}{2}h''(\xi)(F_T - F_t)^2$$

If h is convex and $F_t = E_t[F_T]$, *Jensen's inequality* follows by taking expectations of both sides:

$$E[h(F_T)] \geq h(E[F_T])$$

3.10 Futures Contracts

We have so far seen how dynamic replication can be used for modeling European options and other products with fixed-time payoffs. Later in the book, we discuss many more product types that can be priced with this technique. A particularly common product type to which dynamic replication can be applied is the futures contracts, which is the topic of this section. The arguments in this section are not limited to futures contracts but also apply to any *over-the-counter (OTC)* contract (i.e. non-exchange traded products) for which a clearing house offers themselves to be a central counterpart.

To describe the purpose of futures in financial markets, we first recall some of the main features of the closely related forward contract. Holding this contract gives as much exposure to the underlying as if the underlying itself would have been bought. As $V(t) = 0$ at initiation, it is possible to obtain a non-zero exposure for a zero initial cost. Needless to say, this feature is attractive to speculators. At the same time there is a risk that one of the counterparties fails to satisfy the obligations at maturity, should the market move unfavorably. One of the advantages of futures contracts is that much of this credit exposure is eliminated. As we see below, futures contracts retain many of the attractive features of forward contracts but offer only a small credit risk.

Just as for a forward contract, a *futures contract* is an agreement to exchange an underlying S for a cash amount F' at T. The futures price F' is the fair price, but it is determined in a different way than the forward. If the futures price moves from F'_t to $F'_{t+\delta t}$ from one day till the next, a *variation margin* payment $F'_{t+\delta t} - F'_t$ has to be made between the counterparties. This *daily settlement* reduces the credit exposure and it leads to a futures price F' that for short maturities is close to the forward price F. For longer maturities, on the other hand, the price difference can be substantial.

There are several practical differences between forwards and futures. For instance, futures contracts are exchange traded and the payments and credit exposure are through a clearing house. The credit exposure is further reduced by an initial deposit, the *initial margin*, from each counterparty to the exchange. From a modeling and pricing perspective, however, we are mainly concerned with the effect of the daily settlement. The implication is that a futures contract always has zero value after the settlement has been made. This is in contrast to a forward contract that is worth $S_t - F_0 P_{tT}$ at t (disregarding the credit impact on the pricing), where F_0 is the strike price determined at initialization $t = 0$.

We follow Cox et al. (1981) and determine the fair price F' when daily settlements are made. For this purpose, consider a portfolio U with initial value $F'_0 = F'_{t=0}$. Let this amount be invested in the money market account from today $t = 0$ to tomorrow $t = \delta t$. If the continuously compounded interest rate for this period is r_0, the investment yields the amount $e^{r_0 \delta t} F'_0$ at $t = \delta t$. Assume that we also invest in, at zero cost, $e^{r_0 \delta t}$ futures contracts. If the futures price equals F'_1 at δt, the daily settlement requires a payment of $e^{r_0 \delta t}(F'_1 - F'_0)$. The portfolio is therefore worth $U_1 = e^{r_0 \delta t} F'_1$ at δt. We proceed in the same way by investing the amount U_1 in the money market until $t = 2\delta t$ and enter $e^{r_1 \delta t} e^{r_0 \delta t} - e^{r_0 \delta t}$ more futures contracts, where r_1 is the interest rate for the period $[\delta t, 2\delta t)$. As we already had $e^{r_0 \delta t}$ contracts, the total number of contracts held is $e^{(r_0 + r_1)\delta t}$. The portfolio is therefore worth $U_2 = e^{(r_0 + r_1)\delta t} F'_2$ at $t = 2\delta t$. Repeating the procedure up to maturity gives the amount $U_T = e^{\int_0^T r dt} S_T$ as $F'_T = S_T$.

The above argument implies that today's price of a payment $e^{\int_0^T r dt} S_T$ at T equals the futures price F'_0 at $t = 0$ as the portfolio U that replicated the payment had this initial value. Alternatively, the payment can be priced under the risk-neutral measure:

$$F'_0 = U_0 = E\left[U_T / e^{\int_0^T r dt}\right] = E[S_T] = E[F'_T]$$

from which we conclude that the futures price is a martingale under the risk-neutral measure.

To compare futures and forward prices, we price a forward contract under the risk-neutral measure. The forward F_0 is then given by

$$\begin{aligned} 0 &= E\left[(S_T - F_0)/e^{\int_0^T r dt}\right] = E\left[S_T e^{-\int_0^T r dt}\right] - F_0 E\left[P_{TT}/e^{\int_0^T r dt}\right] \\ &= E\left[S_T e^{-\int_0^T r dt}\right] - F_0 P_{0T} \\ \Leftrightarrow F_0 &= P_{0T}^{-1} E\left[S_T e^{-\int_0^T r dt}\right] \end{aligned}$$

The corresponding computation in the forward measure would, of course, show that the forward is the expectation of the spot value at maturity, i.e. forward prices are martingales in the forward measure. We obtain the relation

$$\begin{aligned} F_0' &= E[S_T] = F_0 + E[S_T] - P_{0T}^{-1} E\left[S_T e^{-\int_0^T r dt}\right] \\ &= F_0 + E[S_T] - P_{0T}^{-1}\left(\mathrm{Cov}\left(S_T, e^{-\int_0^T r dt}\right) + E\left[S_T\right] E\left[e^{-\int_0^T r dt}\right]\right) \\ &= F_0 - P_{0T}^{-1}\mathrm{Cov}\left(S_T, e^{-\int_0^T r dt}\right) \end{aligned}$$

We see that the difference between the futures price and the forward price is determined by the covariance between the underlying and the inverse money market account. In particular, the prices are equal if interest rates are deterministic. For T not too large, $e^{-\int_0^T r dt}$ is close to one and its volatility is therefore small. Thus, the difference between forward and futures prices is most pronounced for long maturities. As the covariance between two variables is independent of the measure, see Appendix, the covariance can be measured in the real-world measure, for example, by a historical analysis. The correction term needed to price a futures contract from the linear instrument of a forward contract is called a convexity adjustment.

There are a couple of practical aspects that often affect the price difference between forward and futures contracts more than that of the convexity adjustment. For instance, the forward price needs to include a premium for the counterparty risk. For the futures contract, on the other hand, there are certain costs associated with the trading. For example, although interest rate is paid by the clearing house on the collateral it is not always the best rate that can be found in the market. There are also fees that need to be paid to the clearing house. Companies therefore often prefer to use forward contracts if they consider their counterparty bank to be safe.

An option on a futures contract pays $(F'_{\bar{T}} - K)_+$ at the maturity $\bar{T}$ of the option. Assuming deterministic interest rates, the futures price is equal to the forward price, meaning that the option can be priced with the techniques that were used earlier in this chapter. The general case can be solved by working in the risk-neutral measure, see, for example, Sect. 12.3.

3.11 Settlement Lag

When purchasing a financial asset V, the payment does in general not take place today t but at a later date t'. The payment date t' is defined in the contract specification as a certain number of business days after the trade date. The number of business days between the two dates is called the *settlement lag* and is typically equal to 0, 1, 2 or 3 but can also be much longer. The corresponding date t' is referred to as the *settlement date*. Each financial market has its own standard convention of the settlement lag that is used unless stated explicitly. The settlement date is then referred to as the *spot date*. As we mainly consider standard contracts, we use the terminology spot date and settlement date interchangeably.

The contract calendar that classifies days into business days and holidays does not have to coincide with the holiday calendar of a specific country. For instance, an exchange can define its own calendar that can be different from the calendar of the country in which it is located. Another example is the *target calendar* which relates to the Euro zone and does not exactly coincide with the calendars of the member countries.

A payment date for a product can either be given explicitly as a date or indirectly via a *tenor*, i.e. a time period, which usually extends from the spot date. For example, a 1M forward contract traded on April 13 has spot date 15 April (assuming no holidays between these dates and a 2 day settlement lag) and the payment takes place May 15. As payments cannot take place on holidays, it might be necessary to *holiday adjust* the payment date with the contract-specific calendar to the *following* or the *previous* business day. In the former case the payment could end up the next month should the spot date be located in the end of the month. This can be avoided by using the previous convention in this particular instance. The resulting holiday adjustment is the most commonly used. It is called *modified following*. Similarly, *modified previous* is defined by using the next business day should the previous business day be located in the preceding month.

If the spot date is one of the last dates in a month, it might be necessary to compute the payment date from the last date in the forward month. For instance, the payment date of a 1M forward contract with spot date January 30 is determined by applying the holiday adjustment to the last day in February, i.e. 28 (non-leap year) or 29 (leap year).

If the spot date is the last business day of a month, the *end-of-month rule* is sometimes used. It means that the payment date is the last business day in the forward month. For example, assume that the spot date is April 29 and that the next day is a holiday. The 1M forward date is then the holiday adjusted date obtained from May 31.

Forward contracts are often cash settled, which means that the amount $S - K$ is exchanged between the counterparties instead of performing the actual purchase of S by K. Just as there is a settlement lag between the trade date t and the spot date t', there is a lag between the date T when S is read off from the market and the date T' when a forward contract pays $S - K$. The *fixing date* or *reset date* T is the date that has T' as its spot date. If there can be more than one date that has T' as its spot date, as for FX markets, it is custom to let T be the earlier of these dates. We conclude that the reset date cannot be obtained directly from the trade date but needs to be computed via the spot date and the payment date.

The above construction of dates is standard and does not only apply to our illustrating example of forward contracts. For instance, a 6M European call option pays $(S_T - K)_+$ at the holiday adjusted date T', 6M from the spot date, where T is the earliest date with spot date T'. For option and forwards, the date T is referred to as the *expiry* or maturity and T' as the *delivery date*.

When entering a contract at the trade date t, the amount V_t that needs to be paid at the spot date t' is called the price of V at t. If it were possible for a customer to make an upfront payment, the fair amount would be $D_{tt'} V_t$, where D is the discount factor. We refer to this amount as the value of V at t. Thus, in our terminology, the value of a contract is different from the price when the settlement lag is non-zero. The price $P_{tT'}$ of a zero-coupon bond is therefore equal to $D_{t'T'} := D_{tT'}/D_{tt'}$ and does not coincide with the discount factor $D_{tT'}$, defined as the t-value of a payment of \$1 at T', which explains the change of notation in this section.

Most often, only the price is of concern. For example, traders are not interested in having values displayed on the trading screen but only want to see the prices as this is what they quote their clients. For a quantitative modeler, on the other hand, it is important to be aware of the difference between these concepts. For instance, even though a certain number of settlement days might be the practice in a specific market, there can be products that violate this convention, e.g. overnight deposits in interest rate markets and cash deals in FX markets.

To explain the impact of the settlement lag on derivatives pricing, we start by considering a forward contract that pays $S_T - K$ at the payment date T'. To replicate this payment, S must be bought today and sold at T at the same time as we go short K zero-coupon bonds maturing at T'. This requires an initial cost of $S_t - KP_{tT'}$ to be paid at the spot date t'. We conclude that the forward price equals $F_t = P_{tT'}^{-1} S_t$.

A derivative can have a different settlement lag than the underlying. To illustrate the consequences of different lags, consider the (unrealistic) example when the forward market has zero lag. The pricing then involves a payment of S_T at T, which through inverse discounting is equivalent with a payment $D_{TT'}^{-1} S_T$ at T'. This latter payment can be approximately priced by assuming deterministic interest rates (or that they are uncorrelated with the underlying), meaning that the contribution from the discount factor can be read off from today's yield curve. This approach is for obvious reasons not possible when the underlying is an interest rate product and this instance is dealt with in Sect. 13.4.

We now consider an option V that pays $(S_T - K)_+$ at the payment date T'. We assume that the option market and the underlying market have equal settlement lag.

We also assume that interest rates are deterministic, which means that the discount factor between two future dates is independent of today's date.

To reconnect with the option pricing techniques discussed earlier in this chapter, we introduce hypothetical zero-lag contracts $\bar{S}$ and $\bar{V}$ that have the same values as S and V. The prices are then related by $\bar{S}_t = D_{tt'} S_t$ and $\bar{V}_t = D_{tt'} V_t$. We have

$$\bar{V}_T = D_{TT'} V_T = D_{TT'} (S_T - K)_+ = (\bar{S}_T - \bar{K})_+$$

where $\bar{K} = D_{TT'} K$. According to Sect. 3.4, the price is

$$\bar{V}(t, S) = \bar{S}_t N\left(d_+\right) - D_{tT} \bar{K} N\left(d_-\right), \quad d_\pm = \frac{\ln(\bar{S}_t/(D_{tT}\bar{K}))}{\sigma\sqrt{T-t}} \pm \frac{1}{2}\sigma\sqrt{T-t}$$

which implies that

$$V(t, S) = S_t N\left(d_+\right) - D_{t'T'} K N\left(d_-\right), \quad d_\pm = \frac{\ln(S_t/(D_{t'T'}K))}{\sigma\sqrt{T-t}} \pm \frac{1}{2}\sigma\sqrt{T-t}$$

Thus, the settlement-lag effect on option pricing is that the discounting needs to be done from T' to t' instead of from T to t. The volatility, on the other hand, still needs to be measured from t to T.

To obtain as simple formulae as possible in the book, we have chosen to only include the effect of the settlement lag in this section. It is, however, important to account for the settlement lag for liquid vanilla products as the impact can exceed the bid-offer spread. For exotics, the lag is of less importance, but as it is simple to account for, we recommend financial modelers to include it in the pricing, not the least to obtain agreement between exotics and vanillas in limiting cases.

Bibliography

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–654

Cox JC, Ingersoll JE, Ross SA (1981) The relation between forward prices and futures prices. J Financial Econ 9:321–346

Gradshteyn IS, Rhyzik IM (2007) Table of integrals, series, and products, 7th edn. Academic Press, New York

Chapter 4
Derivatives Modeling in Practice

This chapter covers general derivatives modeling questions such as:

- What do the various market participants expect from a derivatives model?
- How can the model variables be determined?
- How should a model be used?
- What are the model limitations?
- How should a model be tested?

To avoid too abstract a discussion, we frequently narrow down the subject to the pricing of European call options within the Black–Scholes model. We assume, for simplicity, that interest rates are zero, implying that $F = S$.

We start the chapter by considering the incentives of the various participants in the derivatives market. That provides us with an understanding of the demands on models and how they should be used. We then review various calibration techniques followed by a section on hedging. We discuss model limitations and round off the presentation with an overview of model testing methods.

4.1 Model Applications

To illustrate the various application areas of a derivatives model, we first discuss the motivation for trading derivatives. One of the main uses of derivatives is for speculation. Consider, for example, a portfolio consisting of S/V European call options V. The portfolio is worth as much as the underlying itself, but changes to lowest order by $S/VV_S dS$ when the underlying changes by dS. As the *leverage* S/VV_S is greater than 1, the exposure to the underlying has increased by purchasing call options instead of the underlying itself. In the Black–Scholes model, the leverage is an increasing function with limits 1 for $K \to 0$ and ∞ for $K \to \infty$. This spectrum of the leverage makes options ideal for speculative purposes.

Derivatives can also be used to reduce the risk that comes from price fluctuations. For instance, a copper mining company can protect itself from a fall in the copper

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_4,

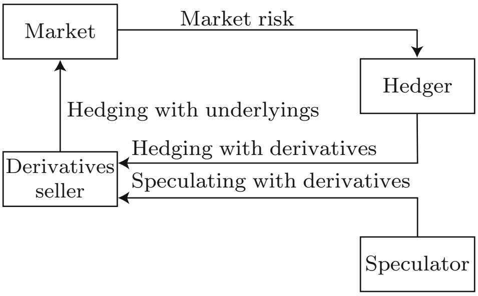

Fig. 4.1 The role of derivatives in the market

price by entering futures contracts or by purchasing put options. Because of the leverage, it is possible to reduce the risk for a limited cost. A strategy that results in the reduction of the risk to a financial variable is referred to as a *hedging* strategy.

The hedging with and speculation in derivatives opens up for another type of market participants: the derivatives sellers. The sellers are mainly banks dealing with derivatives to supply the demand by the hedgers and speculators. They do not want to be exposed to any risk themselves and therefore hedge derivatives by going long the replicating strategy predicted by their models (Fig. 4.1).

There are several reasons why the hedgers and speculators do not go long the replicating strategy themselves. For example, they might not have the interest as it takes them off their core business area. Another reason is the various demands necessary for a successful replication, including the know-how, the access to market places and a trading platform. It is also important to have the volumes required for liquid trading and contacts with counterparties that might be interested in taking the opposite position of the deal. Furthermore, there is a risk of an unsuccessful replication because of human error, system faults, bad choice of modeling assumptions, etc.

A derivatives model results in an analytical or numerical expression of the price in terms of the model variables, for example t, S and σ. The (mathematical) derivative of the price with respect to these variables measures the *sensitivity* of the model, which can be used to compute the *market risk*. Market participants can be interested in only the price, only the risk or both the price and the risk. Let us now look at what the different market participants expect of a derivatives model under various circumstances.

Any derivatives model useful to a hedger should obviously take into account the original market risk, i.e. the exposure that the hedge should reduce. As the nature of this risk is different from business to business, it is impossible to apply a general model in this situation.

For the seller of a derivative, the problem is more isolated and a general derivatives model can often be applied. The model is used in fundamentally different ways depending on whether the derivative is quoted or not in the market. If it is quoted, the model needs to be calibrated to the market price and should be used for hedging purposes, i.e. for the construction of the replicating portfolio. In this instance, the risk of the model is the important component as it states how the derivative should be hedged. If, on the other hand, the price is not quoted, neither for the derivative under consideration nor for any related derivative, the model variables cannot always be calibrated but need to estimated, for example, by a time-series analysis of the underlying. In this situation, both the risk and the price are important.

The situation is often somewhere between the two extreme cases above, i.e. there is no quote for the specific derivative that is going to be sold, but only for related derivatives on the same underlying, for example, with different strikes or maturities. The model can then be calibrated to the quoted derivatives by either matching the quotes exactly or approximately, e.g. through a least squares fit. In this instance, the pricing is simply an interpolation (or approximate fit) and is not as important, i.e. model dependent, as the risk. An approximate calibration of derivatives quotes can also be used by a speculator to identify cheap and expensive derivatives in the market, meaning that the price of the model is of major importance.

The model price can be important even for derivatives quoted in the market. Consider, for example, the situation when 3M options are liquidly traded and we buy one of those. Viewed from the day of purchase, the model price is not important as the correct price can be found in the market. Later on, on the other hand, the option has a maturity below 3M and then it is no longer possible to read off its price from the market. Even though we might not be interested in selling the option to a third party, it is still useful to provide a valuation. For instance, it is often necessary to have a measure of the credit exposure towards the counterparty.

A typical situation is that the bank that sells a derivative knows some other bank that quotes better prices. The bank can then make a *back-to-back deal*, which means that the derivative is simultaneously bought from another bank and sold to the client. The original bank then acts as a broker. Back-to-back deals are often made between local banks, with strong relationship to the industry of their country or region, and investment banks in the major trading centers New York, London and Tokyo, but are also common between investment banks. The model price of a back-to-back deal is unimportant at the trade date as it is provided by the counterparty bank. At later dates the model price becomes important for the measurement of the credit exposure toward the two counterparties. The market risk is less important as two opposite positions have been taken. For this reason, local banks are sometimes exclusively interested in model prices.

4.2 Calibration

A derivatives model depends on three types of variables: contract specific such as K and T for an option, observables such as S and t, and unobservables such as σ in the Black–Scholes model. To be able to use a model, the unobservable variables first need to be determined. This can be done either by backing them out, i.e. *calibrating* them, from market data on derivatives or by estimating them, for example, from a historical analysis of the underlying. The latter approach can, of course, only be pursued if there is a financial interpretation of the model variable.

Care needs to be taken when estimating the variables as the price is then be heavily dependent on how well the model reflects the reality. For instance, historical data reveals that the implied volatility of options is closely related, but still different from the volatility in a time-series analysis of the underlying. The difference can be much bigger for other types of model variables such as the volatility of volatility in stochastic volatility models. Furthermore, the estimation of variables is often backwards looking, i.e. historical data is used, while derivatives are forward looking. Sampling of historical data also depends on the length of the time series and the observation frequency, e.g. daily or weekly. It is therefore dangerous to use the approach of estimation of model variables in isolation. As an attempt of improvement, consider the pricing of an option in the situation when there exists an option quote for a closely related underlying. If the implied volatility of the quoted option is, for example, 1% above the historical volatility, then by assuming the same to be true for the option to be priced, a better estimate is hopefully obtained for the implied volatility.

Because of the fundamental difficulties associated with the estimation of model variables, this approach should only be used when the market is too illiquid to allow for calibration. The advantage of calibrating to market prices is that some of the responsibility put on the model gets removed. Indeed, from a pricing perspective the model then acts as an interpolator between the calibration instruments and the limiting constrains (such as $V \to S - KP_{0T}$ when $K \to 0$ for European call options). Furthermore, as we explain in the next section, much of the risk towards the underlying can be removed by hedging with the calibration instruments.

The calibration of a model amounts to the determination of the unobservable variables $\{\sigma_i\}$ so that the model prices $\{\mathcal{C}_k^{\text{model}}(\{\sigma_i\})\}$ of the calibration instruments match, or are a good approximation of, the market prices $\{\mathcal{C}_k^{\text{market}}\}$. We have used the notation $\{\sigma_i\}$ for the unobservables to relate with the Black–Scholes model for which there is a single unobservable variable σ; it does not mean that these variables have to be volatilities.

Derivatives can be classified into two groups: *vanillas* and *exotics*. The former group consists of European call and put options and other simple products such as their digital counterparts. The group of exotic options comprises everything else and they are typically path dependent or higher dimensional. The model calibration is usually done to liquid vanilla derivatives, such as ATM European options. Once the calibration has been done, the model can be used to price exotic options or non-liquid vanillas such as call options on non-quoted strikes.

The calibration of a model can be done numerically through an iterative process. The first step is to make an initial guess of $\{\sigma_i\}$, which can be based on the last successful calibration. Once the choice has been made, the model prices can be computed. Based on information such as the differences $\{\mathcal{C}_k^{\text{market}} - \mathcal{C}_k^{\text{model}}(\{\sigma_i\})\}$ and the Jacobian $\left\{\frac{\partial \mathcal{C}_k^{\text{model}}}{\partial \sigma_i}\right\}$, a better estimate for $\{\sigma_i\}$ can be found. This procedure is repeated until convergence is reached.

The calibration instruments $\{\mathcal{C}_k\}$ can usually be priced with models of the Black–Scholes type meaning that each of them only depends on a single unobservable $\tilde{\sigma}_k$ that has the interpretation as a volatility. For instance, the LIBOR market model for interest rates is formulated in terms of caplet volatilities $\{\sigma_i\}$ but is often calibrated to swaptions that are quoted in the market in terms of their implied volatilities $\{\tilde{\sigma}_k\}$, see Chap. 13. To put fair weights on OTM and ITM derivatives, the calibration is typically done by minimizing the elements in the vector $\{\tilde{\sigma}_k^{\text{market}} - \tilde{\sigma}_k^{\text{model}}(\{\sigma_i\})\}$ instead of the price differences.

As the calibration instruments are priced repeatedly during the calibration process, not only do they have to be included in the set of products that can be priced by the model, but the pricing needs to be fast as well. These instruments are therefore often priced analytically within the model or, alternatively, through an analytical proxy formula. The calibration instruments should be liquid enough for reliable and up-to-date market quotes to exist and they should be chosen in such a way that the unobservable variables can be backed out. The latter statement means that not only should there be at least as many calibration instruments as there are unobservable variables, but also that the instruments should depend collectively on all the variables.

Let us first focus on the instance when there are as many calibration instruments as there are unobservables. If the calibration instruments are chosen appropriately, there is a unique solution of the unobservables that price them correctly. Furthermore, there exist efficient numerical algorithms for finding the solution, e.g. the Newton-Raphson root finder. The result is model prices that are identical with the market prices for the calibration instruments. At first, this appears to be exactly what we want to achieve. Indeed, the prices of a model can obviously not be trusted should the calibration instruments be mispriced. The problem with the approach originates in the fact that even though the calibration instruments are assumed to be liquid, their market prices are often not in agreement with what one might expect. For instance, some of the quotes can be a bit out of date or someone might just have made a bid that was too high considering the current market conditions. Thus, the market quotes of the calibration instruments are not part of a smooth curve: some quotes appear too high and some too low. A model with an exact pricing of the calibration instruments therefore inherits this type of non-smooth behavior.

An example of the consequences of non-smooth market data can be seen in the method of *bootstrapping*, which means that the unobservable variables are backed out iteratively. The first calibration instrument is chosen to depend on a single unobservable variable which therefore can be backed out. The second calibration instrument is chosen to depend on two variables: the one that was just backed out and one additional unobservable. It is again possible to back out the unknown variable. If the calibration instruments are chosen appropriately, these steps can be applied repeatedly until all variables have been backed out. A bootstrapping method has the advantages of high performance and easy implementation. Indeed, for a numerical calibration it is only necessary to use 1-dimensional root finders instead of being confronted with the harder problem of finding a higher-dimensional root.

To understand the impact of non-smooth market data on bootstrapping, assume that the first instrument is quoted inaccurately. The resulting error propagates, and possibly gets magnified, through the bootstrapping equations. The consequence can be that the bootstrapping equations fail to be solvable. Note that the problem is generic and bootstrapping was only used to illustrate how unsolvable equations can arise. The problem often originates from the fact that the root does not exist, which leads to the failure of any root-finding technique.

The problem of non-smooth market data can partly be circumvented by using more calibration instruments than unobservables. It is then no longer possible to match the market prices exactly. Instead, an approximate method needs to be used, for example, by minimizing the sum of the squared differences between the model and market prices for the calibration instruments. Less emphasis is then put on single data points, resulting in a smoother fit. The scheme can be made more sophisticated by using multiplicative weights on the squared differences in the minimization procedure. The weights can be chosen proportionally to the importance we assign to the various quotes, usually determined by the liquidity of the corresponding products.

The approximate fit to the market data is typically implemented with numerical methods designed to find a minimum value (for the sum of the squared differences). Unfortunately, these methods are usually not as successful as methods designed to find a unique root. The main reason is that the convergence is often towards a local minimum instead of the global minimum. In fact, it is difficult to determine numerically whether the convergence really is towards the global minimum. A serious consequence is that, because of the changing market data, the next calibration can be towards a different minimum, leading to an artificial jump of the model prices. The same effect can occur when computing the risk by bumping the variables. Apart from being unstable and not always returning the correct solution, the methods for finding a global minimum are in general slower than root-finding algorithms.

As an illustrating example of the situation when there are fewer unobservable variables than calibration instruments, consider the calibration of the Black–Scholes model to ATM options maturing at $T_1, T_2, \ldots, T_n$. The model can be calibrated by choosing the volatility that minimizes the sum of the squared differences of the model and market implied volatilities for the calibration instruments. As only a single free parameter is used to match n quotes, the fit is likely to be poor. A better result can be obtained by allowing the unobservable variables to depend on the observable variables. For instance, in our example we can allow the volatility to depend on t. The following parametric form of the volatility is popular:

$$\sigma(t) = \alpha + (\beta + \gamma t)e^{-\delta t}, \quad \delta > 0$$

which can be interpreted as an interpolation between a linear function $\alpha + \beta + \gamma t - \beta\delta t$ for small t and a constant function α for large t. The four parameters can be determined, for example, by a least squares minimization method. The result is a smooth model curve with a relatively close match to the market prices of the calibration instruments. Even more degrees of freedom can be introduced by allowing the volatility to be a continuous and piece-wise linear function with discontinuous derivatives at $T_1, T_2, \dots, T_n$. It is then possible to match the market quotes exactly. Another possibility that leads to a smoother result is to let the volatility be given by a spline with node points at $T_1, T_2, \dots, T_n$.

The technique of allowing the unobservable variables to depend on the observable variables is a powerful tool for calibration. It is easy to implement and can be calibrated to a flexible set of market quotes. In fact, the method is so successful that it is sometimes possible to calibrate models that are an inherently bad representation of the problem. Therefore, it is important to ensure that the functional dependence of the variables is not abused, but only used for small adjustments of a model that we already believe to be good. If the unobservables show too big a functional dependence after the calibration, the model has probably been tweaked too much and its predictive powers are lost. Models for which the unobservable variables are allowed to depend on the observable variables are generalizations of so-called local volatility models, the topic of Chap. 6.

The case of fewer unobservables than calibration instruments can alternatively be handled by allowing the unobservable variables to have different values for different market data points. For instance, in the above example it is for each i possible to correctly price the T_i maturing option through the Black–Scholes model with volatility equal to the corresponding implied volatility $\sigma_{\mathrm{imp},i}$. The pricing of a general T-maturing option can be done with a Black–Scholes volatility obtained through an interpolation or an approximate fit from the set $\{\sigma_{\mathrm{imp},i}\}$. For instance, the pricing of the $T = (T_i + T_{i+1})/2$ maturing option can be done with the volatility $(\sigma_{\mathrm{imp},i} + \sigma_{\mathrm{imp},i+1})/2$.

The two approaches above are different as we either allow the unobservable variables to depend on observable variables or on contract specific variables. In our example, it means that the time interpolation (or approximate fit) is either done in terms of the local volatility σ or in terms of the implied volatility σ_{imp}. By introducing a volatility weighted time variable according to

$$\tilde{t}(t) = \int_0^t \sigma(u)^2 du$$

in the computations in Sect. 3.4, we conclude that the implied volatility and the local volatility are linked by

$$\sigma_{\mathrm{imp}}(T) = \sqrt{\frac{1}{T}\int_0^T \sigma(u)^2 du}$$

The two calibration schemes are therefore closely related. For instance, a linear interpolation of the local volatility corresponds to a non-linear interpolation of the implied volatility and vice versa.

A disadvantage of the local-volatility interpolation is that the local volatilities need to be backed out from the market quotes, which are given by the implied volatilities. Fortunately, these computations are often relatively straightforward.

An advantage of local-volatility interpolations is that the local volatility has a financial meaning: it measures the level of the instantaneous fluctuations of the underlying. This is in contrast to the implied volatility which does not have any financial interpretation and is simply a number. For example, consider the situation when we are given two ATM market quotes at T_i and T_{i+1} and would like to price ATM options with maturities in $[T_i, T_{i+1}]$. Having no additional information, the simplest possible model seems to be to assume that the local volatility varies linearly between these two points. A linear interpolation in the implied volatility, on the other hand, corresponds to a local-volatility interpolation with a less natural model interpretation.

We have seen that once a model has been constructed, we can choose between (1) keeping it the way it is, (2) generalizing it by allowing the unobservable variables to depend on the observable variables, or (3) generalizing it by allowing the unobservable variables to depend on the contract specific variables. The fundamental difference between the approaches lies in the hedge and in the pricing of products with a different flavor than the calibration instruments. The former is covered in the next section while we now turn our attention to the latter. For this purpose, we use the example of European call option pricing and rely on the fundamental theorem of asset pricing to write the price as

$$U(F, K, t, T) = \int (F_T - K)_+ p(T, F_T; t, F) dF_T$$

implying that

$$\frac{d^2}{dK^2} U(F, K, t, T) = p(T, K; t, F)$$

for some Green's function p. We conclude that there is a bijective correspondence between option prices and Green's functions. As prices and implied volatilities also are bijectively related, there exists a bijective correspondence between implied volatilities and Green's functions for fixed t and T. An implied volatility interpolation can therefore be interpreted as an interpolation in the forward coordinates of the Green's function, giving us the function $p(T, F_T; t, F_t)$ for the maturities T to which we calibrate, where t is today's date. This is in contrast to the outcome when using the constant volatility Black–Scholes model, or the generalization to S- and t-dependent volatility, when $p(T, F_T; t', F_{t'})$ is obtained for arbitrary t' in (t, T). As path-dependent derivatives depend on conditional distributions, they cannot be priced when interpolating the implied volatility.

The problems occurring when the unobservables depend on the contract specific variables can also be illustrated by, for example, the pricing of a knock-out option with barrier B and strike K. This product can be priced using a local volatility equal to $\sigma_{\text{imp}}(K)$, i.e. the same volatility that is used for the corresponding call option. Unfortunately, it is not clear whether this is the correct volatility to be used. Indeed, as the barrier option also depends on the volatility when the underlying is close to the barrier, we could just as well have used $\sigma = \sigma_{\text{imp}}(B)$, or perhaps $\sigma = \sigma_{\text{imp}}(\frac{1}{2}(K + B))$. The pricing of barrier options therefore becomes ambiguous when using the method of implied volatility interpolation.

We have seen that the pricing of products that have a type different from the calibration instruments can be problematic when the unobservable variables depend on the contract specific variables. To be fair, there are also some issues with the pricing of such products for the unmodified model, or when the model is generalized so that the unobservable variables depend on the observable variables. The reason for these issues can be understood by considering the pricing of path-dependent derivatives with a model based on a process (an SDE) that has been calibrated to European option quotes, i.e. the Green's function is such that the skews and the smiles are as observed in the market. It is well known, however, that there are several influencing factors, e.g. supply and demand effects, for the implied volatility skew and smile apart from the propagation of the underlying. It means that the calibration assigns unrealistic probability weights to the paths of the underlying, with the consequence that path-dependent derivatives are mispriced, for instance, by using a too high probability that a knock-out level is reached for a barrier option.

We conclude that there are some issues associated with using the same model for vanilla options and path dependent derivatives, as the former product type is typically priced with models for which the full skew and smile behavior is determined by the underlying process. The problems could be solved by using more sophisticated models explaining the skew and smile via other effects as well, but there are currently no popular models of this type. We are therefore sometimes forced to use one model for vanilla options and another for path-dependent derivatives.

To understand the implication of using different models for vanillas and exotics, consider again the pricing of a knock-out option. When the barrier is far away from the spot, the product converges to the corresponding vanilla option. In this instance, we end up with two models for the same product: one vanilla model and one exotic model, resulting in different prices and hedges. This type of arbitrage within a pricing system can be problematic. One way to avoid it is to use the method of adjusters that is introduced in the next section.

In later chapters, we give examples of how it is possible to calibrate skew and smile models with constant variables. When calibrating to several maturities, it is often necessary to use time-dependent variables. One possible approach to calibrate models with time-dependent variables is to identify the corresponding model with constant variables that gives the same prices for a certain fixed maturity. For instance, the Black–Scholes model with constant volatility $\sqrt{\int_0^T \sigma(u)^2 du / T}$ gives the same T-maturity prices as when using the time-dependent volatility. The same kind of idea can be used to replace time-dependent skew and smile variables with constant variables. The most popular technique is currently that of Markovian projection, see, for example, Piterbarg (2006).

Let us turn our attention to the implementation of a calibration, with a focus on the performance. Because of the repetitive pricing of the calibration instruments, the calibration of a model is often much slower than the pricing itself. This is particularly the case if no analytical proxy formula is used for the calibration instruments, or if there are fewer model variables than there are calibration instruments, meaning that an approximate fit has to be made. Depending on the implementation, the majority of the calibration time is spent on computing the Jacobian $\{\frac{dC_k}{d\sigma_i}\}$ for the transformation between the unobservable model variables and the calibration instrument prices. Fortunately, the matrix elements can be computed independently by bumping the variables one at a time. It means that the computations can be done in parallel. Another way to increase the performance is to separate the calibration from the pricing and only calibrate at certain times, typically overnight. Needless to say, this suffers from the risk of a large intra-day market move. Smaller market moves can, however, be accounted for by the method of adjusters, see next section.

The number of pricing iterations in a numerical calibration is typically 10–100. One of the reasons for this high number is that the calibration must be much more accurate than the bump size used for the greeks to avoid any distortion to the risk management (see the next section). It might therefore appear that numerical calibration is unfeasible from a performance perspective and analytical calibration is the only possibility. With the speeding-up methods of the previous paragraph, however, we conclude that this is not necessary the case. Furthermore, development in computing, including multicore technology and the possibility to perform calculations on the graphics card, are also in favor of numerical calibration. We now discuss some further advantages and speeding-up possibilities that can be done with a numerical calibration. We choose to focus on the instance when the pricing and calibration is done with a *Monte Carlo simulation*, which means that a (pseudo) random number generator is used for simulating the paths of the underlyings.

The numerical calibration with Monte Carlo can be speeded up by pricing all calibration instruments in the same simulation. It is also useful to have the same settings (the same random numbers, simulation dates, number of simulation paths, etc.) for the calibration as for the pricing (and for the risk). Once the calibration has been done, the calibration instruments are then priced correctly in the pricing simulation, even if only a single simulation path is used. Viewing pricing as an interpolation of the calibration instrument prices, there are two sources of errors: from the calibration instrument prices and from the interpolation. In a numerical calibration with the same calibration and pricing settings, the first source of error is removed as the calibration instruments are priced accurately in the pricing simulation. This is in contrast to analytical calibration for which the errors in the calibration instrument prices originate both in the Monte Carlo noise of the pricing simulation and in the approximate value from the closed-form proxy formulae. The numerical noise means that, for a given pricing accuracy, fewer number of paths are needed for the numerical calibration than for an analytical calibration. Furthermore, the numerical calibration does not introduce any systematic error like those coming from approximative closed-form expressions.

Although the calibration instruments are relatively simple, typically being call options, the pricing models can be complex, meaning that suitable closed-form formulae can only be found for special types of SDEs. When doing a numerical calibration, on the other hand, there is no such constraint and the SDE can be chosen according to our needs. With a clever choice of SDE, it can also be possible to implement a substantially faster pricer, see Sect. 13.16 for a specific application.

There exists various approaches for speeding up the computation of parameter sensitivities in Monte Carlo simulations, such as the adjoint method by Giles and Glassermann (2006). With such techniques it is also possible to improve the performance for computing $\{\frac{d\mathcal{C}_k}{d\sigma_i}\}$.

A practical advantage of calibrating through simulation is that the pricer itself is used for the calibration, meaning that only little extra implementation work needs to be done. Furthermore, it is possible to calibrate to any product that can be priced with the pricer. For instance, it is straightforward to calibrate a yield curve model to skew and smile, something that is problematic to do analytically and has triggered a lot of research.

Some of the unobservable variables can be left uncalibrated. The user of the model can then decide on the purpose of these variables. For instance, the variables can be used to manually calibrate the model to one or several instruments that are considered to be of particular importance for the product that is priced. Alternatively, a variable might have a financial interpretation for which we have a specific view of the value. For instance, even though the correlation matrix between LIBOR rates can be calibrated from cap and swaption prices, it is usually left as a free input to yield curve models. It is popular to assume a functional form of the correlation and then manually calibrate or estimate the functional parameters, see Sect. 12.4. As an additional example, consider the situation when a seller of an exotic derivative finds out that the model price is incorrect. This information can come from, for instance, a more advanced (but slower) model or from a derivatives buyer claiming that other clients offer prices in a different range. With a free variable, the model can be made to match what is believed to be the correct price. Of course, the model parameters should not vary too much (or should not vary at all) from deal to deal as this might be a sign of fundamental weaknesses with the model. It is then useful to have a financial interpretation of the model variables to get an intuition of the size of the adjustments that are made.

4.3 Risk Management

A bank that sells a derivative is exposed to a market risk. By analyzing how the value of a derivative changes with the market moves, we now describe how the risk can be partially removed by hedging appropriately. We consider a fixed derivative, meaning that the contract specific variables are constant and the price only depends on the unobservable variables $\{\sigma_i\}$ and on the observables variables, e.g. the current time t and the underlying S. To simplify notation, we assume that there is only a single underlying S and a single unobservable σ, which means that the model price can be written as $V(S,\sigma,t)$.

Assume that at a later time $t+dt$, the underlying value is $S+dS$ and the calibrated unobservable variable is $\sigma + d\sigma$. It means, according to the model, that the seller of the derivative has made a gain (or loss) of

$$\begin{aligned}&-V(S+dS,\sigma+d\sigma,t+dt)-(-V(S,\sigma,t))\approx\\&-\frac{dV}{dS}dS-\frac{dV}{d\sigma}d\sigma-\frac{dV}{dt}dt-\frac{1}{2}\frac{d^2V}{dS^2}(dS)^2-\frac{d^2V}{dSd\sigma}dSd\sigma-\dots\end{aligned}$$

We observe that the risk is measured by the (mathematical) derivatives $\frac{dV}{dS}$, $\frac{dV}{d\sigma}$, $\frac{dV}{dt}$, $\frac{d^2V}{dS^2}$, $\frac{d^2V}{dSd\sigma}$, ... The collective name for these derivatives is *greeks* as they are usually denoted by Greek symbols. We now discuss the role of the greeks in the Black-Scholes model, one at a time.

The derivative's exposure to the underlying can be partially eliminated by holding $\frac{dV}{dS}$ number of underlyings. A change in the underlying then implies a change $\frac{dV}{dS}dS$ which compensates the lowest order gain of the derivative from the move of the underlying. $\frac{dV}{dS}$ is called the *delta* Δ and the strategy of going long Δ number of underlyings is referred to as delta hedging the derivative. The delta is equal to $N(d_+)$ in the Black–Scholes model.

The exposure to the unobservable variable σ is to the lowest order measured by $\frac{dV}{d\sigma}$, called the *vega* Λ of the derivative. Observe that the underlying has a delta: $\frac{dS}{dS}=1$, and can be used to delta hedge the derivative, but not a vega as $\frac{dS}{d\sigma}=0$. An instrument depending on σ, i.e. another derivative, is therefore necessary for a vega hedge.

Dynamic hedging of derivatives involves frequent buying and selling of the hedging instruments. Because of the bid-offer spread, this can be a costly affair. These transaction costs can, however, be limited by choosing appropriate hedging instruments. For instance, the hedging instruments should be as liquid as possible since this implies small bid-offer spreads. Also, note that the underlying is usually more liquid than derivatives, which means that it is cheaper to delta hedge than to vega hedge. The cost of the vega hedge can be reduced by hedging with derivatives that have a large vega. To explain how this can be done, consider the Black–Scholes model where we have

$$\frac{dV}{d\sigma}=S\sqrt{T-t}n(d_+)$$

Plotting the vega as a function of K shows a bell-shaped form that goes to zero for both large and small K. Indeed, for large K we know that the call option is almost worthless and the volatility is therefore irrelevant. For small K, on the other hand, we are certain to exercise the option and again there is only a weak volatility dependence. The maximum of the vega is given by

$$\frac{d^2V}{d\sigma dK}=0\Leftrightarrow d_+=0\Leftrightarrow K=Se^{\sigma\sqrt{T-t}/2}$$

The quantity $\sigma\sqrt{T-t}$ is in general small, which leads to a maximum at $K \approx S$. If the computation had been done with non-zero interest rates, the same result would have been obtained with S replaced with F, i.e. the vega assumes its maximum close to the forward. As options are usually most liquid for $K = F$, ATM instruments are ideal for the vega hedge as only a small amount is needed at the same time as bid-offer spreads are small.

Let us now turn our attention to the derivative $\frac{dV}{dt}$, called the *theta* Θ. It is a measure of the gain (or loss) in a derivative from t to $t + dt$ when the market quotes are unchanged between these two points in time. The role of Θ in risk management is fundamentally different from that of Δ and Λ. For a detailed explanation of Θ, assume σ to be constant. According to Sect. 3.2, the delta hedge then gives an exact replication of the derivative, if performed continuously in time. It means, in particular, that the replicating portfolio and the derivative must have an equally large theta. The role of the theta can therefore be understood by analyzing the replicating portfolio.

To explain how the replicating portfolio can gain or lose money from t to $t + dt$ when the market is unchanged: $S(t + dt) = S(t)$, consider the discretization $t_0 = t, t_1 = t+dt/2, t_2 = t+dt$, where we for simplicity have used only one intermediate time point. We use the notation $S_i = S(t_i)$ and $\Delta_i = \Delta(t_i)$, which means that $S_2 = S_0$. The replicating portfolio consists of Δ_0 underlyings S_0 at t_0. At t_1, this part of the portfolio is worth

$$\Delta_0 S_1 = \Delta_1 S_1 + (\Delta_0 - \Delta_1) S_1$$

i.e. it consists of Δ_1 number of underlyings and $(\Delta_0 - \Delta_1)S_1$ cash. At t_2, the underlying is assumed to have moved back to S_0, so the new value is

$$\Delta_1 S_0 + (\Delta_0 - \Delta_1) S_1 = \Delta_0 S_0 - (\Delta_1 - \Delta_0)(S_1 - S_0) \to \Delta_0 S_0 - \frac{d^2V}{dS^2}(S_1 - S_0)^2$$

Assuming a lognormal process: $dS = \sigma S dW$, we obtain the gain

$$-\frac{d^2V}{dS^2}(S_1 - S_0)^2 = -\sigma^2 S^2 \frac{d^2V}{dS^2}(W_1 - W_0)^2 \sim -\frac{1}{2}\sigma^2 S^2 \frac{d^2V}{dS^2} dt$$

since the Brownian motion changed from W_0 to W_1 in the time $\frac{1}{2}dt$. This result can be made exact by using more intermediate points and approaching the continuous limit. By definition, the gain must be equal to the theta, implying the derivatives pricing equation

$$\frac{dV}{dt} = -\frac{1}{2}\sigma^2 S^2 \frac{d^2V}{dS^2}$$

which was initially derived in Sect. 3.4.

The theta is in general negative for European call options. For example, if the volatility is time-independent in the Black–Scholes framework, the option depends

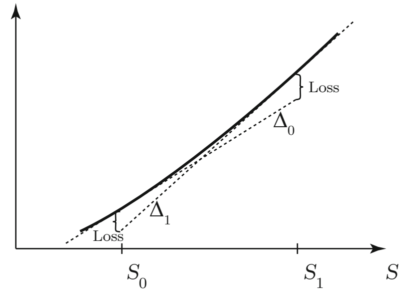

Fig. 4.2 The theta loss in the hedging strategy occurs because the underlying is bought at high values and sold at low values

on t and T only through the combination $T - t$. Thus, the no-arbitrage relation $\frac{dV}{dT} \geq 0$ implies a negative theta: $\frac{dV}{dt} \leq 0$. According to the pricing equation, $\frac{d^2V}{dS^2}$ must be positive, i.e. the option price is a convex function of S.

To understand in detail why the term $\frac{d^2V}{dS^2}(dS)^2$ leads to a loss, assume that $S_1 > S_0$ in the above discretization. The positivity of $\frac{d^2V}{dS^2}$ implies that $\Delta_1 > \Delta_0$, i.e. the holdings of the underlying in the replicating portfolio increase when the underlying goes up and decrease when it goes down. It is this strategy of buying more underlyings at high values and selling off at low values that leads to a loss in the time value when replicating options (Fig. 4.2). The same result is obtained in the case $S_0 > S_1$. As the loss is proportional to the *gamma* $\Gamma = \frac{d^2V}{dS^2}$, it follows that the higher the gamma, the higher the gain in the hedging strategy and the higher the loss in being long an option. Since the Black–Scholes gamma

$$\frac{d^2V}{dS^2} = \frac{1}{S\sigma\sqrt{T-t}} n(d_+)$$

has its maximum close to ATM, it is for these options that the time-value is the greatest.

According to the Black–Scholes equation, options decrease in value by

$$\frac{dV}{dt} = -\frac{1}{2}\sigma_{\text{imp}}^2 S^2 \Gamma$$

if the underlying value does not move between two time steps. As the corresponding loss for the hedging strategy is

$$\frac{dV}{dt} = -\frac{1}{2}\sigma^2 S^2 \Gamma$$

where σ is the realized volatility, the combined strategy of selling an option and delta hedging it leads to the gain

$$\frac{1}{2}\int_0^T \left(\sigma_{\text{imp}}^2 - \sigma^2\right) S^2 \Gamma dt$$

We conclude that the gain is positive if the realized volatility is lower than the implied volatility at times when $S^2\Gamma$ is large.

For a deterministic volatility σ, the delta hedge provides a perfect replication of a derivative if continuous-time trading is permitted. In practice, however, the restructuring of the hedge is done discretely in time, typically daily or when a large market move occurs. The reason is not only because it is impossible to hedge continuously in reality, but also since it would imply large transaction costs. When limited to a discrete-time hedge, the underlying moves a finite amount between hedging times and not infinitesimally as previously assumed. To avoid as much transaction costs as possible, the best practice is to only rehedge when Δ has changed by a certain amount. It means that rehedging should be done frequently when Δ changes a lot, that is when Γ is large. For administrative reasons, however, rehedging is typically done daily.

Assume that the underlying moves from S_0 to S_1 during a small, but finite, time period dt. Being short an option leads to a gain

$$\begin{aligned}-V(t+dt,S_1)+V(t,S_0) &= -\Theta dt - V(t,S_1)+V(t,S_0)-\dots\\ &= -\Theta dt-(S_1-S_0)\Delta-\frac{1}{2}(S_1-S_0)^2\Gamma-\dots\end{aligned}$$

where the higher order terms have been omitted. Being long the hedging strategy leads to a gain $(S_1-S_0)\Delta$. The total position therefore has the gain

$$-\Theta dt-\frac{1}{2}(S_1-S_0)^2\Gamma-\dots$$

which means that the break-even points where no gain or loss is made is to the lowest order given by

$$S_1=S_0\pm\sqrt{-\frac{2\Theta}{\Gamma}dt}=S_0(1\pm\sigma_{\text{imp}}\sqrt{dt})$$

We conclude that the lowest order gain has a maximum at S_0 and zeros located one standard deviation away from S_0 in the Black-Scholes model. The daily hedge of an option therefore leads to positive gains 68% of the time. The gains cannot exceed Θdt while the losses can be unlimited (Fig. 4.3).

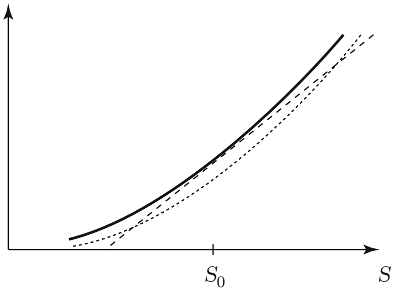

Fig. 4.3 Comparing the change in the option price with the delta hedging strategy for a small time step

To avoid the remaining exposure for a delta hedged option it is common to also gamma hedge. The underlying has zero gamma: $\frac{d^2S}{dS^2} = 0$, which means that a derivative is required for this purpose. As the maximum of the Black–Scholes gamma occurs close to ATM, it is standard to use ATM options for this purpose. The *vanna* $\frac{d^2V}{dSd\sigma}$ and the *volga* $\frac{d^2V}{d\sigma d\sigma}$ are also often hedged, but their contribution is typically smaller than that of the gamma.

The hedging could have been done for even higher orders such as V_{SSS}. This is usually not done as the model already contains several approximating assumptions, e.g. the form of the underlying process, that most likely have a higher impact than these higher order derivatives.

A derivatives dealer typically manages a large portfolio of derivatives, with long and short positions, all priced and hedged within the same model. For practical reasons, and to minimize the cost of hedging, the deals are not hedged individually, but the risk parameters, such as delta, vega and gamma, are aggregated to a hedge for the whole portfolio. Observe that the Black–Scholes gamma and vega only differ by a factor containing the time to maturity and the value of the underlying. It means that for a portfolio of options with the same underlying and with similar maturities, a vega hedge automatically implies an approximate gamma hedge.

The hedge values can be computed analytically for simple models as the Black–Scholes model. For more advanced models, it is necessary to compute the risk by bumping the variables one by one and then revalue the deal (or portfolio). The hedge values are usually computed daily or when a large market move occurs. They also need to be computed at the time a deal is made in order to do an immediate hedge.

It happens frequently that a derivatives trader chooses not to fully hedge a derivative. One reason could be that the vega hedge is expensive and the trader instead accepts the risk associated with a volatility move. Another reason could be that the trading desk believe in increasing volatility and therefore choose not to vega hedge as this would lead to a loss, assuming that we are long volatility.

Recall from the previous section that some variables are only (manually) calibrated or estimated at pricing date. Since they are not recalibrated from day to day, it is not possible to dynamically hedge the exposure. The dealer is then subjected to unhedged price fluctuations. Furthermore, incorrect model greeks are obtained for the other variables. To illustrate this fact, assume that the Black–Scholes volatility is determined at pricing date and then used throughout the deal. At a later date, not only the volatility is incorrect but also the Black–Scholes delta, as it depends on the volatility. For an additional example, consider the pricing of a deal that pays the floored difference $(S_1 - S_2)_+$ of two assets at time T. Assume that both S_1 and S_2 follow lognormal processes where the volatilities are calibrated and hedged, while the correlation ρ is estimated only at the pricing date. To understand the impact on the greeks, assume that the deal is ITM and that the correlation between the underlyings is high. The deal is approximately replicated by going long 1 unit of S_1 and short 1 unit of S_2. If the correlation decays during the lifetime of the trade, the volatility of the difference $S_1 - S_2$ increases with the consequence that fewer underlyings are needed for the hedge as the chance of the deal ending up OTM at maturity is greater. Thus, chosing not to calibrate the correlation leads to an incorrect delta hedge.

To make the theory behind hedging more explicit, consider the situation when we are short a European option $V(K, T)$ and would like to hedge this exposure. We assume that hedging is done in continuous time so that only first-order hedges are of interest. The corresponding ATM option is usually liquidly traded, which means that the purchase of $b = \frac{dV(K,T)}{d\sigma} / \frac{dV(K_{\mathrm{ATM}},T)}{d\sigma}$ ATM options gives a volatility-independent portfolio $-V(K, T) + bV(K_{\mathrm{ATM}}, T)$. The next step is to purchase $c = \frac{dV(K,T)}{dS} - b\frac{dV(K_{\mathrm{ATM}},T)}{dS}$ underlyings to obtain a portfolio where the S-dependence has been eliminated. The result is that

$$V(K, T) = bV(K_{\mathrm{ATM}}, T) + cS + A$$

where A is both underlying- and volatility-independent. The value of A can be computed from the model price of the difference between the left-hand side and the first two terms on the right-hand side. It should be small, assuming that the derivative is well replicated by its hedging instruments.

The above formula instructs us that once $V(K, T)$ has been sold, we should immediately purchase b ATM options and c number of underlyings for the hedge. To cover the true hedging costs, we need to charge the customer for the market price of the ATM option, leading to the price

$$V(K, T) = bV^{\mathrm{market}}(K_{\mathrm{ATM}}, T) + cS + A$$

By the same token, the market price of S should be used and not the model price, which could be different, for example, in a Monte Carlo implementation. In summary:

$$V(K,T) = V^{\mathrm{model}}(K,T) + b(V^{\mathrm{market}}(K_{\mathrm{ATM}},T) - V^{\mathrm{model}}(K_{\mathrm{ATM}},T)) + c(S - S^{\mathrm{model}})$$

from which we see how the last two terms use the market prices of the hedging instruments to adjust the model price of the derivative.

We move on to the general theory of hedging and let $\{h_j\}$ denote the hedging instruments. We determine the optional hedge, i.e. the weights $\{b_j\}$ such that $V - \sum_j b_j h_j$ is as independent as possible of the market. We continue to assume continuous-time hedging to avoid complications from higher-order hedges. As the market data only enters the pricing via the calibration instruments $\{\mathcal{C}_k\}$, the weights $\{b_j\}$ should be chosen so that the vector $\left\{\frac{dV}{d\mathcal{C}_k} - \sum_j b_j \frac{dh_j}{d\mathcal{C}_k}\right\}$ has as small entries as possible.

Let us first consider the situation when the hedging instruments are used to calibrate the model, i.e. $\mathcal{C}_k = h_k$. If they also are independent, $\frac{dh_j}{d\mathcal{C}_k} = \delta_{jk}$, the hedge weights are given by $b_j = \frac{dV}{d\mathcal{C}_j}$. When the calibration instruments are volatility dependent and can be written as $\mathcal{C}_k = \mathcal{C}_k(\tilde{\sigma}_k)$, for some volatility $\tilde{\sigma}_k$, the weights can be computed from $\frac{dV}{d\mathcal{C}_j} = \frac{dV}{d\tilde{\sigma}_j}\frac{d\tilde{\sigma}_j}{d\mathcal{C}_j}$. The latter factor is typically known analytically (it is the inverse vega in the Black–Scholes model) while the former can be computed, for example, by a bump and revalue approach.

There are two schools of modeling techniques when $\mathcal{C}_k = h_k$. One of them uses a general model for a large class of products while the other one uses models that are tailor-made for each product. We now discuss these techniques in detail by considering European option pricing in the Black–Scholes model.

The first technique can be illustrated by using the Black–Scholes model, calibrated to an ATM option, for the pricing of options with the same maturity but arbitrary strikes. As the hedging instrument is assumed to coincide with the calibration instrument, it is given by the ATM option, which is not an ideal hedge for ITM and OTM options. In the second technique, the calibration is product specific. It means that the calibration is done to the liquid option that has strike closest to the option that we want to price. We then end up with several Black–Scholes models, calibrated to options with different strikes. Although the pricing is more accurate with this approach, the risk (hedge) aggregation and netting is not consistent as different models are combined. Furthermore, a product-specific calibration needs to be done at the same time as the pricing, which can have a performance impact. In our example of the Black–Scholes model, these problems can be avoided by using more advanced models that takes skew and smile into account. However, there exist other examples for which it is not easy to find more advanced models, for instance, the use of the LMM to price ITM and OTM Bermudan swaptions, see a further discussion later in this section.

One attempt to get the best of both worlds is to let the hedging instruments be different from the calibration instruments. The calibration instruments can then be chosen to be general enough so that a large class of products can be priced within the model. At the same time, the hedging instruments can be tailor-made to each product resulting in accurate hedging and, as we explain below, it leads to accurate pricing as well. Furthermore, the calibration can be separated from the pricing and done overnight.

Let us describe how to find the weights $\{b_j\}$ when the calibration instruments are different from the hedging instruments. It is then necessary to minimize the vector $\left\{\frac{dV}{d\mathcal{C}_k} - \sum_j b_j \frac{dh_j}{d\mathcal{C}_k}\right\}$, or equivalently $\left\{\frac{dV}{d\tilde{\sigma}_k} - \sum_j b_j \frac{dh_j}{d\tilde{\sigma}_k}\right\}$. Introducing the vector $(\Lambda_V)_k = \frac{dV}{d\tilde{\sigma}_k}$ and the matrix $(\Lambda_h)_{kj} = \frac{dh_j}{d\tilde{\sigma}_k}$, we see that if there are as many hedging instruments as there are calibration instruments and if Λ_h is invertible, the choice $b = \Lambda_h^{-1}\Lambda_V$ makes the derivative fully hedged. If the matrix is non-invertible (for example, if there are fewer hedging instruments than there are calibrated variables), b can be chosen to minimize the sum of squares in the vector. The result is $b = (\Lambda_h^T\Lambda_h)^{-1}\Lambda_h^T\Lambda_V$. There are also other possible criteria that can be used to determine b in this situation. For instance, the squared sum of a weighted vector $\left\{\lambda_k(\frac{dV}{d\tilde{\sigma}_k} - \sum_j b_j \frac{dh_j}{d\tilde{\sigma}_k})\right\}$ can be minimized. If we believe that a variable $\tilde{\sigma}_k$ varies widely during the lifetime of the derivative, the corresponding weight λ_k should be chosen to be a large number.

If there are more hedging instruments than there are calibration instruments, there are often several combinations of $\{b_j\}$ that eliminate the risk vector. A natural solution is the smallest hedge with this property, i.e. the minimum of $b^T b$ under the condition $\Lambda_V - \Lambda_h b = 0$. The result is $b = \Lambda_h^T(\Lambda_h \Lambda_h^T)^{-1}\Lambda_V$. A straightforward generalization is to minimize the squared sum of a weighted vector $\lambda^T b$. The entries in the vector λ can then be chosen to be large if the corresponding hedging instrument has a large bid-offer spread. Alternatively, the additional degrees of freedom can be used to minimize higher-order terms such as the gamma, vanna or volga.

When the hedging instruments are chosen different from the calibration instruments, their model prices may no longer agree with the market prices, even when an exact calibration is done. This is important to account for as the product price can obviously not be completely trusted if the hedging instruments are priced inaccurately. The incorrect pricing of the hedging instruments can also happen when they are equal to the calibration instruments, for instance, when an approximate calibration is made (although this has the advantage of smoothing out rough market data). Furthermore, the hedging instruments are mispriced when an intra-day market move occurs after an overnight calibration. We now follow Hagan (2004) and extend the method of model adjustment to a more general setting.

After selling a derivative, we immediately hedge it by purchasing b_j quantities of the hedging instruments for the price $\sum_j b_j h_j^{\text{market}}$. We charge the buyer the cost of the hedge plus the price of the remaining exposure. The latter is given by $V - \sum_j b_j h_j$ and must be computed with a derivatives model. It results in the price

$$V = \sum_j b_j h_j^{\text{market}} + (V - \sum_j b_j h_j)^{\text{model}} = V^{\text{model}} + \sum_j b_j \left(h_j^{\text{market}} - h_j^{\text{model}}\right)$$

being charged to the customer. We see that apart from the computation of the hedge weights $\{b_j\}$, the model is only used for correcting the price through the difference between the derivatives model price and the model price of the hedge. Certainly, it is also possible to use an alternative model for the computation of the difference. If the hedging instruments $\{h_j\}$ and the model are chosen appropriately, it is possible for each instance in time to find a vector $\{b_j\}$ such that the linear combination $\sum_j b_j h_j$ approximately represents the market behavior of the derivative, resulting in a weak model dependence of the difference.

The adjustment technique can mistakenly appear not to have any impact on the risk as in both the non-adjusted and the adjusted case, $\{b_j\}$ hedging instruments are bought. To show the advantage of adjustment for risk management, consider the situation of holding a derivative that is hedged with $\{b_j\}$ hedging instruments, resulting in the portfolio $V - \sum_j b_j h_j$. As market prices exist for the hedging instruments, they can be used for the pricing which leads to a portfolio of value $V - \sum_j b_j h_j^{\text{market}}$. Now, the non-adjusted price is $V^{\text{model}} - \sum_j b_j h_j^{\text{market}}$ and changes by $dV^{\text{model}} - \sum_j b_j dh_j^{\text{market}}$ when the market moves. This change is not necessarily zero, which implies that the hedge is imperfect. For the adjusted portfolio value, the change is given by

$$\begin{aligned} dV &- \sum_j b_j dh_j^{\text{market}} \\ &= dV^{\text{model}} - \sum_j b_j dh_j^{\text{model}} + \sum_j \left(h_j^{\text{market}} - h_j^{\text{model}} \right) db_j \end{aligned}$$

It follows from the defining expression for $\{b_j\}$ that the difference between the first two terms on the right-hand side is relatively small. Furthermore, the model is assumed to be good, which means that the model prices of the hedging instruments are close to the market prices. We conclude that the last term is small and thereby the whole right-hand side. The hedge weights $\{b_j\}$ therefore provide an accurate description of the risk of the adjusted price.

To fully understand the adjustment method when the hedging instruments are different from the calibration instruments, we show how a general model can be tailor-made using a particular set of hedging instruments suitable for the product that is priced. More specifically, we consider the example of pricing a European call option $V(K, T)$ with the Black–Scholes model calibrated to the ATM quote with the same maturity. We assume that there is a liquidly traded option $V(K', T)$ with strike that is close to K. To account for the skew and smile effects, we hedge with this option rather than the ATM option. The adjusted price is equal to

$$V(K,T) = V^{\text{model}}(K,T) + b(V^{\text{market}}(K',T) - V^{\text{model}}(K',T)) + c(S - S^{\text{model}})$$

As we are working in the Black–Scholes model, the model price of the underlying agrees with the market price: $S^{\text{model}} = S$. With σ being the calibrated model volatility, i.e. the implied volatility of the ATM option, and σ' the implied volatility of the K' option, we obtain

$$V(K,T) = \text{BS}(K,\sigma) + \frac{\frac{d}{d\sigma}\text{BS}(K,\sigma)}{\frac{d}{d\sigma}\text{BS}(K',\sigma)} \left(\text{BS}(K',\sigma') - \text{BS}(K',\sigma) \right)$$

If σ' is close to σ, the expression within the brackets can to the lowest order be written as $(\sigma' - \sigma)\frac{d}{d\sigma}\text{BS}(K', \sigma)$, resulting in

$$V(K,T) \approx \text{BS}(K,\sigma) + (\sigma' - \sigma)\frac{d}{d\sigma}\text{BS}(K,\sigma) \approx \text{BS}(K,\sigma')$$

We see that for σ' close to σ, the adjustment has moved the price from $\mathrm{BS}(K,\sigma)$ to $\mathrm{BS}(K,\sigma')$ which is an improvement as the hedging instrument was assumed to be close to $V(K,T)$.

We now look at an alternative application of the adjustment formula. To avoid too abstract a discussion, we illustrate the method of Ekstrand (2010) by an example from interest rate derivatives. More specifically, we consider the pricing of a callable floored CMS spread product. We assume that counterparty A makes quarterly payments to counterparty B that are proportional to the difference between two swap rates. We let the payments be floored at 0 to ensure that they always are positive. In return, B makes payments proportional to a LIBOR rate. At each payment date (or a couple of days before), counterparty A has the right to call the deal, i.e. to decide for the future payments to stop. Such deals are typically priced with yield curve models such as the LMM model, see Chap. 13. This type of models is usually good in deciding when it is optimal for A to call the deal and in determining the price for having the callability embedded into the deal. With regards to the pricing of the individual cash flows, there exist other models that do a better job. For instance, the pricing of a single payment proportional to a swap rate is often done by static replication with swaptions, see Sect. 13.4. The pricing of a floored CMS spread payment can then be done by combining a static replication approach with a copula model, see Sect. 10.1.

The above example illustrates a situation that frequently occurs when pricing exotic derivatives: the derivative contains two (or more) important features, but the models can each only take proper account of one of the features. In our case there is a copula model with static replication that can price the individual cash flows but not the callability and there is a yield curve model that can price both features but we are not too confident about the quality of the cash-flow pricing. We now explain how adjusters can be used to get the best of both worlds.

For an abstract description of the problem, let M be a model that can price a product V as well as a set $\{V_j\}$ of simpler products for which there exists a more accurate model M'. For reasons that will soon become clear, we refer to M' as an adjustment model. In the above example, V is the callable floored CMS spread product and V_j the corresponding floored CMS spread payment at time T_j. M is a yield curve model while the adjusted model M' is a copula model with marginal distributions based on static replication.

In parallel with how the hedging coefficients were determined earlier in this section, we choose $\{b_j\}$ to be such that $V - \sum_j b_j V_j$ is as independent as possible of the model parameters in M. More precisely, they are computed by minimizing the (weighted) vector $\left\{\frac{dV}{d\tilde{\sigma}_k} - \sum_j b_j \frac{dV_j}{d\tilde{\sigma}_k}\right\}$, where $\{\tilde{\sigma}_k\}$ are the implied volatilities of the calibration instruments that belong to M. It is then possible to use the more accurate model M' for $\sum_j b_j V_j$ to obtain

$$V = \sum_j b_j V_j^{\text{adj.model}} + (V - \sum_j b_j V_j)^{\text{model}} = V^{\text{model}} + \sum_j b_j \left(V_j^{\text{adj.model}} - V_j^{\text{model}} \right)$$

This method is essentially identical with the approach earlier in this section where adjusters were used with the market quotes of the hedging instruments. Indeed, the main idea is to approximate the model dependence of a product with a simpler set of products for which more accurate prices can be obtained, either from the market or from more specialized models. In the example of the callable floored CMS spread, the weights $\{b_j\}$ can be chosen as the probabilities of calling the deal at the corresponding dates. The advantage is that these probabilities are often a byproduct of the pricing which leads to a reduced computing time.

The adjustment improves not only the price, but also the risk. This is because the adjustment models are more accurate and can often take into account skew, smile and the dynamics of the volatility curve. Furthermore, the numerical stability is usually improved. To understand the details behind this statement, let us return to the example of the callable floored CMS spread and assume for the sake of the argument that the payments are digital, i.e. a fixed amount is paid if the swap spread is positive. The implementation of yield curve models then contains unwanted numerical noise deriving, for example, from Monte Carlo simulations. The advantage of the adjustment method is that the pricing can usually be done analytically for cash flow models, leading to more stable results.

We further illustrate the benefits of adjustment models by looking at an additional example: the pricing of Bermudan swaptions. We assume that floating and fixed interest rate payments are exchanged (e.g. 6M LIBOR versus 5%) and that one of the counterparties has the right to call the deal. This product is also typically priced with yield curve models as the LMM. Unfortunately, it is difficult to account for the skew, smile and the proper dynamics in this kind of model, see Chap. 13. It means that in the extreme case when the Bermudan swaption is sure to be called at a specific date, we effectively end up with a rather poor swaption model. As typical pricing software already have an advanced model for vanilla products as swaptions, the result is two vanilla models: one poor model derived from the extreme case of the exotics pricer, and a good model tailor-made for vanillas. Being short an exotic product that is certain to be called at a specific date and long the corresponding swaption does therefore not lead to a perfect hedge in the pricing system. This inconsistency between exotic models and vanilla models has troubled quantitative analysts for a long time. The inconsistency disappears when using adjusters as this approach can be viewed as a natural way to calibrate exotics models to vanilla models. Furthermore, the inconsistency in the greeks also disappears as the skew, smile and dynamics of vanilla models become induced into exotic models via the adjustment.

When using an adjustment model,

$$V = \sum_j b_j V_j^{\text{adj.model}} + (V - \sum_j b_j V_j)^{\text{model}}$$

it is possible to adjust the expression further, for example, with adjusters obtained from the market quotes of the hedging instruments. It means that $V_j^{\text{adj.model}}$ should be replaced with

$$V_j^{\text{adj.model}} + \sum_i c_{ji} \left(h_i^{\text{market}} - h_i^{\text{adj.model}} \right)$$

and $(V - \sum_j b_j V_j)^{\text{model}}$ with

$$(V - \sum_j b_j V_j)^{\text{model}} + \sum_i a_i \left(h_i^{\text{market}} - h_i^{\text{model}} \right)$$

where $\{c_{ji}\}$ and $\{a_i\}$ are determined as usual. The resulting price is then given by

$$V = V^{\text{model}} + \sum_j b_j \left(V_j^{\text{adj.model}} - V_j^{\text{model}} \right)$$
$$+ \sum_{ji} b_j c_{ji} \left(h_i^{\text{market}} - h_i^{\text{adj.model}} \right) + \sum_i a_i \left(h_i^{\text{market}} - h_i^{\text{model}} \right)$$

We have assumed that the adjustment model has the same hedging instruments as the original model. This is clearly not necessary and it is straightforward to extend the computations to the general case.

To understand the computational cost associated with the adjustment formula, note that three components are needed to adjust the classical result V^{model}, namely h_j^{market} or $h_j^{\text{adj.model}}$, h_j^{model} and b_j. If the former equals h_j^{market}, it can be immediately read off from the market. If, on the other hand, it is equal to $h_j^{\text{adj.model}}$, the performance cost is negligible as the adjustment model typically is (semi-)analytic. For the model prices of the hedging instruments, there is typically a performance cost. However, if the pricer is such that the hedging instruments can be priced simultaneously with V, for example, in a common Monte Carlo simulation, the cost is low.

Whether there is a performance impact from the computation of the hedge weights $\{b_j\}$ depends on the situation. Let us first consider some instances for which the impact is negligible. The first example is when the hedge weights are obtained as a byproduct of the pricing, for instance, when they represent exercise probabilities. Another example regards the adjustment of a whole portfolio with the hedging instruments. In this situation the hedge weights are also obtained for free as they are already computed for the risk management. In general, the hedge weights are not known and as their computation is a magnitude slower than that of the non-adjusted price, the performance can suffer.

There are also several reasons why the performance should be improved by the adjustment. For instance, since much of the pricing responsibility is transferred to the adjustment model, which typically is high performing, the original model does no longer have to be a state-of-the-art model and it is instead possible to use a high-speed model. Furthermore, much of the risk is also transferred to the more accurate adjustment model. As the risk is usually the most performance demanding part of a model, substantial improvement can be achieved by, for example, reducing the number of paths in a Monte Carlo simulation.

We leave the method of adjusters and return to the example of the previous section where the Black–Scholes model was calibrated to ATM options maturing at $T_1, T_2, \ldots, T_n$. We assume that these calibration instruments are used for the hedging as well. In the instance of an approximate calibration through a parametric function, a change in any of the T_i option prices affects the option prices for all maturities. It means that an option needs to be hedged with a combination of all the calibration instruments. We immediately see the absurdity: an option with maturity before T_1 should obviously be hedged with the T_1 option and possibly with the T_2 and T_3 option but certainly not with the T_n option. The same effect occurs for an exact calibration when using a spline for the interpolation as a change in a single point impacts the whole curve. This unwanted non-local risk is the price to pay when using smooth curves obtained from global interpolation, i.e. an interpolation depending on all the node points, with and not only the neighboring.

The alternative is to do an exact calibration using a local interpolation scheme as the linear interpolation. The change of a node point then only affects the curve between the two neighboring node points. The disadvantage of this approach is a less smooth curve.

The two extreme cases we described above both have their shortcomings and it is therefore natural to look for hybrids of them so that the negative effects are less pronounced. We consider exact calibrations and discuss an interpolation technique referred to as a *tension spline*. This interpolation method has a free parameter which on the one extreme turns the interpolation linear and on the other extreme turns the interpolation into a spline. The tension spline can be visualized as a rubber band connecting the node points, with the free parameter being the tension.

To derive the mathematical expression for the tension spline, observe that the linear interpolation can be obtained by minimizing the length $\int (y'(x))^2 dx$ of the curve while the cubic spline minimizes the curvature $\int (y''(x))^2 dx$. The tension spline, on the other hand, minimizes a linear combination $\int \big(\theta(y'(x))^2 + (1-\theta)(y''(x))^2\big)\, dx$ of these quantities, where the free parameter $\theta \in [0, 1]$ is the tension. Using calculus of variations, we obtain

$$y(x) = Ae^{\omega x} + Be^{-\omega x} + Cx + D, \quad \omega = \sqrt{\theta/(1-\theta)}$$

Just as for ordinary cubic splines, see Press et al. (2002), the tension spline is patched together at the node points in the unique (up to choices at the end points) way making the curve, its derivative and second derivative continuous. This leads to a curve that can be tuned between local and global interpolation and thereby controlling locality of the risk.

We now change topic and consider the concept of dynamics for an option model, defined as the dependence of the implied volatility on the underlying. From

$$\frac{dV}{dS} = \frac{\partial V}{\partial S} + \frac{\partial V}{\partial \sigma_{\text{imp}}} \frac{\partial \sigma_{\text{imp}}}{\partial S}$$

we see how the dynamics affect the delta and thereby the hedging. As mentioned in Sect. 3.6, the dynamics are usually observed to be somewhere between sticky strike and sticky delta. For hedging purposes, it is therefore important to have a model that reflects this behavior.

Consider the example of the Black–Scholes type model with a volatility depending on the underlying. It follows that

$$\frac{dV}{dS} = \frac{\partial V}{\partial S} + \frac{\partial V}{\partial \sigma} \frac{\partial \sigma}{\partial S}$$

It is interesting to note that the function $\sigma(S)$, which is determined through the calibration to market quotes, also gives the dynamics. Thus, the choice of model type, and the corresponding calibration, determines the dynamics. For this purpose, we discuss the three most popular fundamental model types, local volatility models, stochastic volatility models and Lévy models, and their dynamics in Chaps. 6–8.

We return to the discussion of models for which the unobservable variables are allowed to depend on the contract specific variables and consider the example of implied volatility interpolation. Recall that the calibration gives $p(T, F_T; t, F_t)$ for various T and F_T but only for the fixed t and F_t given by today's values. It means that we have no information about the F_t-dependence and it therefore needs to be imposed externally. We described in Sect. 3.6 how this can be done for a pure sticky-strike or sticky-delta behavior. It is also possible to impose a mixture of these extremes, for example, by setting $\sigma_{\text{imp}}(T, K; t, F) = \sigma_{\text{imp}}(T, K/F^{\beta}, t)$.

Care needs to be taken when imposing the dynamics, as we now show with an example when the implied volatility is interpolated in the strike direction. For the sake of argument, we impose a sticky-delta behavior $\sigma_{\text{imp}} = \sigma_{\text{imp}}(K/S)$, and assume that the implied volatility at $K = (K_i + K_{i+1})/2$ is obtained from linear interpolation, i.e. $\sigma_{\text{imp}} = (\sigma_{\text{imp},i} + \sigma_{\text{imp},i+1})/2$. Taking the S-derivative on both sides and using the fact that $S\partial_S g(K/S) = -K\partial_K g(K/S)$ for any differentiable function g, we obtain:

$$\begin{aligned} &K\partial_K \sigma_{\text{imp}}(K/S)\big|_{K=(K_1+K_2)/2} \\ &\quad = \frac{1}{2}\Big(K_1 \partial_K \sigma_{\text{imp}}(K/S)\big|_{K=K_1} + K_2 \partial_K \sigma_{\text{imp}}(K/S)\big|_{K=K_2}\Big) \end{aligned}$$

This constraint means that it is not possible to impose arbitrary dynamics on volatility curves. Indeed, the imposed dynamics do not necessarily commute with the interpolation, i.e. different results are obtained depending on whether we first take the S derivative at the node points and then interpolate, or vice versa. To retain consistency, the dynamics can be imposed only at the calibration points $K_1, \ldots, K_n$, from which the dynamics at a general K follow.

As the dynamics modify the delta through the dependence of the implied volatility on the underlying value, one might expect this effect to cancel out if we also vega hedge. To investigate whether this is true, assume that we are long an option V that is vega hedged with an ATM option V_{ATM}, resulting in a vega-neutral portfolio: $\frac{d}{d\sigma}(V - bV_{\text{ATM}}) = 0$. When the underlying moves, the implied volatilities of the two options change in certain ways while the dynamics of the model predict a different change. The outcome is a mispricing with amounts $d\sigma$ and $d\sigma_{\text{ATM}}$. If these two changes are equal, the hedged portfolio is not mispriced to the lowest order:

$$\begin{aligned} V(\sigma + d\sigma) - bV(\sigma_{\text{ATM}} + d\sigma) &\approx V(\sigma) - bV(\sigma_{\text{ATM}}) + d\sigma \frac{d}{d\sigma}(V - bV_{\text{ATM}}) \\ &= V(\sigma) - bV(\sigma_{\text{ATM}}) \end{aligned}$$

If, on the other hand, the changes in volatilities are unequal, the dynamics are of importance for vega hedged portfolios as well.

The implied volatility curve shifts often almost in parallel when the underlying changes, for instance when there is a pure skew and no smile. Conditional on that we vega hedge, the above discussion implies that we obtain a good delta hedge as long as we use a model that predicts a parallel shift, even if the shift size incorrect. This is important because we later see that the most popular skew and smile models often give a parallel shift in the implied volatility that is fundamentally different from the observed shift. It means that these models yield good delta hedges only if they are vega hedged.

There is an interesting connection between the dynamics of a model and the pricing of path-dependent derivatives. Because the price of path-dependent derivatives depends on conditional distributions, it depends on the Green's function $p(T, F_T; t', F_{t'})$ both in the forward and backward coordinates. This is in contrast to vanilla options that only depend on the backward coordinates for t' equal to today's date, with the dependence only important for hedging and not for pricing. As the dynamics of a model are given by the dependence on the backward variables, we see that the choice of dynamics affects both the pricing and hedging of path-dependent derivatives.

In reality, the most popular skew and smile models have dynamics that are in disagreement with the observed market behavior. Fortunately, the incorrect dynamics are usually only of minor importance for vanillas as we vega hedge them or impose the dynamics externally. For path-dependent derivatives, on the other hand, there is no such simple solution and we are stuck with incorrect prices unless we find a model with appropriate dynamics. If imposing the dynamics for vanillas, but not for path-dependent derivatives (as this is not possible), the outcome is two models that agree on the price but not on the hedge.

We round off by discussing the implementation of the risk computations that for advanced models has to be done via a bump-and-revalue approach. If it takes as long time to do the revaluation as it does to do the pricing, the risk computation is slower than the pricing by a factor given by the number of calibration instruments. The performance bottleneck in a derivatives pricing system is for this reason often found in the risk computation. Fortunately, there are several possible ways to speed up the computation of the risk and we now describe some of the more efficient techniques.

First of all, the bumping of the prices (or implied volatilities) of the calibration instruments can be done independently. Thus, the performance can be improved by using parallel computing. Secondly, as mentioned earlier, the risk does not have to be computed for the individual products but can be done on a portfolio level.

The performance can be improved substantially in the instance when the pricer depends on a low-performing calibration. Indeed, evaluating the risk amounts to computing

$$\frac{dV}{d\tilde{\sigma}_k} = \sum_j \frac{dV}{d\sigma_j}\frac{d\sigma_j}{d\tilde{\sigma}_k}$$

where the first factor is fast to evaluate as it measures how much the price varies with the model parameters and is therefore independent of the calibrations. The second factor, on the other hand, is slow as it requires a calibration for each calibration instrument. Fortunately, in the common situation when there are as many calibration instruments as there are model variables, the second factor is the inverse of the matrix $\left\{\frac{d\tilde{\sigma}_k}{d\sigma_j}\right\}$. This matrix only involves the pricing of the calibration instruments and does not depend on the calibration. The computation of the risk can therefore be done efficiently via the formula

$$\frac{dV}{d\tilde{\sigma}_k} = \sum_j \frac{dV}{d\sigma_j}\left\{\frac{d\tilde{\sigma}}{d\sigma}\right\}^{-1}_{jk}$$

4.4 Model Limitations

Derivatives models can fail to produce accurate results for various reasons. We go through some of the most important cases and discuss how they can be improved.

We start by looking at some situations where the method of dynamic replication breaks down due to unrealistic hedges. The standard example is that of a digital option paying $\theta(S_T - K)$ at maturity. The present value can be computed with the traditional lognormal Black–Scholes framework or with static replication to account for the skew and smile. In the Black–Scholes model, a digital option is worth $-\frac{dV}{dK} = N(d_-)$, where V is the corresponding call option. The Black–Scholes delta is then given by

$$\frac{dN(d_-)}{dS} = \frac{1}{S\sigma\sqrt{T-t}}n(d_-)$$

Observe that for ATM options, $K = S$, the delta tends to infinity when $t \to T$. It means that ATM options with short maturities need a large number of underlyings for the hedge. This hedging strategy is associated with a large gamma risk as it in practice only can be done discretely in time.

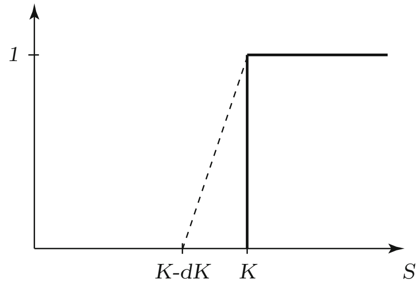

Fig. 4.4 Pricing a digital call option with a call spread

A standard technique for pricing a product for which dynamic replication fails is to find another product, worth a little bit more, but with a well-behaved hedge. Assume, for example, that a client wants to buy a digital call option. We do not want to charge the Black–Scholes price as this means that we take on the problem with the ill-behaved hedge for free. Instead, we purchase $\frac{1}{dK}$ call options with strike $K - dK$ and sell $\frac{1}{dK}$ call options with strike K and charge the client the resulting price, see Fig. 4.4. If the underlying is below $K - dK$ at maturity, no payments occur as we make no money on the call options and do not have to pay the client anything for the digital call. If the underlying is above K at maturity, we make $\frac{1}{dK}((S - K + dK) - (S - K)) = 1$ on the call options which is just the amount needed to be paid to the client. Finally, if the underlying is between $K - dK$ and K at maturity, we make a profit $\frac{1}{dK}(S - K + dK)$ on the call options while no payment need to be made to the client. We have therefore charged the client the price for a call spread that has equal or greater value than the digital option in all scenarios.

As dK is typically small, it is not possible to find liquid call options in the market that have such a small difference in strike. It means that the call spread has to be dynamically replicated. But as long as dK is not too small, the call spread has a well-behaved hedge, which means that the dynamic replication approach works.

The only non-trivial in the strategy is the choice of dK. In the case $dK \to 0$, we take on the problem with the ill-defined hedge for free, while in the case of a large dK, the client gets overcharged and rather deals with someone else. The delicate value of dK must therefore be such that we are satisfied with the hedge at the same time as the client finds the price attractive.

For another example of ill-behaved hedges in dynamic replication, consider a knock-out option of call type with an upper barrier B. If the underlying is far away from the barrier, the price of the option increases when the underlying increases. When the value of the underlying comes close enough to the barrier, the price starts to decrease and tends to zero, see Fig. 4.5. When pricing the option with, for example, a lognormal model for the underlying, the price curve is steep close to the barrier. The large negative value of the delta leads to hedging problems.

In the same way as for digital options, the problem is usually solved by finding a product with a more conservative price and better behaved greeks, typically an option with a slightly higher barrier $B + dB$. We charge the client the price of the higher barrier and hedge the option accordingly. If the underlying never hits the

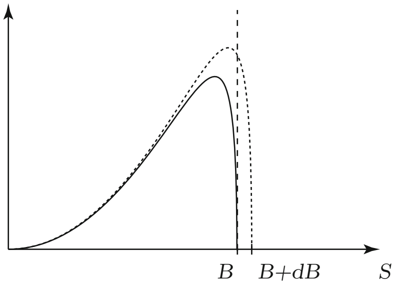

Fig. 4.5 Pricing a knock-out option with a barrier shift

barrier, the final payoff of the hedging strategy has the same value as the payment required by the client. If the underlying hits the barrier, on the other hand, we terminate the positively valued hedge while no payment is needed to the client. For *dB* not too small, the greeks behave well as long as the underlying is below B.

An additional example of derivative models having problems in practice is when they include parameters that cannot be hedged by liquid products. A correlation is a typical example of a parameter for which there are limited hedging possibilities. For a bank it is therefore important to sell products that are both long and short correlation to avoid a naked exposure to this type of risk. This has not been the situation traditionally and banks have sold long correlation products such as options with and without barriers on baskets of equity shares or products depending jointly on credit defaults. In times of uncertainty, equity correlations increase, which results in losses for the banks. An example of popular equity products with short correlation is given by payoffs that depend positively on the performance of the best (or worst) share(s) in a basket.

It is not always obvious whether the parameters in a model can be hedged with liquid instruments. Consider, for example, the situation when a bank sells a derivative to a client. To minimize the exposure to market movements, the bank purchases the hedging instruments and constructs a replicating portfolio. However, when markets go through a crisis, there is a general escape to products that are considered safe. The consequence is that the liquidity disappears for certain products, resulting in large bid-offer spreads. If this happens to the hedging instruments, it may no longer be possible to hedge the derivative effectively meaning that the issuer is exposed to the risk of unfavorable market moves.

Another property to account for is the feedback effect the hedge can have on the price: a price move can lead to hedges of the market participants which moves the price even more which results in new hedges, and so on. The feedback takes place until the market has established itself at a new equilibrium. This effect appears when the positions in the market are large compared to the liquidity or when the products involved have high leverage.

An example of the feedback effect can be seen in the events taking place following the unexpected announcement on June 5, 2008 that the European Central Bank was likely to raise the interest rate at its next meeting. This led to an increase of the short rates in a yield curve that was already relatively flat. As investment banks had sold exotic products containing zero-strike digital options on calendar spreads of the yield curve, they all had to hedge themselves simultaneously. The result was a fast, and dramatic, inversion of the euro interest rate curve. In this situation, all investment banks had to hedge themselves in a similar way, which is an example of how liquidity can disappear from the market. The result was substantial losses for the European parts of the investment banks.

Quantitative analysts face two conflicting requirements on the models they construct: they should be simple enough for practical use but complex enough to account for realistic market behavior. The first requirement is obviously more important and simplifying assumptions such as those in Sect. 1.2 are therefore made. It is then left to the traders to account for, and charge extra premium for, the inaccuracies in the models. Unfortunately, the competition between traders at different banks is sometimes so strong that the charged price does not cover the risk.

To give an example of a simplifying assumption that has not been properly accounted for historically, consider the situation when a model contains the problem of merging marginal distributions to a joint distribution. This can be solved with the Gaussian copula, which only uses the correlation as an input, see Sect. 10.1 for details. Unfortunately, after calibrating the correlation to normal market scenarios, the correlation between extreme events is too weak. The solution to this problem would require a copula with at least one additional parameter, which cannot be accurately calibrated as extreme events are rare. Banks therefore took the simple approach and traditionally used the Gaussian copula. Another reason for its popularity was that the merging of distributions often was only one of many problems to be solved in a complex model, thereby making it hard to motivate anything but the simplest non-trivial copula. The result of the simplification was that banks significantly mispriced the impact of extreme scenarios.

The most important simplifying assumption of Sect. 1.2 that needs to be accounted for in practice is arguably that of negligible credit exposure. For instance, when trading a product that involves future cash flows from a client, it is necessary to include the counterparty credit exposure in the price. Although the adjustment in principle can be done ad hoc by the traders, the responsibility has lately shifted to models developed by quants. It is then possible to quantify and hedge the credit risk. Investment banks nowadays have many quantitative analysts working on *credit value adjustment (CVA)*. CVA is the value that needs to be added to the price of a product to account for the risk of the counterparty to default.

Let $e(t)$ be the expected loss of a contract conditional on that a counterparty default happens at t. We assume $e(t)$ to be discounted to today and contain the factor 1-R, where R is the recovery rate. The CVA for a product with the last cash flow taking place at T can then be written as

$$\int_0^T e(t)p(t)dt$$

where $p(t)$ is the risk-neutral probability density function that the counterparty defaults at t. For major counterparties, $p(t)$ can be obtained from quotes on *credit default swaps (CDSs)*.

The expected loss is in general correlated with the probability of the counterparty defaulting. The risk is said to be in the *wrong way* if $e(t)$ is large when the probability of default is high. The reverse situation is known as *right way risk*.

The simplest possible CVA computation is done by assuming the expected loss to be independent of the probability of default. $e(t)$ is then equal to $(1 - R)U(t)_+$ where U is the discounted contract value. The computation becomes particularly simple when the contract value always is positive, for instance, for a bond. The outcome of the CVA is then nothing more than adding an appropriate spread on top of the discount curve.

When U can be positive as well as negative, such as for a swap, the pricing is more complex. Indeed, the CVA then looks like an option payout. Thus, CVA introduces a volatility-dependence even in the situation when the contract itself is volatility independent. Because of the complexities, CVA often has to be priced with numerical methods such as Monte Carlo simulations.

There are often netting agreements in place regarding credit defaults. It means that if $\{U_i\}_i$ are parts of such an agreement, the loss on default is $(1 - R)(\sum_i U_i)_+$, which is smaller than the un-netted amount $(1 - R) \sum_i (U_i)_+$. The complexity with netting agreements is that it is no longer sufficient to compute CVA for individual contracts, but all contracts belonging to the same netting agreement need to be accounted for.

It is also common to include the effect of own default. The resulting price impact is referred to as *debt value adjustment (DVA)*. It can be accounted for in the same way as for CVA. For instance, consider a simulation of the contract value $U(t)$ and of the default probabilities for us and our counterparty. Should the counterparty default first, we consider $U(t)_+$ while if we default first, $U(t)_-$ should be used. One of the main problems with DVA is that hedging it involves trading CDSs on ourselves.

Another example where the credit component affects the pricing regards the traded product itself. Consider, for instance, a product that depends on the performance of a basket of equities. Account should then be taken of the fact that any of the shares in the basket can default. In fact, it does not have to be as dramatic as a default; even a trade halt in one of the shares is sufficient to have an impact on derivatives. For example, the suspension of trades in Fortis in 2008 had a significant impact on barrier options on Euro Stoxx 50.

4.5 Testing

When a model has been implemented, it needs to go through rigorous testing. The goal of the testing is to verify that the model:

- Gives reasonable prices and risk values
- Gives correct prices and risk values in special cases (for example, when the volatility is zero or when the calibration instruments are priced)

- Is in agreement with other models that can be used for (a subset of) the supported products
- Is arbitrage free
- Has appropriate dynamics

In this section we show that there exist several different testing techniques and that a combination of them is necessary for completeness.

There are three components that can cause unwanted behavior in an implementation of a model. They originate in the model itself, the method by which the model is solved, or in the implementation. A typical error from the model is the existence of arbitrage strategies. For an example of a method error, consider the pricing of a European option using a Crank-Nicholson PDE solver. It is well known that the price has an oscillatory dependence on the spot value when the option has a short maturity and is close to ATM. In this case there is nothing wrong with the model (the PDE), but the problem is instead that the PDE solver performs badly when the payoff has a discontinuous derivative (which can be taken care of by smoothing out the payoff, for example, by using an implicit PDE solver near the maturity date). Finally, by errors in the implementation we mean bugs in the computer code.

The first test a model should go through is a theoretical investigation, i.e. an analysis with a pen and paper. It is then often possible to verify if a model is arbitrage free, has correct asymptotics and behaves well qualitatively. This testing is also necessary for a proper understanding of the model. It is not, however, sufficient to meet all the above conditions.

A model can only be thoroughly tested once it has been implemented. As it is the implementation that is used by the traders and the risk managers, this is where the testing becomes particularly important. The disadvantage of such a test is that it can be hard to determine if a problem has its roots in the model, the method of solving the model or if it comes from a bug. Many times, the origin of a problem can be determined by varying the model parameters (e.g. the shift parameter in a shifted lognormal model) or the parameters of the solution method (e.g. the number of paths in a Monte Carlo simulation). A theoretical analysis of a model can also help us to find the origin of a problem as it guides us to instances for which the behavior of the model is known.

We now discuss in more detail how the implementation of a model can be tested. The initial test is to verify that the implementation is stable and has proper error handling. This can be done by using unexpected arguments. For instance, in the field expecting a strike value for a European option, a string "test" can be inserted. The system should then not crash but instead return an appropriate error message such as: "Please insert a numerical strike value". Other tests of this type are the use of a negative number of paths in a Monte Carlo simulator, a correlation value outside the interval $[-1, 1]$, or a negative or zero FX rate.

The second step is to test the model behavior when the arguments are strange, but anyway acceptable, e.g. a negative strike for a European call option. From one point of view, this is unnatural and an error is expected. On the other hand, viewing a call option as a contract with payoff $(S - K)_+$, there is nothing that prevents the strike from becoming negative. In fact, permitting negative strikes allows us to price a larger set of products. Thus, in this situation it is not clear whether an error should be thrown or not. Another example concerns the value of the volatility. One philosophy is to limit volatilities to the range $[0, 1]$ to avoid the user inserting a volatility of five, for example, mistakenly interpreting it as $5\% = 0.05$. By making such restrictions, we might not support certain markets such as the electricity market for which the spot volatility can be several hundred percent, which leads to high implied volatilities. For yet another example, assume that 1435.7 number of simulations is inserted in a Monte Carlo simulation. One solution is to interpret this as 1436 simulations by rounding upwards. However, it seems to be a plausible explanation that the user made a typo and possibly meant 14357 simulations, or something else. It therefore appears more natural to throw an error here.

The dependency on the input values should also be tested to ensure that sensible values (i.e. not an error) are returned for special cases, for example, when the strike or the volatility is zero. Consider, for instance, the Black–Scholes formula which contains a division with the volatility, meaning that the zero-volatility instance has to be implemented in a special way.

It is necessary to test whether the result is independent of how and when the pricing is done. For instance, if we first price a specific product, then a couple of intermediate products and thereafter the original product again, the same result should obviously be returned, assuming that the market data remains unchanged. This type of errors occurs relatively often because some variable in the computer code has not been reset.

The testing described so far relates only to the functionality of the implementation but not to the correctness of the returned values: the present value, the greeks and other output variables. These values are non-trivial to validate as the answer is not known in general. For instance, the price is model dependent when obtained from dynamic replication, meaning that there is no such thing as a true price. For this reason, much of the testing has to be limited to special cases for which the correct price and/or risk values are known. For example, the tests can be on special products types such as the calibration instruments or zero strike products. Another approach is to let the model parameters take extreme values. For instance, setting the volatility to zero leads to a known price. Observe that it is a good idea to let the parameters be very small rather than exactly equal to zero as these limiting cases can be implemented differently.

Despite the difficulty of testing a model in non-special cases, some limited testing can anyway be done. First of all, the model should be tested to be free of arbitrage. For instance, a callable deal should always be verified to be worth more that the corresponding non-callable deal. It is also a good idea to verify that the parity relations are fulfilled. Secondly, when implementing a new model, there often already exists a model in the pricing software that can be used for at least a subset of the intended products to be priced. The two models can then be compared on this subset to verify that they do not deviate more than expected. If there is not an already existing model, it is a good idea to invent an alternative model (preferably not by the same person that came up with the original model to minimize the risk of the same mistake occuring twice) or choose it from the literature. When testing a new method of solving an existing model, it is obviously possible to use the full set of products for the test. When it comes to the testing of the implementation, it is useful to have an alternative implementation of the model, for example, in a spreadsheet, preferably by someone other than the creator of the original code.

When testing two different models, methods or implementations, a suitable set of test cases needs to be chosen. The test cases consist of a choice of products, market data, model parameters and method parameters. It is common to use (a subset of) the deals already booked as test cases. Additionally, it is useful to include manufactured test cases designed to target suspected weaknesses in the model, the solution method and the implementation. Furthermore, randomly generated test cases are useful in catching unexpected errors.

We have so far discussed tests that can be performed to validate a new model or method. This testing is done at only one instance in time. Apart from this, there are tests that need to be run each time there is a change in the implementation of a model. These recurring tests are simpler as they can be done by comparing the new version with the older version. Therefore, the focus is mainly on the choice of appropriate test cases. Typically, the test cases are chosen in a similar way as for the initial testing, i.e. they are a combination of currently booked deals, manufactured deals and randomly generated deals. It is useful to have the same set of test cases throughout the lifetime of the implementation as the changes in the PV and the risk values then can be tracked through the different versions of the model. When the computer code has been modified, it should be verified that only the deals that were expected to change values did so, and no other. It is also necessary to investigate if the changes were as anticipated.

Observe that for financial pricing models it is often required that all test cases succeed and all PV and risk value differences from one version to the next are understood. This is in contrast to many other industries, such as IT, where only a subset (for example, 99%) of the test cases have to succeed. This is understandable as the impact of a bug in a pricing model can lead to a substantial loss while a bug in a mobile telephone, for instance, can often be resolved by turning the device off and then on again.

Apart from testing the PV and the greeks separately, it is also important to verify that these numbers are consistent with each other, i.e. whether the greeks accurately describe the changes in the price. This can be done by letting the model undergo a hedging simulation. In such a simulation we use a data set that describes how the market can change from one day to the next (or whichever hedging time interval that is used). The data set can be obtained either from a historical database or by generating it through a model. For each scenario of the simulation, the profit or loss from being short the derivative and long the hedge (or vice versa) is recorded. By aggregating the results for all simulation paths, the correctness of the model can be measured. It is also possible to investigate how the product together with the hedge behaves under large moves of the market data.

The hedging simulation is typically done between two dates and not over a succession of dates covering a longer time period such as a month. The reason is that the hedging instruments are typically defined in terms of time periods from the trade date such as 3M or 1Y, and are therefore different on the various simulation dates. This linear explosion of hedging instruments often complicates the implementation of a hedging simulation unless restricting to a single simulation step. An exception to the rule is the analysis of a static hedge for which the hedging is only done at the first date. Another exception is when the hedge is done by futures contracts or by options on futures, which is typically the situation for commodities markets, for which the maturity is defined as a specific date or month and not as a time period from the trade date.

To understand the precise meaning of a hedging simulation, let us restrict ourselves to the situation when the calibration instruments are independent and equal to the hedging instruments, $\mathcal{C}_j = h_j$, meaning that the weights of the replicating portfolio are given by $b_j = \frac{dV}{dh_j}$. The hedging portfolio $\sum_j \frac{dV}{dh_j} h_j$ then changes by $\sum_j \frac{dV}{dh_j} dh_j$ when h_j changes by dh_j. The change in h_j, in turn, comes from a change $d\tilde{\sigma_j}$ in the volatility, assuming for simplicity that the underlying is constant so that only vega risk is present. The change in the hedging portfolio can be written as

$$\sum_j \frac{dV}{dh_j} \frac{dh_j}{d\tilde{\sigma_j}} d\tilde{\sigma_j} = \sum_j \frac{dV}{d\tilde{\sigma_j}} d\tilde{\sigma_j}$$

A hedging simulation therefore amounts to comparing the price difference

$$V^{\text{model}} \left(\{\tilde{\sigma_j} + d\tilde{\sigma_j}\}\right) - V^{\text{model}} \left(\{\tilde{\sigma_j}\}\right)$$

with the change

$$\sum_j \left(\frac{dV}{d\tilde{\sigma_j}}\right)^{\text{model}} d\tilde{\sigma_j}$$

of the replicating portfolio. If market quotes exist for V, another useful test is obtained by replacing $V^{\text{model}}(\cdot)$ with $V^{\text{market}}(\cdot)$ in the above difference. In general, the change in the hedging portfolio is different from the change in V and there are a couple of reasons for this behavior. One explanation is that the market move is so large that the lowest-order hedge term does not suffice to explain the price change. In this situation the hedging should ideally have been done by including higher-order terms such as the volga. Another reason can be that the numerical computations of the risk are unstable and give a poor value of $\frac{dV}{d\tilde{\sigma_j}}$. The difference can also depend on a poor choice of model that does not correctly take into account the moves in the market, e.g. the move in $V^{\text{market}}(\cdot)$ is not only described by the values of $\{\tilde{\sigma_j}\}$, but also by other variables.

The hedging simulation has the advantage that it is a good reflection of the reality. Unfortunately, the reality is usually quite complex, meaning that it can be complicated to analyze any peculiarities or errors that show up in the testing. To obtain a more tractable simulation, recall that V depends on the market quotes $\{\tilde{\sigma_j}\}$ via the model variables $\{\sigma_i\}$. The change in the replicating portfolio can therefore be written as

$$\sum_j \frac{dV}{d\tilde{\sigma_j}} d\tilde{\sigma_j} = \sum_{ij} \frac{dV}{d\sigma_i} \frac{d\sigma_i}{d\tilde{\sigma_j}} d\tilde{\sigma_j} = \sum_i \frac{dV}{d\sigma_i} d\sigma_i$$

assuming that there is a bijective relation between $\{\tilde{\sigma_j}\}$ and $\{\sigma_i\}$. We can therefore compare

$$V^{\text{model}}(\{\sigma_i + d\sigma_i\}) - V^{\text{model}}(\{\sigma_i\})$$

with

$$\sum_i \left(\frac{dV}{d\sigma_i}\right)^{\text{model}} d\sigma_i$$

This is similar to the hedging simulation with the difference that the dependence on the matrix $\left\{\frac{d\sigma_i}{d\tilde{\sigma_j}}\right\}$, relating the model to the calibration, has been eliminated. This test is therefore not as comprehensive as the hedging simulation, but it is anyway useful as it isolates the test to only depend on the pricing and not the calibration. Thus, if this test behaves well while there are problems with the hedging simulation, the error must lie in the calibration.

Even when a model has gone through thorough testing, it might still happen that it fails when in use. The worst type of failure is arguably when an incorrect price is returned that is so close to the correct price that the trader does not realize that something is wrong. The cause of this can be the model, the solution method or the implementation. An example is when the method for determining the early exercise boundary (when using a Monte Carlo simulator) does not work properly for certain products. To prevent such mispricing, it is useful to have at least one more model at the trading desk to verify the sanity of the result. Should the prices of the models differ too much, it is necessary to investigate the reason and determine which model (if any) that returns an appropriate price.

Bibliography

Ekstrand C (2010) Calibrating exotic models to vanilla models. WILMOTT Magazine April: 109–116

Giles M, Glassermann P (2006) Smoking adjoints: fast monte carlo greeks. Risk January:92–96

Hagan P (2004) Adjusters: turning good prices into great prices. The best of WILMOTT 1. Wiley, England

Piterbarg V (2006) Markovian projection method for volatility calibration. Social Science Research Network. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=906473. Accessed 16 May 2011

Piterbarg V, Renedo M (2004) Eurodollar futures convexity adjustments in stochastic volatility models. Social Science Research Network. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=610223. Accessed 16 May 2011

Part II
Skew and Smile Techniques

Chapter 5
Continuous Stochastic Processes

In Chap. 3 we saw that stochastic differential equations play an important role in the pricing of derivatives. We here discuss various types of SDEs that can be used for the underlying process. We focus on SDEs that have simple solutions and thereby allow for an efficient implementation. As the drift term does not enter the pricing formula for derivatives, we often assume it to be zero. The European call option price formula is derived for the driftless SDEs. This is useful not only as call options are important by themselves, but also because they through static replication can be used to price any other fixed-time payment. Furthermore, exotics models are most often calibrated to European call options.

We consider equations of the type

$$F_t = \int_0^t \mu(u, F_u)du + \int_0^t \sigma(u, F_u)dW_u$$

where an integral with respect to a Brownian motion can be defined from limits of integrands that are piece-wise constant functions of t, see Appendix. It is common to write SDEs in the equivalent differential form:

$$dF_t = \mu(t, F_t)dt + \sigma(t, F_t)dW_t$$

We have chosen to only discuss SDEs for which the stochastic part is represented by a Brownian motion. The more general case of jump processes is the topic of Chap. 8. Furthermore, we are only concerned with SDEs that admit strong solutions, i.e. that are adapted to the filtration of the Brownian motion, see the Appendix for the definition of these concepts. It can be shown that under certain growth restrictions for the functions μ and σ, a strong solution exists and is unique.

We start with an introduction to the linear SDE and the special cases of the normal SDE, the lognormal SDE, the shifted lognormal SDE and the Ornstein-Uhlenbeck SDE. We also cover the quadratic SDE, the Brownian bridge, the Bessel process and the closely related CEV process. Finally, we describe how certain non-analytic SDEs can be used for derivatives pricing.

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_5,

5.1 The Linear SDE

The *linear SDE*

$$dF_t = (\mu_1(t) + \mu_2(t)F_t)dt + (\sigma_1(t) + \sigma_2(t)F_t)dW_t$$

can be solved by first solving the corresponding geometric process with Ito's lemma:

$$\begin{aligned} d\tilde{F}_t &= \mu_2\tilde{F}_t dt + \sigma_2\tilde{F}_t dW_t \\ \Rightarrow d\ln\tilde{F}_t &= \tilde{F}_t^{-1}d\tilde{F}_t - \frac{1}{2}\tilde{F}_t^{-2}(d\tilde{F}_t)^2 = \mu_2 dt + \sigma_2 dW_t - \frac{1}{2}\sigma_2^2 dt \\ \Rightarrow \tilde{F}_T &= \tilde{F}_0 \exp\left(\int_0^T (\mu_2 - \frac{1}{2}\sigma_2^2)dt + \int_0^T \sigma_2 dW_t\right) \end{aligned}$$

With $d\tilde{F}_t^{-1} = (-\mu_2 + \sigma_2^2)\tilde{F}_t^{-1}dt - \sigma_2\tilde{F}_t^{-1}dW_t$ we obtain the solution

$$\begin{aligned} d(F_t\tilde{F}_t^{-1}) &= (dF_t)\tilde{F}_t^{-1} + F_t d\tilde{F}_t^{-1} + (dF_t)d\tilde{F}_t^{-1} \\ &= (\mu_1 + \mu_2 F_t)\tilde{F}_t^{-1}dt + (\sigma_1 + \sigma_2 F_t)\tilde{F}_t^{-1}dW_t + F_t(-\mu_2 + \sigma_2^2)\tilde{F}_t^{-1}dt \\ &\quad -F_t\sigma_2\tilde{F}_t^{-1}dW_t - (\sigma_1 + \sigma_2 F_t)\sigma_2\tilde{F}_t^{-1}dt \\ &= (\mu_1 - \sigma_1\sigma_2)\tilde{F}_t^{-1}dt + \sigma_1\tilde{F}_t^{-1}dW_t \\ \Rightarrow F_T &= \tilde{F}_T\left(F_0\tilde{F}_0^{-1} + \int_0^T (\mu_1 - \sigma_1\sigma_2)\tilde{F}_t^{-1}dt + \int_0^T \sigma_1\tilde{F}_t^{-1}dW_t\right) \end{aligned}$$

Important special cases of the linear SDE are the lognormal SDE, the normal SDE, the shifted lognormal SDE and the Ornstein-Uhlenbeck process, all to be covered in the following sections.

5.2 The Lognormal SDE

We saw in the previous section that the *lognormal SDE*

$$dFt = \sigma F_t dW_t$$

has the solution

$$F_T = F_0 \exp\left(-\frac{1}{2}\int_0^T \sigma^2 dt + \int_0^T \sigma dW_t\right)$$

The integral $\int_0^t \sigma dW_t$ is normally distributed with mean 0 and variance $\bar{\sigma}^2 T := \int_0^T \sigma^2 dt$, see the Appendix. It follows that

$$F_T = F_0 e^{-\bar{\sigma}^2 T/2 + \bar{\sigma}\sqrt{T} X}$$

for X a $\mathcal{N}(0,1)$ variable, i.e. $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. If the volatility had been constant, we would have obtained a similar expression:

$$F_T = F_0 e^{-\sigma^2 T/2 + \sigma\sqrt{T} X}$$

We therefore see that it is possible to assume, without loss of generality, that the volatility is constant because if this is not the case, we can simply replace the variance with the integrated variance.

We compute the Green's function $p_F(T, F_T, F_0)$, where we by abuse of notation have used F_T to denote the distribution of F at T as well as the value of F at T. Introducing the variable x by

$$F_T = F_0 e^{-\sigma^2 T/2 + \sigma\sqrt{T} x} \Leftrightarrow x = \frac{\ln(F_T/F_0)}{\sigma\sqrt{T}} + \frac{1}{2}\sigma\sqrt{T}$$

it follows that

$$\begin{aligned} & p_F(T, F_T, F_0) dF_T = p_X(x) dx \\ \Leftrightarrow\ & p_F(T, F_T, F_0) = p_X(x) / \frac{dF_T}{dx} = \frac{1}{\sqrt{2\pi\sigma^2 T} F_T} e^{-\left(\frac{\ln(F_T/F_0)}{\sigma\sqrt{T}} + \frac{1}{2}\sigma\sqrt{T}\right)^2/2} \end{aligned}$$

The forward European call option price is equal to

$$\begin{aligned} E[(F_T - K)_+] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(F_0 e^{-\sigma^2 T/2 + \sigma\sqrt{T} x} - K\right)_+ e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{(\ln(K/F_0) + \sigma^2 T/2)/\sigma\sqrt{T}}^{\infty} \left(F_0 e^{-(x - \sigma\sqrt{T})^2/2} - K e^{-x^2/2}\right) dx \\ &= F_0 N(d_+) - K N(d_-), \quad d_\pm = \frac{\ln(F_0/K)}{\sigma\sqrt{T}} \pm \frac{1}{2}\sigma\sqrt{T} \end{aligned}$$

The price of an ATM option, $K = F_0$, is for small $\sigma\sqrt{T}$ given by

$$\begin{aligned} E[(F_T - K)_+]|_{K=F_0} &= F_0 \left(N\left(\frac{1}{2}\sigma\sqrt{T}\right) - N\left(-\frac{1}{2}\sigma\sqrt{T}\right)\right) \\ &\approx F_0 \sigma\sqrt{T} \frac{d}{dx} N(x)|_{x=0} = \frac{1}{\sqrt{2\pi}} \sigma\sqrt{T} F_0 \end{aligned}$$

5.3 The Normal SDE

The *normal SDE*

$$dF_t = \sigma dW_t$$

is solved by

$$F_T = F_0 + \int_0^T \sigma dW_t = F_0 + \sigma\sqrt{T}X, \quad X \sim \mathcal{N}(0,1)$$

The Green's function

$$p_F(T, F_T, F_0) = p_X(x)/\frac{dF_T}{dx} = \frac{1}{\sqrt{2\pi\sigma^2 T}}e^{-(F_T - F_0)^2/2\sigma^2 T}$$

is fundamentally different from the lognormal version as it supports negative values for F_T. Some financial processes can be negative and for these it is certainly possible to use this type of SDE. For other financial variables, such as FX rates or equity stocks, the positivity of a process can be a necessity. One way to avoid negative values is to impose *boundary conditions* at $F = 0$. The most commonly used boundary conditions are the *absorbing* (Dirichlet) and the *reflecting* (Neumann) conditions. The reflecting boundary condition states that if the process hits 0 then it bounces back out again along the positive axis, just as a ball would bounce against a wall. The absorbing boundary condition states that if the process hits 0 then it gets stuck there forever, just as a ball thrown at a wall covered with glue would stick.

The boundary conditions can be represented mathematically by the *method of images*. It means that we use a mirror process that has the same probability to be located at $-F$ as the original process has to be located at F. In the instance of an absorbing condition, the image process is given a negative value, which means that it annihilates the original process if they hit each other (which happens at $F = 0$). For the reflecting condition, the processes have the same sign which can be interpreted as if they bounce off each other at $F = 0$. The Green's function therefore takes the form:

$$p_F(T, F_T, F_0) = \frac{1}{\sqrt{2\pi\sigma^2 T}}e^{-(F_T - F_0)^2/2\sigma^2 T} + \eta\frac{1}{\sqrt{2\pi\sigma^2 T}}e^{-(F_T + F_0)^2/2\sigma^2 T}$$

where

$$\eta = \begin{cases} 0 & \text{No boundary condition} \\ -1 & \text{Absorbing boundary condition} \\ 1 & \text{Reflective boundary condition} \end{cases}$$

For a more detailed discussion of the method of images, consult any standard textbook on PDEs.

The choice of boundary condition depends on the type of problem under consideration. We have already seen that the disadvantage of having no boundary conditions is that the underlying can be negative. The probability of this happening is

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\sigma^2 T}} e^{-(F_T-F_0)^2/2\sigma^2 T} dF_T = N(-F_0/\sigma\sqrt{T})$$

With absorbing boundary conditions the possibility of a negative F is avoided but instead there is a finite probability that $F_T = 0$:

$$\begin{aligned} P(F_T = 0) &= 1 - P(F_T > 0) \\ &= 1 - \int_0^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2 T}} e^{-(F_T-F_0)^2/2\sigma^2 T} - \frac{1}{\sqrt{2\pi\sigma^2 T}} e^{-(F_T+F_0)^2/2\sigma^2 T} \right) dF_T \\ &= 1 - N(F_0/\sigma\sqrt{T}) + N(-F_0/\sigma\sqrt{T}) = 2N(-F_0/\sigma\sqrt{T}) \end{aligned}$$

A finite probability for the underlying to be equal to zero can also ruin the model. For example, consider the situation when F represents an FX rate and we need to convert a currency via multiplication by F^{-1}. This problem can be partially solved by moving the boundary to a small positive value $F = \epsilon > 0$.

The probability of ending up at $F = 0$ is small for short maturities, meaning that a boundary condition only has a minor effect. For large maturities, on the other hand, the contribution can be important with a mispricing of far out-of-the-money put options (or equivalently, by put-call parity, far in-the-money call options) as a consequence.

A reflecting boundary avoids negative values as well as assigning a zero probability to the event $F = 0$ (but the PDF is not equal to zero at this point). Unfortunately, this boundary condition does often not have a realistic financial interpretation. For example, when F models an equity forward, it is counterintuitive that F bounces back up along the positive axis after it has hit the zero point. In this instance the absorbing condition is more natural as if $F = 0$ is hit, the equity forward stays there forever.

We conclude that if positivity is essential, the introduction of boundary conditions is often not sufficient to rescue a model for long or intermediate maturities. It is usually better to change the underlying process so that only positive values are supported. Examples of such processes are given later in this chapter.

The inclusion of boundary conditions is not limited to the normal SDE but can be used for any of the processes in this chapter. We have chosen to have the discussion here as boundary conditions are commonly applied to the normal SDE.

In the absence of boundary conditions, the European call option price is equal to

$$\begin{aligned} E[(F_T - K)_+] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(F_0 + \sigma\sqrt{T}x - K \right)_+ e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{(K-F_0)/\sigma\sqrt{T}}^{\infty} \left((F_0 - K)e^{-x^2/2} - \sigma\sqrt{T}\frac{d}{dx}e^{-x^2/2} \right) dx \\ &= (F_0 - K)N(d_0) + \sigma\sqrt{T}n(d_0), \quad d_0 = \frac{F_0 - K}{\sigma\sqrt{T}} \end{aligned}$$

From the form of the Green's function with boundary conditions, we conclude that the general solution is obtained by adding a term with F_0 replaced by $-F_0$, multiplied by η:

$$\begin{aligned}
E[(F_T - K)_+] &= (F_0 - K)N(d_0) + \sigma\sqrt{T}n(d_0) \\
&\quad + \eta(-F_0 - K)N(d_0^*) + \eta\sigma\sqrt{T}n(d_0^*), \\
d_0 &= \frac{F_0 - K}{\sigma\sqrt{T}}, \quad d_0^* = \frac{-F_0 - K}{\sigma\sqrt{T}}
\end{aligned}$$

Just as for the lognormal SDE, a time-dependent volatility is handled by replacing $\sigma^2 T$ with the integrated variance $\int_0^T \sigma^2 dt$. We end this section with the simple expression obtained for ATM options in the absence of boundary conditions:

$$E[(F_T - K)_+]|_{K=F_0} = \frac{1}{\sqrt{2\pi}}\sigma\sqrt{T}$$

5.4 The Shifted Lognormal SDE

It is useful to have an SDE that interpolates (and extrapolates) between the lognormal SDE and the normal SDE. We would like this SDE to be such that the ATM volatility is independent of the interpolation parameter when $\sigma\sqrt{T}$ is small. As the normal model

$$dF_t = \sigma F_0 dW_t$$

has the same ATM volatility as the lognormal model for small $\sigma\sqrt{T}$, this suggests the form

$$dF_t = \sigma(F_0 + \beta(F_t - F_0))dW_t$$

which reduces to the lognormal SDE for $\beta = 1$ and the normal SDE for $\beta = 0$. Writing the SDE as

$$dF_t = \beta\sigma(F_t - \tilde{F})dW_t, \quad \tilde{F} = F_0\left(1 - \frac{1}{\beta}\right)$$

for $\beta \neq 0$, we conclude that $G = F - \tilde{F}$ satisfies

$$dG_t = dF_t = \beta\sigma(F_t - \tilde{F})dW_t = \beta\sigma G_t dW_t$$

i.e. G is lognormal. As F is a constant added to a lognormal process, the SDE followed by F is referred to as a *shifted lognormal SDE*. G_t is lognormal and therefore positive, which implies that $F_t > \tilde{F}$.

The solution for the lognormal process G gives us

$$F_T = \tilde{F} + \left(F_0 - \tilde{F}\right)e^{-\beta^2\sigma^2 T/2 + \beta\sigma\sqrt{T}X}, \quad X \sim \mathcal{N}(0,1)$$

From

$$
\begin{aligned}
x &= \frac{\ln\left(\left(F_T - \tilde{F}\right) / \left(F_0 - \tilde{F}\right)\right)}{\beta\sigma\sqrt{T}} + \frac{1}{2}\beta\sigma\sqrt{T} \\
\frac{dF_T}{dx} &= \beta\sigma\sqrt{T}\left(F_T - \tilde{F}\right)
\end{aligned}
$$

we obtain

$$
\begin{aligned}
p_F(T, F_T, F_0) &= p_X(x) / \frac{dF_T}{dx} \\
&= \frac{1}{\sqrt{2\pi\beta^2\sigma^2 T}\left(F_T - \tilde{F}\right)} e^{-\left(\frac{\ln\left(\left(F_T - \tilde{F}\right)/\left(F_0 - \tilde{F}\right)\right)}{\beta\sigma\sqrt{T}} + \frac{1}{2}\beta\sigma\sqrt{T}\right)^2/2}
\end{aligned}
$$

The call option price is

$$
\begin{aligned}
E[(F_T - K)_+] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\tilde{F} + \left(F_0 - \tilde{F}\right) e^{-\beta^2\sigma^2 T/2 + \beta\sigma\sqrt{T}x} - K\right)_+ e^{-x^2/2} dx \\
&= \begin{cases} \left(F_0 - \tilde{F}\right) N(d_+) - \left(K - \tilde{F}\right) N(d_-), & K > \tilde{F} \\ F_0 - K, & K \leq \tilde{F} \end{cases}
\end{aligned}
$$

$$
d_{\pm} = \frac{\ln\left(\frac{F_0 - \tilde{F}}{K - \tilde{F}}\right)}{\beta\sigma\sqrt{T}} \pm \frac{1}{2}\beta\sigma\sqrt{T}
$$

As usual, a time-dependent volatility is handled by replacing $\sigma^2 T$ with the integrated variance $\int_0^T \sigma^2 dt$.

We have considered two SDEs for option pricing earlier in this chapter: the lognormal SDE and the normal SDE. They have each one free parameter, σ, and can therefore only be calibrated to ATM options. The shifted lognormal SDE, on the other hand, has two free parameters and can be calibrated to both ATM options and the skew. In the next chapter we discuss in a more general setting how β determines the skew, but this behavior can also be verified directly from the above European call option formula. The skew in the shifted lognormal model is such that the implied volatility decreases with the strike for $\beta < 1$. This behavior is useful when modeling underlyings for which the volatility decreases with increasing F. This is typically the situation for equities and sometimes for interest rates and FX rates (as there is one currency on each side of an FX rate, half of the FX rates have skews in one direction and half in the other direction, unless there is no skew at all).

Observe that the ATM price without boundary condition and with small $\sigma\sqrt{T}$ is given by

$$
\begin{aligned}
E[(F_T - K)_+]|_{K=F_0} &= \frac{F_0}{\beta}\left(N\left(\frac{1}{2}\beta\sigma\sqrt{T}\right) - N\left(-\frac{1}{2}\beta\sigma\sqrt{T}\right)\right) \\
&\approx \frac{F_0}{\beta}\beta\sigma\sqrt{T}\frac{d}{dx}N(x)|_{x=0} = \frac{1}{\sqrt{2\pi}}\sigma\sqrt{T}F_0
\end{aligned}
$$

which indeed is independent of β. It can also be shown that σ only has a small effect on the skew (this, of course, depends on the exact definition of skew). As σ mainly affects the ATM price and β mainly affects the skew, the calibration is easy both to implement and to debug.

5.5 The Quadratic SDE

We here consider the *quadratic SDE*

$$
dF_t = \sigma(F_t - a)(1 - F_t/b)dW_t, \quad a < b
$$

The PDE for the corresponding call option problem is

$$
\begin{cases}
U_\tau &= \frac{1}{2}(F-a)^2(1-F/b)^2 U_{FF}, \quad \tau = \sigma^2(T-t) \\
U(\tau = 0) &= (F-K)_+
\end{cases}
$$

Following Ingersoll (1996), we use the transformation

$$
U(\tau, F) = \frac{(1-K/b)(1-F/b)}{1-a/b}\Psi(\tau, x), \quad x = \frac{F-a}{1-F/b}
$$

to obtain

$$
\begin{aligned}
U_F &= \frac{1-K/b}{1-a/b}\left(-\frac{1}{b}\Psi + (1-F/b)\Psi_x\frac{1-a/b}{(1-F/b)^2}\right) \\
\Rightarrow U_{FF} &= \left(-\frac{1-K/b}{b-a}\Psi_x + \frac{1-K/b}{1-F/b}\Psi_{xx}\right)\frac{1-a/b}{(1-F/b)^2} + \frac{1-K/b}{(1-F/b)^2}\frac{1}{b}\Psi_x \\
&= \frac{(1-K/b)(1-a/b)}{(1-F/b)^3}\Psi_{xx}
\end{aligned}
$$

The PDE takes the form

$$
\begin{aligned}
\frac{(1-K/b)(1-F/b)}{1-a/b}\Psi_\tau &= \frac{1}{2}(F-a)^2(1-F/b)^2\frac{(1-K/b)(1-a/b)}{(1-F/b)^3}\Psi_{xx} \\
\Leftrightarrow \Psi_\tau &= \frac{1}{2}(1-a/b)^2 x^2 \Psi_{xx}
\end{aligned}
$$

with initial condition

$$\begin{aligned}\Psi(\tau=0) &= \frac{1-a/b}{(1-K/b)(1-F/b)}(F-K)_+ \\ &= \left(\frac{(F-a)(1-K/b)-(1-F/b)(K-a)}{(1-K/b)(1-F/b)}\right)_+ = \left(x-\frac{K-a}{1-K/b}\right)_+\end{aligned}$$

As this looks like the lognormal problem for $F_0, K \in (a,b)$, we obtain

$$\Psi = x_0 N(d_+) - \frac{K-a}{1-K/b}N(d_-), \quad d_\pm = \frac{\ln\left(x_0/\left(\frac{K-a}{1-K/b}\right)\right)}{(1-a/b)\sigma\sqrt{T}} \pm \frac{1}{2}(1-a/b)\sigma\sqrt{T}$$

Transforming back to the original coordinates gives

$$\begin{aligned}U &= \frac{1}{1-a/b}\left((F_0-a)(1-K/b)N(d_+)-(K-a)(1-F_0/b)N(d_-)\right), \\ d_\pm &= \frac{\ln\left(\left(\frac{F_0-a}{1-F_0/b}\right)/\left(\frac{K-a}{1-K/b}\right)\right)}{(1-a/b)\sigma\sqrt{T}} \pm \frac{1}{2}(1-a/b)\sigma\sqrt{T}\end{aligned}$$

The cases when $F_0 < a$ and $F_0 > b$ can be solved in a similar way. For example, when $F_0, K > b$, we obtain

$$\begin{aligned}U &= \frac{1}{1-a/b}\left((F_0-a)(1-K/b)N(-d_+)-(K-a)(1-F_0/b)N(-d_-)\right), \\ d_\pm &= \frac{\ln\left(\left(\frac{F_0-a}{1-F_0/b}\right)/\left(\frac{K-a}{1-K/b}\right)\right)}{(1-a/b)\sigma\sqrt{T}} \pm \frac{1}{2}(1-a/b)\sigma\sqrt{T}\end{aligned}$$

The Green's function can be computed by taking the second derivative of V with respect to K, see Sect. 6.2. It follows from the form of the Green's function that if F starts in one of the regions $(-\infty, a)$, (a,b) or (b,∞), it stays there forever.

To simplify the calibration, we want the variables a and b to have as small an effect as possible on the ATM price, which suggests the rescaling

$$dF_t = \sigma\frac{(F_t-a)(1-F_t/b)}{(F_0-a)(1-F_0/b)}F_0 dW_t, \quad a<b$$

It is then necessary to replace σ with $\sigma\frac{F_0}{(F_0-a)(1-F_0/b)}$ in the option pricing formula. As the quadratic volatility process has three free parameters: σ, a and b, it can be used not only to control the skew, but the smile as well.

The quadratic volatility model can be used for $F_0 \in (a,b)$ if we are relatively confident that the underlying F stays within the region (a,b). For instance, based on the observation that most of the major currency pairs have historically been within certain bounds, this model was used in Ingersoll (1996) for FX rates. Care needs to be taken when using the quadratic volatility model as the implied volatility has a peculiar dependence on the strike when F_0 is close to a or b, or when the maturity is long.

The instance $F_0 > b > 0$ can be used to model skew for which implied volatility increases with the strike. This skew is in the opposite direction to the shifted lognormal model and is useful when the volatility increases with the underlying. This is sometimes the situation for commodities and FX rates.

5.6 The Ornstein-Uhlenbeck Process

The *Ornstein-Uhlenbeck process* has the form

$$dF_t = \lambda_t(\tilde{F}_t - F_t)dt + \sigma_t dW_t$$

We derive and analyze the Green's function, but we do not compute the option price as the SDE is not driftless. We have chosen to explicitly express the time-dependence of the variables as the general solution cannot be obtained from the constant coefficient case by simple substitutions such as $\sigma^2 T \to \int_0^T \sigma^2 du$.

For $F_t > \tilde{F}_t$ the drift term is negative while it is positive for $F_t < \tilde{F}_t$. It means that F_t is pulled towards $\tilde{F}_t$ and the strength of the pull is determined by the factor λ_t. For this reason, $\tilde{F}_t$ is referred to as the *mean-reversion level* while λ_t is called the *mean-reversion factor*. Clearly, for $\lambda_t = 0$ the process reduces to a Brownian motion and is independent of the mean-reversion level. For $\lambda_t \to \infty$, on the other hand, the process immediately becomes equal to $\tilde{F}_t$ and stays there forever.

The solution to the Ornstein-Uhlenbeck process is obtained from

$$\begin{aligned}
d\left(e^{\int_0^t \lambda_u du} F_t\right) &= e^{\int_0^t \lambda_u du} \lambda_t \tilde{F}_t dt + e^{\int_0^t \lambda_u du} \sigma_t dW_t \\
\Rightarrow F_T &= e^{-\int_0^T \lambda_u du}\left(F_0 + \int_0^T e^{\int_0^t \lambda_u du}\lambda_t \tilde{F}_t dt + \int_0^T e^{\int_0^t \lambda_u du}\sigma_t dW_t\right) \\
&= e^{-\int_0^T \lambda_u du}\left(F_0 + \int_0^T e^{\int_0^t \lambda_u du}\lambda_t \tilde{F}_t dt + \bar{\omega}_T \sqrt{T} X\right)
\end{aligned}$$

where $X \sim \mathcal{N}(0,1)$, $\omega_t = e^{\int_0^t \lambda_u du}\sigma_t$, and $\bar{\omega}_T$ is defined by $\bar{\omega}_T^2 T = \int_0^T \omega^2 du$. As F_T is normally distributed, it is completely determined by its mean and variance. To analyze the solution in more detail, let us consider the situation when the coefficients are constant. We then obtain

$$F_T = e^{-\lambda T}\left(F_0 + \lambda \tilde{F} \int_0^T e^{\lambda t} dt + \sigma \int_0^T e^{\lambda t} dW_t\right)$$

with mean and variance given by

$$
\begin{aligned}
E\left[F_T\right] &= \tilde{F} + e^{-\lambda T}\left(F_0 - \tilde{F}\right) \\
\operatorname{Var}\left(F_T\right) &= \frac{\sigma^2}{2\lambda}\left(1 - e^{-2\lambda T}\right)
\end{aligned}
$$

For T or λ small, the expressions simplify to

$$
\begin{aligned}
E\left[F_T\right] &\to F_0 \\
\operatorname{Var}\left(F_T\right) &\to \sigma^2 T
\end{aligned}
$$

which is the same result as would have been obtained for the corresponding non mean-reverting process. As mentioned above, for large λ the process becomes constant and equal to the mean-reversion level:

$$
\begin{aligned}
E\left[F_T\right] &\to \tilde{F} \\
\operatorname{Var}\left(F_T\right) &\to 0
\end{aligned}
$$

Finally, we consider the limit of large T. Then

$$
\begin{aligned}
E\left[F_T\right] &\to \tilde{F} \\
\operatorname{Var}\left(F_T\right) &\to \frac{\sigma^2}{2\lambda}
\end{aligned}
$$

As expected, the mean is equal to $\tilde{F}$. It is interesting to observe that the variance becomes independent of T. The reason is that for large T a state of equilibrium is established between the variance increasing effect from the Brownian driver and the variance decreasing effect from the pull towards the mean-reversion level $\tilde{F}$.

The Ornstein-Uhlenbeck belongs to the class of *mean-reverting processes*. Additional examples of processes belonging to this class are the geometric Ornstein-Uhlenbeck model, obtained by replacing σ_t with $\sigma_t F_t$, and the Cox-Ingersoll-Ross (CIR) process, obtained by replacing σ_t with $\sigma_t \sqrt{F_t}$ in the equation for dF_t.

5.7 The Brownian Bridge

To explain the purpose of *Brownian bridge processes*, we first consider the distribution of a Brownian motion conditional on that its value W_T at a future time T is known. Since

$$\begin{aligned}&\text{Covar}\left(W_t - \frac{t}{T}W_T, W_T\right)\\&= \text{Covar}\left(W_t, W_T - W_t\right) + \text{Covar}\left(W_t, W_t\right) - \frac{t}{T}\text{Covar}\left(W_T, W_T\right)\\&= t - \frac{t}{T}T = 0\end{aligned}$$

and as Gaussian variables are characterized by their covariance, $W_t - \frac{t}{T}W_T$ and W_T must be independent. Conditional on that $W_T = x$, W_t and $W_t - \frac{t}{T}W_T + \frac{t}{T}x$ have equal distribution. It means that the conditional distribution of W_t is equal to $\frac{t}{T}x$ added to the conditional distribution of $W_t - \frac{t}{T}W_T$. However, the latter is equal to the non-conditional distribution of $W_t - \frac{t}{T}W_T$ according to the above discussion. As this is a normal variable with mean 0 and variance $t(T-t)/T$, it follows that

$$W_t|_{W_T} \sim \mathcal{N}\left(\frac{t}{T}W_T, \frac{t(T-t)}{T}\right)$$

This argument can be generalized to obtain the distribution of W_t conditional on W_s and W_T for $s < t < T$. Using the Brownian motion $W_t' = W_{s+t} - W_s$ in the above result gives

$$W_t|_{W_s, W_T} \sim \mathcal{N}\left(W_s + \frac{t-s}{T-s}(W_T - W_s), \frac{(t-s)(T-t)}{T-s}\right)$$

Consider now

$$X_t = \frac{T-t}{\sqrt{T}}W_{t/(T-t)} + \frac{t}{T}x$$

To analyze this process, we use the fact that

$$\tilde{W}_t = \int_0^{g^{-1}(t)} \sigma_u dW_u, \quad g(t) = \int_0^t \sigma_u^2 du$$

is a Brownian motion for arbitrary functions σ_u. Indeed, it is straightforward to show that the defining properties for a Brownian motion (see Appendix) is satisfied. For example,

$$\begin{aligned}\text{Var}\left(\tilde{W}_t - \tilde{W}_s\right) &= \text{Var}\left(\int_{g^{-1}(s)}^{g^{-1}(t)} \sigma_u dW_u\right) = \int_{g^{-1}(s)}^{g^{-1}(t)} \sigma_u^2 du\\&= \int_0^{g^{-1}(t)} \sigma_u^2 du - \int_0^{g^{-1}(s)} \sigma_u^2 du = t - s\end{aligned}$$

From this general statement we obtain

$$X_t = (T-t)\int_0^t (T-u)^{-1} dW_u + \frac{t}{T}x$$

It follows that

$$\begin{aligned} X_t &= (T-t)\int_0^s (T-u)^{-1}dW_u + (T-t)\int_s^t (T-u)^{-1}dW_u + \frac{t}{T}x \\ &= \frac{T-t}{T-s}\left(X_s - \frac{s}{T}x\right) + \frac{t}{T}x + (T-t)\int_s^t (T-u)^{-1}dW_u \\ &= X_s + \frac{t-s}{T-s}(x - X_s) + (T-t)\int_s^t (T-u)^{-1}dW_u \end{aligned}$$

from which we conclude that $X_t|_{X_s}$ has the same distribution as $W_t|_{W_s,W_T}$ if $X_s = W_s$ and $x = W_T$. In fact, it can be proven that X_T is the unique continuous-path process with this property. This process is called a Brownian bridge.

From the computation

$$\begin{aligned} dX_t &= -\left(\int_0^t (T-u)^{-1}dW_u\right)dt + \frac{1}{T}xdt + (T-t)(T-t)^{-1}dW_t \\ &= -(T-t)^{-1}\left(X_t - \frac{t}{T}x\right)dt + (T-t)^{-1}\frac{T-t}{T}xdt + dW_t \\ &= \frac{x - X_t}{T-t}dt + dW_t \end{aligned}$$

we conclude that a Brownian bride process is nothing more than an Ornstein-Uhlenbeck process with time-dependent parameters. The mean reversion is such that the process converges to the point x when $t \to T$.

5.8 The CEV Process

The *constant elasticity of variance (CEV) process*

$$dF_t = \sigma F_t^\beta dW_t$$

can be used as an alternative to the shifted lognormal model to interpolate between the normal ($\beta = 0$) and the lognormal ($\beta = 1$) processes. We solve this model by using the corresponding backward Kolmogorov equation

$$U_\tau = \frac{1}{2}F^{2\beta}U_{FF}, \quad \tau = \sigma^2(T-t)$$

An alternative solution method is given in Sect. 5.9. The above PDE can be transformed to have integer powers by setting $\Omega(\tau, x) = U(\tau, F)$, where $x =$

$F^{1-\beta}$. Using

$$U_F = \Omega_x(1-\beta)F^{-\beta} \Rightarrow U_{FF} = \Omega_{xx}(1-\beta)^2 F^{-2\beta} - \Omega_x \beta(1-\beta)F^{-\beta-1}$$

we obtain

$$\Omega_\tau = U_\tau = \frac{1}{2}F^{2\beta}U_{FF} = \frac{1}{2}(1-\beta)^2\Omega_{xx} - \frac{1}{2}\beta(1-\beta)x^{-1}\Omega_x$$

The similarity between this PDE and the operator

$$\mathcal{D}_x^\gamma = \frac{d^2}{dx^2} + x^{-1}\frac{d}{dx} - \gamma^{-2}x^{-2}$$

appearing in Bessel's differential equation is clear. To make this explicit, we use

$$\begin{aligned} \Omega(\tau, x) &= x^{1/\gamma}\Phi(\tau, x), \quad \gamma = 2(1-\beta) \\ \Rightarrow \Omega_x &= x^{1/\gamma}\Phi_x + \frac{1}{\gamma}x^{1/\gamma-1}\Phi \\ \Rightarrow \Omega_{xx} &= x^{1/\gamma}\Phi_{xx} + \frac{2}{\gamma}x^{1/\gamma-1}\Phi_x + \frac{1}{\gamma}\left(\frac{1}{\gamma}-1\right)x^{1/\gamma-2}\Phi \end{aligned}$$

to get

$$\begin{aligned} \Phi_\tau &= x^{-1/\gamma}\Omega_\tau = \frac{1}{8}x^{-1/\gamma}\left(\gamma^2\Omega_{xx} + (\gamma-2)\gamma x^{-1}\Omega_x\right) \\ &= \frac{1}{8}\left(\gamma^2\Phi_{xx} + 2\gamma x^{-1}\Phi_x + (1-\gamma)x^{-2}\Phi + (\gamma-2)\gamma x^{-1}\Phi_x + (\gamma-2)x^{-2}\Phi\right) \\ &= \frac{1}{8}\gamma^2\left(\Phi_{xx} + x^{-1}\Phi_x - \gamma^{-2}x^{-2}\Phi\right) = \frac{1}{8}\gamma^2\mathcal{D}_x^\gamma\Phi \end{aligned}$$

Finally, by a time-scaling

$$\Psi(\omega, x) = \Phi(\tau, x), \quad \omega = \frac{\gamma^2}{4}\tau$$

we arrive at

$$\Psi_\omega = \frac{1}{2}\mathcal{D}_x^\gamma\Psi$$

In summary, the transformation that has been made is

$$\begin{cases} U(\tau, F) = F^{1/2}\Psi\left(\frac{\gamma^2}{4}\tau, F^{\gamma/2}\right) \\ \Psi(\omega, x) = x^{-1/\gamma}U\left(\frac{4}{\gamma^2}\omega, x^{2/\gamma}\right) \end{cases}$$

To solve the resulting PDE we use the Hankel transformations

$$\begin{cases} \Psi(\omega, x) = \int_0^\infty k\hat{\Psi}(\omega, k) J_{1/\gamma}(kx)dk \\ \hat{\Psi}(\omega, k) = \int_0^\infty x\Psi(\omega, x) J_{1/\gamma}(kx)dx \end{cases}$$

which can be verified from the orthogonality relation between Bessel functions

$$\int_0^\infty kJ_{1/\gamma}(kx')J_{1/\gamma}(kx)dk = \delta(x - x')/x$$

As Bessel functions solve Bessel's differential equation

$$\mathcal{D}_x^\gamma J_{1/\gamma}(kx) = -k^2 J_{1/\gamma}(kx)$$

the transformed problem becomes

$$\hat{\Psi}_\omega = -\frac{1}{2}k^2\hat{\Psi} \Leftrightarrow \hat{\Psi}(\omega, k) = e^{-k^2\omega/2}\hat{\Psi}(\omega = 0, k)$$

The Green's function $q(\omega, x, x')$ for the PDE satisfied by $\Psi(\omega, x)$ is by definition the solution to the problem

$$\begin{cases} q_\omega = \frac{1}{2}\mathcal{D}_x^\gamma q \\ q(\omega = 0, x, x') = \delta(x - x') \end{cases}$$

The transformed problem then has the initial condition

$$\hat{q}(\omega = 0, k, x') = \int_0^\infty x\delta(x - x')J_{1/\gamma}(kx)dx = x'J_{1/\gamma}(kx')$$

from which we obtain

$$q(\omega, x, x') = \int_0^\infty ke^{-k^2\omega/2}x'J_{1/\gamma}(kx')J_{1/\gamma}(kx)dk$$

We change integration variable from k to kx' and use the equation

$$\int_0^\infty ke^{-k^2\eta/2}J_{1/\gamma}(k)J_{1/\gamma}(k\chi)dk = \frac{1}{\eta}I_{1/\gamma}(\chi/\eta)e^{-(1+\chi^2)/2\eta}$$

which can be found in, for example, Luke (1962). The result is

$$\begin{aligned} q(\omega, x, x') &= x'^{-1}\int_0^\infty k e^{-k^2\omega/2x'^2} J_{1/\gamma}(k) J_{1/\gamma}(kx/x')dk \\ &= x'^{-1}\frac{x'^2}{\omega} I_{1/\gamma}\left(\frac{x'^2}{\omega}\frac{x}{x'}\right) e^{-(1+(\frac{x}{x'})^2)x'^2/2\omega} = \frac{x'}{\omega} I_{1/\gamma}\left(\frac{xx'}{\omega}\right) e^{-(x^2+x'^2)/2\omega} \end{aligned}$$

The Green's function can be used to find the solution

$$\Psi(\omega, x) = \int_0^\infty h(x')q(\omega, x, x')dx'$$

to the general problem

$$\begin{cases} \Psi_\omega = \frac{1}{2}\mathcal{D}_x^\gamma \Psi \\ \Psi(\omega = 0, x) = h(x) \end{cases}$$

As an example of the applicability of the above method, let us find the Green's function to the original PDE satisfied by U. We have the initial condition

$$\begin{aligned} & U(\tau = 0, F) = \delta(F - F') \\ \Leftrightarrow\ & \Psi(\omega = 0, x) = x^{-1/\gamma} U\left(\tau = 0, x^{2/\gamma}\right) = x^{-1/\gamma}\delta(x^{2/\gamma} - F') \\ & \qquad = x^{-1/\gamma}\frac{\gamma}{2}x^{1-2/\gamma}\delta(x - F'^{\gamma/2}) = \frac{\gamma}{2}x^{1-3/\gamma}\delta(x - F'^{\gamma/2}) \end{aligned}$$

which gives

$$\Psi(\omega, x) = \frac{\gamma}{2\omega}F'^{\gamma-3/2} I_{1/\gamma}\left(\frac{xF'^{\gamma/2}}{\omega}\right) e^{-(x^2+F'^\gamma)/2\omega}$$

The Green's function $p(T, F'; t, F)$ for the CEV process is therefore given by

$$\begin{aligned} p(T, F'; t, F) &= F^{1/2}\frac{4}{\gamma^2\tau}\frac{\gamma}{2}F'^{\gamma-3/2} I_{1/\gamma}\left(4\frac{(FF')^{\gamma/2}}{\gamma^2\tau}\right) e^{-2(F^\gamma+F'^\gamma)/\gamma^2\tau} \\ &= \frac{2}{\gamma\tau}F'^{\gamma-2}(FF')^{1/2} I_{1/\gamma}\left(4\frac{(FF')^{\gamma/2}}{\gamma^2\tau}\right) e^{-2(F^\gamma+F'^\gamma)/\gamma^2\tau} \\ \tau &= \sigma^2(T - t), \quad \gamma = 2(1 - \beta) \end{aligned}$$

Observe that we solved the PDE by using the fact that $J_{1/\gamma}$ is a solution of the Bessel PDE. Recall that a second-order differential equation can have more than one solution and the choice of solution determines the boundary condition. Our choice of solution is such that $p(F', T; F, t)$ is zero at the boundary $F' = 0$. This condition was enforced by our choice of solution for $\beta < 1/2$ while it is satisfied automatically for $\beta \geq 1/2$.

In the normal case $\gamma = 2$ we can use

$$I_{1/2}(z) = \sqrt{\frac{2}{\pi z}}\sinh(z)$$

to obtain

$$\begin{aligned}p_{\beta=0}(T, F'; t, F) &= \frac{1}{\tau}(FF')^{1/2}\sqrt{\frac{2\tau}{\pi FF'}}\frac{1}{2}\left(e^{FF'/\tau} - e^{-FF'/\tau}\right)e^{-(F^2+F'^2)/2\tau}\\ &= \frac{1}{\sqrt{2\pi\tau}}\left(e^{-(F-F')^2/2\tau} - e^{-(F+F')^2/2\tau}\right)\end{aligned}$$

which as expected is the normal Green's function with an absorbing boundary condition. Recall from Sect. 5.3 that in this situation there is a finite probability that $F = 0$. This is in fact true for all $\beta < 1$ and it can be shown that the probability is given by

$$P(F' = 0) = G\left(\frac{1}{\gamma}, \frac{2}{\tau\gamma^2}F^{\gamma}\right)$$

where G is the *complementary gamma distribution function*

$$G(y, x) = \Gamma(y)^{-1}\int_x^{\infty} e^{-z}z^{y-1}dz$$

and Γ is the *gamma function*

$$\Gamma(y) = \int_0^{\infty} e^{-z}z^{y-1}dz$$

satisfying $\Gamma(n + 1) = n!$ for positive integers n.

The way F ends up in the absorbing point $F = 0$ depends on the value of β. Indeed, by using the asymptotic limit

$$I_\alpha(z) \approx \frac{1}{\Gamma(\alpha+1)}(z/2)^{\alpha}, \quad \text{for } 0 < z \ll \sqrt{\alpha+1}$$

we obtain the Green's function for small F':

$$\begin{aligned}p(T, F'; t, F) &\approx \frac{2}{\gamma\tau}F'^{\gamma-2}(FF')^{1/2}\frac{1}{\Gamma(1/\gamma+1)}(2\frac{(FF')^{\gamma/2}}{\gamma^2\tau})^{1/\gamma}e^{-2F^{\gamma}/\gamma^2\tau}\\ &= \left(\frac{2}{\gamma^2\tau}\right)^{1/\gamma+1}\frac{\gamma F}{\Gamma(1/\gamma+1)}e^{-2F^{\gamma}/\gamma^2\tau}F'^{\gamma-1}\end{aligned}$$

We conclude that p has the asymptotic behavior

$$\begin{cases} p \to 0 & \text{for } \beta < 1/2 \\ p \to \text{const} & \text{for } \beta = 1/2 \\ p \to \infty & \text{for } 1/2 < \beta < 1 \end{cases}$$

when $F' \to 0$.

The call option price can be computed by integrating the payoff over the Green's function. Unfortunately, the series expansion of the modified Bessel function converges slowly, which means that there is a performance impact for accurate calculation of option prices. An alternative method, described in Lipton (2001), makes use of the fact that I_α is particularly simple for α a half integer: $\alpha = 1/2, 3/2, 5/2, \ldots$, corresponding to $\beta = 0, 2/3, 4/5, 5/6, \ldots$. The modified Bessel functions are then simple expressions of the hyperbolic functions (sinh and cosh), which means that it is straightforward, though a bit cumbersome, to compute the option price. The option price for a general value of β can then approximately be obtained by interpolation.

5.9 The Bessel Process

The *Bessel process* is closely related to the CEV process and is reviewed in detail in Revuz and Yor (1999). Several of the variables that appear in our treatment of Bessel processes are gamma distributed so we start our discussion there.

The *gamma distribution* is the two-parameter family $\Gamma_{\alpha\beta}$ of random variables with PDF

$$p_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

The moment generating function is given by

$$\begin{aligned} M_X(k) = E\left[e^{kX}\right] &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty z^{\alpha-1} e^{-(1-k\beta)z/\beta} dz \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)(1-k\beta)^\alpha} \int_0^\infty z^{\alpha-1} e^{-z/\beta} dz = \frac{1}{(1-k\beta)^\alpha} \end{aligned}$$

As an example of a gamma distributed variable, let $X \sim \mathcal{N}(0,1)$ and consider

$$\begin{aligned} P(X^2 < x) = 2P(0 < X < \sqrt{x}) &= 2\frac{1}{\sqrt{2\pi}} \int_0^{\sqrt{x}} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_0^x z^{-1/2} e^{-z/2} dz = \frac{1}{2^{1/2}\Gamma(\frac{1}{2})} \int_0^x z^{-1/2} e^{-z/2} dz \end{aligned}$$

from which it follows that $X^2 \sim \Gamma_{\frac{1}{2},2}$.

By multiplying together the generating functions of $\Gamma_{\alpha_i \beta}$ distributed variables $\{X_i\}$, it follows that

$$M_{\sum_{i=1}^{\delta} X_i}(k) = \frac{1}{(1-k\beta)^{\sum_{i=1}^{\delta}\alpha_i}}$$

which means that $\sum_{i=1}^{\delta} X_i \sim \Gamma_{\sum_{i=1}^{\delta}\alpha_i,\beta}$. If $\{X_i\}$ are normally distributed, then $\sum_{i=1}^{\delta} X_i^2 \sim \Gamma_{\frac{\delta}{2},2}$. In the special case $\beta = 2$, the gamma distribution is called the *chi-square distribution* $\chi^2_\delta = \Gamma_{\frac{\delta}{2},2}$.

Let X be equal to a constant $\sqrt{b}$ added to a $\mathcal{N}(0,1)$ distributed variable. The moment generating function for X^2 is

$$\begin{aligned}
M_{X^2}(k) = E\left[e^{kX^2}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{k(\sqrt{b}+z)^2} e^{-z^2/2} dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(1-2k)z^2/2+2k\sqrt{b}z+kb} dz \\
&= \frac{1}{\sqrt{1-2k}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2+\frac{2k\sqrt{b}}{\sqrt{1-2k}}z+kb} dz \\
&= \frac{1}{\sqrt{1-2k}} e^{kb/(1-2k)} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\left(z-\frac{2k\sqrt{b}}{\sqrt{1-2k}}\right)^2/2} dz = \frac{1}{\sqrt{1-2k}} e^{kb/(1-2k)}
\end{aligned}$$

By multiplying the individual moment generating functions, we obtain the moment generating function for a sum $\sum_{i=1}^{\delta} X_i^2$, where the X_is are independent random variables that can be written as a sum of a constant $\sqrt{b_i}$ added to a $\mathcal{N}(0,1)$ distributed variable:

$$M_{\sum_{i=1}^{\delta} X_i^2}(k) = \frac{1}{(1-2k)^{\delta/2}} e^{kb/(1-2k)}, \quad b = \sum_{i=1}^{\delta} b_i$$

This distribution is called the *non-central chi-square distribution*. To find its PDF, we rewrite the moment generating function as

$$M_{\sum_{i=1}^{\delta} X_i^2}(k) = \frac{1}{(1-2k)^{\delta/2}} e^{-b/2} e^{\frac{b}{2}\frac{1}{1-2k}} = \sum_{n=0}^{\infty} \frac{b^n e^{-b/2}}{n! 2^n} \frac{1}{(1-2k)^{n+\delta/2}}$$

Using a term-by-term argument, we obtain the PDF

$$p_{\sum_{i=1}^{\delta} X_i^2}(x) = \sum_{n=0}^{\infty} \frac{b^n e^{-b/2}}{n! 2^{2n+\delta/2}\Gamma(n+\delta/2)} x^{n+\delta/2-1} e^{-x/2}$$

which can be written as

$$p_{\sum_{i=1}^{\delta} X_i^2}(x) = \frac{1}{2} e^{-(b+x)/2} \left(\frac{x}{b}\right)^{\delta/4-1/2} \sum_{n=0}^{\infty} \frac{(\sqrt{bx})^{2n+\delta/2-1}}{n! 2^{2n+\delta/2-1} \Gamma(n+\delta/2)}$$
$$= \frac{1}{2} e^{-(b+x)/2} \left(\frac{x}{b}\right)^{\delta/4-1/2} I_{\delta/2-1}(\sqrt{bx})$$

where we have used a well-known expression for the modified Bessel function:

$$I_\nu(z) = \sum_{n=0}^{\infty} \frac{(z/2)^{\nu+2n}}{n!\Gamma(n+\nu+1)}$$

The above result can be further extended to a weighted sum of squares of the X_is, see Sect. 10.4. Here we instead generalize the X_is to Brownian motions W_i with starting points $\sqrt{b_i}$. We also discuss a natural extension to positive non-integers δ.

Using the time-scaling $cZ_{t/c^2} \sim Z_t$ for a standard Brownian motion, we can write $W_{i,t} = Z_t + \sqrt{b_i} \sim \sqrt{t}\left(Z_1 + \sqrt{b_i/t}\right)$, where $Z_1 \sim \mathcal{N}(0,1)$. It then follows from the above that the moment generating function for $F_t = \sum_{i=1}^{\delta} W_{i,t}^2$ is given by

$$M_F(k) = \frac{1}{(1-2kt)^{\delta/2}} e^{kb/(1-2kt)}$$

This process is called the *squared Bessel process*. It depends on two parameters δ and b, and is denoted by $\mathrm{BESQ}^\delta(b)$. Observe that

$$F_t = F_0 + \int_0^t dF_s = b + 2\int_0^t \sum_{i=1}^{\delta} W_{i,s} dW_{i,s} + \int_0^t \sum_{i=1}^{\delta} ds$$
$$= b + 2\int_0^t \sqrt{\sum_{i=1}^{\delta} W_{i,s}^2} dZ_s + \delta t = b + 2\int_0^t \sqrt{|F_s|} dZ_s + \delta t$$

where Z_t is a standard Brownian motion. This formula provides us with a natural extension of squared Bessel processes to positive non-integers δ. It can be proven that the SDE has a unique strong solution when $b, \delta \geq 0$. Furthermore, since $F_t = 0$ is the solution when $b, \delta = 0$, it is possible to use comparison SDE theorems to prove that $F_t \geq 0$ when $b, \delta \geq 0$. The absolute value under the square root is therefore not necessary. The positivity of the solution to an SDE is often a desirable property in mathematical finance and this is one of the main reasons for the popularity of squared Bessel processes.

Using the defining SDE, we see that if $F \sim \mathrm{BESQ}^\delta(b)$ and $F' \sim \mathrm{BESQ}^{\delta'}(b')$ are independent, then $F + F' \sim \mathrm{BESQ}^{\delta+\delta'}(b+b')$. Therefore, if $M(\delta, b)$ is the moment generating function of $\mathrm{BESQ}^\delta(b)$, then

$$M(\delta, b)M(\delta', b') = M(\delta + \delta', b + b')$$

From this relation we conclude that $M(\delta, 0) = \alpha A^{\delta}$ and $M(0, b) = \beta B^{b}$, and since $M(\delta, b) = M(\delta, 0)M(0, b)$ we obtain $M(\delta, b) = \gamma A^{\delta} B^{b}$. It implies that $M_F(k)$ must be of the above form also for positive non-integers δ. The PDF for the squared Bessel process can be derived in exactly the same way as was done for the corresponding random variables. The result is

$$p_{F_t}(x) = \frac{1}{2t} e^{-(b+x)/2t} \left(\frac{x}{b}\right)^{\delta/4-1/2} I_{\delta/2-1}\left(\frac{\sqrt{bx}}{t}\right)$$

We now show that it is possible to obtain several familiar processes from the squared Bessel process. For all these processes, the PDF and the moment generating function can easily be obtained from the corresponding results for the squared Bessel process.

The square root G of the squared Bessel process is called the Bessel process $\mathrm{BES}^{\delta}(b)$. For $\delta \geq 2$, it can be shown that the point $x = 0$ is unattainable. We can therefore apply Ito's lemma and obtain the SDE satisfied by the Bessel process:

$$G_t = G_0 + Z_t + \frac{\delta - 1}{2} \int_0^t G_s^{-1} ds$$

Well-known examples of Bessel processes are $\max_{0 \leq u \leq t} W_u - W_t \sim \mathrm{BES}^1(0)$ and $2\max_{0 \leq u \leq t} W_u - W_t \sim \mathrm{BES}^3(0)$ for W a standard Brownian motion, see Pitman (1975).

Consider the inclusion of mean-reversion in the squared Bessel process:

$$dG_t = 2\sqrt{|G_t|} dZ_t + (2\beta G_t + \delta)dt$$

This is a popular SDE, in particular for the short rate in interest rate modeling where it is called the *Cox-Ingersoll-Ross model*. It is also used for the volatility in the Heston model, see Sect. 7.4. Setting $F_t' = e^{-2\beta t} G_t$ gives

$$dF_t' = 2e^{-\beta t} \sqrt{|F_t'|} dZ_t + e^{-2\beta t} \delta dt$$

We make use of the fact that $\int_0^t \sigma_u dZ_u$ and $Z_{\int_0^t \sigma_u^2 du}$ describe the same process for Z a standard Brownian motion, to arrive at

$$\begin{aligned} dF_t' &= 2\sqrt{|F_t'|} dZ(\int_0^t e^{-2\beta u} du) + e^{-2\beta t} \delta dt \\ &= 2\sqrt{|F_t'|} dZ((1 - e^{-2\beta t})/2\beta) + \delta d(1 - e^{-2\beta t})/2\beta \end{aligned}$$

from which we see that $F_t' = F((1 - e^{-2\beta t})/2\beta)$, where $F \sim \mathrm{BESQ}^{\delta}(\cdot)$. The distribution for the mean-reverting process can then easily be computed.

The squared Bessel process is closely related to the CEV process. Indeed, for

$$dG_t = \sigma G_t^{\beta} dW_t$$

and $F' = G^{2(1-\beta)}$, we obtain

$$\begin{aligned} dF'_t &= 2(1-\beta)G_t^{1-2\beta}\sigma G_t^{\beta} dW_t + \frac{1}{2}2(1-\beta)(1-2\beta)G_t^{-2\beta}\sigma^2 G_t^{2\beta} dt \\ &= 2(1-\beta)\sigma\sqrt{F'_t} dW_t + (1-\beta)(1-2\beta)\sigma^2 dt \\ &= 2\sqrt{F'_t} dW_{(1-\beta)^2\sigma^2 t} + \frac{1-2\beta}{1-\beta} d((1-\beta)^2\sigma^2 t) \end{aligned}$$

Thus, $G_t = F_{(1-\beta)^2\sigma^2 t}^{1/2(1-\beta)}$ for $F \sim \mathrm{BESQ}^{(1-2\beta)/(1-\beta)}(\cdot)$. With $\gamma = 2(1-\beta)$ and $\tau = \sigma^2 t$ we get $G_t = F_{\gamma^2\tau/4}^{1/\gamma}$ for $F \sim \mathrm{BESQ}^{2(\gamma-1)/\gamma}(b)$, with $b = x_0^{\gamma}$ where $G_{t=0} = x_0$. Using the PDF for the squared Bessel function gives

$$\begin{aligned} p_{G_t}(x) &= p_{F_{\gamma^2\tau/4}}(x^{\gamma})\frac{dx^{\gamma}}{dx} \\ &= \frac{1}{2\gamma^2\tau/4} e^{(x_0^{\gamma}+x^{\gamma})/(2\gamma^2\tau/4)} \left(\frac{x^{\gamma}}{x_0^{\gamma}}\right)^{-1/2\gamma} I_{-1/\gamma}\left(\frac{\sqrt{x_0^{\gamma}x^{\gamma}}}{\gamma^2\tau/4}\right)\gamma x^{\gamma-1} \\ &= \frac{2}{\gamma\tau} x^{\gamma-2}\left(x_0^{\gamma}x^{\gamma}\right)^{1/2} I_{1/\gamma}\left(4\frac{(x_0 x)^{\gamma/2}}{\gamma^2\tau}\right) e^{-2(x_0^{\gamma}+x^{\gamma})/\gamma^2\tau} \end{aligned}$$

which is identical to the expression in Sect. 5.8.

5.10 Non-Analytic SDEs

The SDEs considered so far have all been analytically solvable. This is useful when pricing vanilla products for which the performance is often important. However, there are many situations in derivatives pricing where analyticity is of minor importance. As an example of this, we now explain how exotic derivatives can be priced by simulating SDEs that do not have closed-form solutions.

A simulation is often not done over a single time step but over a discrete set of dates. The reason can be that an exotic product has several payments or that the model requires multiple simulation points, see, for instance, the lognormal LMM model of Sect. 13.17 that in general is set up so that the simulation is done according to the frequency of the underlying LIBOR rates. It is in this situation necessary to use an SDE that can be simulated over discrete time steps, typically of size less than a year. The SDE also has to be such that it can be calibrated to vanilla products. We now show how both these demands can be satisfied by SDEs that do not have known analytical solutions.

As an illustrating example, we consider an SDE that gives a volatility skew without permitting negative values for the underlying, as opposed to the shifted lognormal SDE, or by allowing a non-zero probability for the underlying to end up at 0, as opposed to the CEV process. The precise specification of the SDE we have in mind is

$$dF = \frac{\omega(\beta, F)}{\omega(\beta, F_0)} \hat{\sigma} F_0 dW_t = \omega(\beta, F)\sigma dW_t, \quad \omega(\beta, F) = \beta^{-1}\left(1 - e^{-\beta F}\right)$$

where $\hat{\sigma}$ is a constant.

Just as for the shifted lognormal process and the CEV process, the β parameter controls the skew. For instance, $\beta \to 0$ gives $\omega(\beta, F) \to F$ meaning that the process turns lognormal. If $\beta \to \infty$, on the other hand, $\omega(\beta, F)/\omega(\beta, F_0) \to 1$ giving a normal process. We also note that the process becomes lognormal in the limit $F \to 0$, which implies that the underlying always stays positive. Furthermore, the process is normal in the limit $F \to \infty$. We conclude that the skew is such that the implied volatility decreases with increasing strike.

The simulation of the process can be done by using the SDE the way it is and taking small time steps. This is quite time consuming though. As an alternative, consider the transformation

$$G = \beta^{-1}\left(e^{\beta F} - 1\right)$$

Ito's lemma gives

$$dG = G\left(\sigma dW_t + \frac{1}{2}\frac{\beta G}{1 + \beta G}\sigma^2 dt\right)$$

so the non-linearity is transferred from the stochastic part of the SDE to the deterministic part (the drift). The reason for doing this transformation is that the stochastic part of the SDE can be viewed as the lowest-order term while the drift is of higher order, which can be formally understood from the relation $dW^2 = dt$. With the non-linearity in the drift, there exist simulation schemes of high accuracy.

For the simulation between two time steps t_n and t_{n+1}, consider first the situation when the G-dependence of the drift is frozen to its value G_n just before the simulation. The SDE can then be solved as:

$$dG = G\left(\sigma dW_t + \frac{1}{2}\frac{\beta G_n}{1 + \beta G_n}\sigma^2 dt\right)$$

$$\Rightarrow G_{n+1} = G_n \exp\left(\sigma\sqrt{\Delta t}X - \frac{1}{2}\frac{1}{1 + \beta G_n}\sigma^2 \Delta t\right), \quad X \sim \mathcal{N}(0, 1)$$

To understand the consequence of the freezing, transform back to the original variable

$$dF = \omega(\beta, F)\sigma dW_t + \frac{1}{2}\beta\omega(\beta, F)\left(\omega(\beta, F_n) - \omega(\beta, F)\right)\sigma^2 dt$$

We see that a mean-reverting drift has been added to the SDE, which gives a smaller variance. To regain the variance, we apply the well-known *predictor-corrector technique*. In this method the simulation is done an additional time by using the result G_{n+1} of the first simulation as the freezing value in the SDE. The final result is the arithmetic mean of the result of the two simulations. With this method, accurate results are obtained for simulations on time steps of at least one year.

The above SDE reveals that the advantage of first transferring the non-linearity to the drift and then doing the freezing is that it keeps the stochastic part of the SDE unchanged and the only effect is a mean-reverting term. This should be compared with the strategy of a naive freezing of the original SDE, $dF = \omega(\beta, F_n)\sigma dW_t$, which gives nothing but a normal process. We could, of course, have frozen the coefficients in different ways such as $dF = F(\omega(\beta, F_n)/F_n)\sigma dW_t$ leading to a lognormal SDE. The point is that no matter how the stochastic part is frozen, it does not come close to the original SDE when using a single simulation step. This is in contrast to freezing the drift which allows long simulation steps, especially when using techniques such as the predictor-corrector method.

Although we transformed the SDE to become lognormal in the diffusion part, it is also possible to transform to any solvable SDE. The reason for choosing the lognormal version is that it is a particularly simple SDE that coincides with the original one in the limits $F \to 0$ and $\beta \to 0$. Furthermore, the transformation was such that F is positive if and only if G is positive and the approximate solution of G (through the predictor-corrector method) preserved the positivity.

The calibration can, for example, be done by approximating the process with a quadratic SDE. As the quadratic SDE has three free parameters, it is possible to match the level, the tilt and the curvature of the implied volatility curve at the forward F_0. The SDE and its approximating quadratic SDE are therefore in close agreement in a region around F_0 which implies that the calibration to ATM options is accurate for maturities of at least 10 years. The computations for determining the approximate quadratic process can be done through Taylor expansion around F_0:

$$\begin{aligned} dF &= \beta^{-1}\left(1 - e^{-\beta F}\right)\sigma dW_t \\ &\approx \left(\beta^{-1}\left(1 - e^{-\beta F_0}\right) + e^{-\beta F_0}(F - F_0) - \frac{1}{2}\beta e^{-\beta F_0}(F - F_0)^2\right)\sigma dW_t \\ &= \left((F - F_-)(1 - F/F_+)\right)\sigma' dW_t \end{aligned}$$

where

$$\begin{aligned} F_\pm &= F_0 + \beta^{-1}\left(1 \pm \sqrt{1 e^{\beta F_0} - 1}\right) \\ \sigma' &= \frac{1}{2}\beta e^{-\beta F_0} F_+ \sigma \end{aligned}$$

Bibliography

Ingersoll JE (1996) Valuing foreign exchange rate derivatives with a bounded exchange process. Rev Derivatives Res 1:159–181

Lipton A (2001) Mathematical methods for foreign exchange. World Scientific Publishing, Singapore

Luke YL (1962) Integrals of Bessel functions. McGraw-Hill, New York

Pitman JW (1975) One-dimensional brownian motion and the three-dimensional bessel process. Adv Appl Probab 7:511–526

Revuz D, Yor M (1999) Continuous Martingales and Brownian Motion. Springer, Berlin

Chapter 6
Local Volatility Models

One way to construct models more complex than the Black–Scholes model is to allow the volatility to depend on the current time and on the value of the underlying:

$$dF_t = \sigma(t, F_t) F_t dW_t$$

Models of this type are called *local volatility models*. By choosing the volatility $\sigma(t, F)$ appropriately it is possible to match the prices of any arbitrage free implied-volatility surface $\sigma_{\text{imp}}(T, K)$. The corresponding option pricing PDE has the form

$$U_t + \frac{1}{2}\sigma(t, F)^2 F^2 U_{FF} = 0$$

We investigate the relation between $\sigma_{\text{imp}}(T, K)$ and $\sigma(t, F)$. Obviously, for $\sigma(t, F) = \sigma(t)$ the introduction of a volatility weighted time

$$\tilde{t}(t) = \int_0^t \sigma(u)^2 du$$

immediately reveals that

$$\sigma_{\text{imp}}(T) = \sqrt{\frac{1}{T}\int_0^T \sigma(u)^2 du} \Leftrightarrow \sigma(T) = \sqrt{\sigma_{\text{imp}}(T)^2 + 2\sigma_{\text{imp}}(T)\sigma'_{\text{imp}}(T)T}$$

For the general case, we use Dupire's formula to express $\sigma(t, F)$ in terms of $\sigma_{\text{imp}}(T, K)$. The inverse relation, i.e. to express $\sigma_{\text{imp}}(T, K)$ as a function of $\sigma(t, F)$, is done by perturbative expansion techniques. Thanks to Dupire's formula, local volatility models are easy to calibrate and have become popular. Unfortunately, their dynamics are not in agreement with market behavior.

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_6,

6.1 ATM Perturbation

We solve the European call option problem

$$\begin{cases} U_t + \frac{1}{2}\sigma(F)^2 F^2 U_{FF} = 0 \\ U(t = T, F) = (F - K)_+ \end{cases}$$

by doing an expansion around the ATM point: $F = (1 + \epsilon x)K$, where ϵ is small. We obtain

$$U_t + \frac{1}{2}\frac{(1 + \epsilon x)^2}{\epsilon^2}\sigma((1 + \epsilon x)K)^2 U_{xx} = 0$$

The form of the PDE suggests that we should consider times to maturity $(T - t)$ of the order ϵ^2. We therefore change variables:

$$\begin{cases} \omega = \sigma(K)^2 \frac{T - t}{\epsilon^2} \\ \Phi(\omega, x) = U(t, F)/\epsilon K \end{cases}$$

which leads to

$$\begin{cases} \Phi_\omega = \frac{1}{2} g(\epsilon x)\Phi_{xx} \\ \Phi(\omega = 0, x) = x_+ \end{cases}$$

where

$$g(z) = \frac{\sigma((1 + z)K)^2}{\sigma(K)^2}(1 + z)^2$$

Taylor expanding g and using $g(0) = 1$ gives

$$\mathcal{D}\Phi = (g_1 \epsilon x + g_2 \epsilon^2 x^2 + \ldots)\Phi_{xx}, \quad \mathcal{D} = \partial_\omega - \frac{1}{2}\partial_x^2, \quad g_n = \frac{1}{2}\frac{1}{n!}\partial_z^n g|_{z=0}$$

Doing a perturbative expansion

$$\Phi = \Phi^0 + \epsilon \Phi^1 + \epsilon^2 \Phi^2 + \ldots$$

and equating equal powers of ϵ gives a chain of PDEs

$$\begin{cases} \mathcal{D}\Phi^0 = 0 \\ \mathcal{D}\Phi^1 = g_1 x \Phi^0_{xx} \\ \mathcal{D}\Phi^2 = g_1 x \Phi^1_{xx} + g_2 x^2 \Phi^0_{xx} \\ \ldots \end{cases}$$

with boundary conditions

$$\Phi^n(\omega = 0, x) = \begin{cases} x_+ & n = 0 \\ 0 & n > 0 \end{cases}$$

To solve the above equations, first note that

$$\Psi = \frac{1}{n+1}\omega^{n+1}\partial_x^m \partial_\omega^k \Omega$$

solves

$$\begin{cases} \mathcal{D}\Psi & = \omega^n \partial_x^m \partial_\omega^k \Omega \\ \Psi(\omega = 0) = 0 \end{cases}$$

if $\mathcal{D}\Omega = 0$ and $n \geq 0$. Indeed,

$$\begin{aligned} \mathcal{D}\Psi &= \left[\mathcal{D}, \frac{1}{n+1}\omega^{n+1}\partial_x^m \partial_\omega^k\right]\Omega \\ &= \left[\partial_\omega, \omega^{n+1}\right]\frac{1}{n+1}\partial_x^m \partial_\omega^k \Omega = \omega^n \partial_x^m \partial_\omega^k \Omega \end{aligned}$$

where $[A, B] = AB - BA$ is the commutator. The reason why we managed to find the solution in such a simple way was that the factor in front of Ω did not contain any powers of x but only derivatives of x. Because the right-hand side in the chain of PDEs contains powers of x, we must first find a way to convert these powers of x into powers of ω together with x- and ω-derivatives.

As Φ^0 is the solution to the normal SDE,

$$\Phi^0(\omega, x) = xN(x/\sqrt{\omega}) + \sqrt{\omega}n(x/\sqrt{\omega})$$

we can use the equations

$$\Phi^0 = x\Phi_x^0 + 2\omega\Phi_\omega^0 \Rightarrow x\Phi_{xx}^0 = -2\omega\Phi_{x\omega}^0$$

to convert powers of x in front of Φ_{xx}^0 and Φ_{xxx}^0 into powers of ω and ω-derivatives:

$$\begin{cases} x^n\Phi_{xx}^0 & = x^{n-1}x\Phi_{xx}^0 = -2\omega x^{n-1}\Phi_{x\omega}^0 = -\omega x^{n-1}\Phi_{xxx}^0 \\ x^{n-1}\Phi_{xxx}^0 & = 2x^{n-2}\partial_\omega x\Phi_x^0 = 2x^{n-2}\partial_\omega(\Phi^0 - 2\omega\Phi_\omega^0) \\ & = 2x^{n-2}(-\Phi_\omega^0 - 2\omega\Phi_{\omega\omega}^0) = -2(1 + 2\omega\partial_\omega)x^{n-2}\Phi_\omega^0 \\ & = -(1 + 2\omega\partial_\omega)x^{n-2}\Phi_{xx}^0 \end{cases}$$

$$\Rightarrow \begin{cases} x^{2n}\Phi_{xx}^0 & = (\omega + 2\omega^2\partial_\omega)^n\Phi_{xx}^0 \\ x^{2n+1}\Phi_{xx}^0 & = -(\omega + 2\omega^2\partial_\omega)^n\omega\Phi_{xxx}^0 \\ x^{2n}\Phi_{xxx}^0 & = (1 + 2\omega\partial_\omega)(\omega + 2\omega^2\partial_\omega)^{n-1}\omega\Phi_{xx}^0 \\ x^{2n+1}\Phi_{xxx}^0 & = -(1 + 2\omega\partial_\omega)(\omega + 2\omega^2\partial_\omega)^n\Phi_{xx}^0 \end{cases}$$

We are now prepared to solve the chain of PDEs. For Φ^1, we get

$$\mathcal{D}\Phi^1 = g_1 x \Phi^0_{xx} = -g_1 \omega \Phi^0_{xxx} \Rightarrow \Phi^1 = -\frac{1}{2} g_1 \omega^2 \Phi^0_{xxx} = -g_1 \omega^2 \Phi^0_{x\omega}$$

Using this result we can solve for Φ^2:

$$\begin{aligned}
\mathcal{D}\Phi^2 &= g_1 x \Phi^1_{xx} + g_2 x^2 \Phi^0_{xx} = -g_1^2 \omega^2 x \Phi^0_{xxx\omega} + g_2 x^2 \Phi^0_{xx} \\
&= g_1^2 \omega^2 \partial_\omega (1 + 2\omega\partial_\omega) \Phi^0_{xx} + g_2 (\omega + 2\omega^2 \partial_\omega) \Phi^0_{xx} \\
&= g_2 \omega \Phi^0_{xx} + (3g_1^2 + 2g_2) \omega^2 \Phi^0_{xx\omega} + 2g_1^2 \omega^3 \Phi^0_{xx\omega\omega} \\
\Rightarrow \Phi^2 &= \frac{1}{2} g_2 \omega^2 \Phi^0_{xx} + \left(g_1^2 + \frac{2}{3} g_2 \right) \omega^3 \Phi^0_{xx\omega} + \frac{1}{2} g_1^2 \omega^4 \Phi^0_{xx\omega\omega} \\
&= g_2 \omega^2 \Phi^0_{\omega} + 2 \left(g_1^2 + \frac{2}{3} g_2 \right) \omega^3 \Phi^0_{\omega\omega} + g_1^2 \omega^4 \Phi^0_{\omega\omega\omega}
\end{aligned}$$

The solution method can be summarized as follows: Assume that the chain of equations for Φ^0, Φ^1, ..., Φ^{n-1} has been solved and that the solutions have been expressed as a sum of terms of the canonical form $\omega^n \partial_x^m \partial_\omega^k \Phi^0$, where m equals 0 or 1. Inserting these solutions into the equation for Φ^n, it is possible to get rid of the x powers on the right-hand side of the PDE with the method described above. The resulting PDE can then be solved by taking the ω anti-derivative. Finally, using $\mathcal{D}\Phi^0 = 0$, the x derivatives can be converted to ω derivatives so that there remains no more than one x derivative. The expression for Φ^n is then of the canonical form and the method can be repeated to find the solutions of higher orders.

We now explicitly compute the solution to order ϵ^1 when $t = 0$. With

$$\Phi^1 = -g_1 \omega^2 \Phi^0_{x\omega} = \frac{1}{2} g_1 \omega x \Phi^0_{xx} = g_1 \omega x \Phi^0_{\omega}$$

we obtain

$$\begin{aligned}
&\Phi \approx \Phi^0 + \epsilon \Phi^1 = \Phi^0 + \epsilon g_1 \omega x \Phi^0_{\omega} \\
&g_1 = 1 + K \frac{\sigma'(K)}{\sigma(K)}, \quad \omega = \sigma(K)^2 \frac{T}{\epsilon^2} \quad x = \frac{1}{\epsilon}(F/K - 1)
\end{aligned}$$

which is approximately true for short maturity options with strike close to the forward. The expression for the implied volatility can be found by comparing with an option $\tilde{\Phi}$ that is priced with a strike-dependent lognormal volatility: $\tilde{\sigma}(F, K) = \sigma_{\text{imp}}(K)$ for some function σ_{imp}. The generalization to higher orders of ϵ is straightforward. As

$$
\begin{aligned}
\sigma_{\text{imp}}(K) &= \sigma_{\text{imp}}(F/(1+\epsilon x)) \approx \sigma_{\text{imp}}(F) - \epsilon x F \sigma'_{\text{imp}}(F) \\
\Rightarrow \tilde{\omega} &= \sigma_{\text{imp}}(K)^2 \frac{T}{\epsilon^2} \approx \left(\sigma_{\text{imp}}(F)^2 - 2\epsilon x F \sigma_{\text{imp}}(F)\sigma'_{\text{imp}}(F)\right) \frac{T}{\epsilon^2} \\
&= \tilde{\omega}_0 + \epsilon \tilde{\omega}_1
\end{aligned}
$$

and $g_1 = 1$ for a lognormal model, we obtain

$$
\begin{aligned}
\tilde{\Phi}(\tilde{\omega}, x) &\approx \Phi^0(\tilde{\omega}, x) + \epsilon \tilde{\omega} x \Phi^0_\omega(\tilde{\omega}, x) \\
&\approx \Phi^0(\tilde{\omega}_0 + \epsilon\tilde{\omega}_1, x) + \epsilon(\tilde{\omega}_0 + \epsilon\tilde{\omega}_1) x \Phi^0_\omega(\tilde{\omega}_0 + \epsilon\tilde{\omega}_1, x) \\
&\approx \Phi^0(\tilde{\omega}_0, x) + \epsilon(\tilde{\omega}_1 + \tilde{\omega}_0 x)\Phi^0_\omega(\tilde{\omega}_0, x)
\end{aligned}
$$

Comparing with the general formula above gives the lowest-order expression

$$
\Phi^0(\tilde{\omega}_0, x) = \Phi^0(\omega, x) \Leftrightarrow \tilde{\omega}_0 = \omega \Leftrightarrow \sigma_{\text{imp}}(F) = \sigma(K)
$$

and the first-order (in ϵ) expression

$$
\begin{aligned}
&\tilde{\omega}_1 + \tilde{\omega}_0 x = g_1 \omega x \Leftrightarrow \tilde{\omega}_1 = (g_1 - 1)\omega x \\
&\Leftrightarrow -2xF\sigma_{\text{imp}}(F)\sigma'_{\text{imp}}(F)\frac{T}{\epsilon^2} = K\frac{\sigma'(K)}{\sigma(K)}\sigma(K)^2\frac{T}{\epsilon^2}x \\
&\Leftrightarrow \sigma'_{\text{imp}}(F) = -\frac{K}{2F}\sigma'(K)
\end{aligned}
$$

We therefore finally arrive at

$$
\begin{aligned}
\sigma_{\text{imp}}(K) &\approx \sigma_{\text{imp}}(F) - \epsilon x F \sigma'_{\text{imp}}(F) \\
&= \sigma(K) + \epsilon x F \frac{K}{2F}\sigma'(K) \\
&\approx \sigma\left(K + \frac{1}{2}(F - K)\right) \\
&= \sigma\left(\frac{1}{2}(F + K)\right)
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\sigma_{\text{imp}}(F) &= \sigma(F) \\
\sigma'_{\text{imp}}(F) &= \frac{1}{2}\sigma'(F)
\end{aligned}
$$

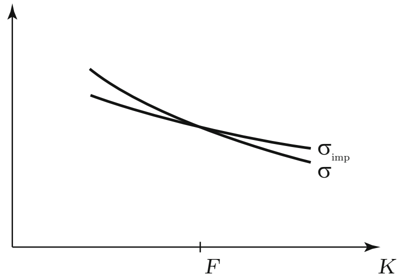

Fig. 6.1 Relation between the local volatility and the implied volatility skew for short maturities and close to ATM

We conclude that the implied volatility is equal to the local volatility at $K = F$, but the derivative is only half the size, see Fig. 6.1.

With a vanishing ATM skew, i.e. $\sigma'_{\text{imp}}(F) = 0$, the lowest order non-vanishing contribution is of order ϵ^2 and can be shown to be given by

$$\sigma_{\text{imp}}(K) \approx \sigma\left(F + \frac{1}{\sqrt{3}}(K - F)\right)$$

In particular, we obtain

$$\sigma''_{\text{imp}}(F) = \frac{1}{3}\sigma''(F)$$

which means that the curvature of the local volatility is three times the size of the ATM implied volatility smile.

Observe that we did not include boundary conditions in the computations. As Φ^0 is a solution to the normal PDE, the underlying can assume negative values. By imposing appropriate boundary conditions in the solution technique described above, it is possible to obtain a solution Φ that only supports positive values of the underlying. From a theoretical point of view, the existence of boundary conditions is irrelevant as they affect the distribution in the tails while the perturbative expansion technique is concerned with the ATM region. Different choices of boundary conditions do not therefore have a significant impact on the prices in the region of interest (unless the maturity is very long). From a practical point of view, however, boundary conditions can be important. When implementing a successful model, the users sooner or later use the model far away from the domain where it is supposed to be valid, in particular if there is no other suitable model. It is then important that even though we no longer have a state-of-the-art model in this region, the model should still give reasonable prices and be arbitrage-free.

We have so far assumed that $\sigma(t, F)$ is a function only of F. If it is of product form $\sigma(t, F) = \sigma_1(t)\sigma_2(F)$, the introduction of a volatility weighted time:

$$\tilde{t}(t) = \int_0^t \sigma_1^2(u)du$$

reduces the problem to the situation where $\sigma(t, F)$ only depends on F. A general dependence of σ on t and F can be solved with perturbation. The calculations are roughly the same as above but more cumbersome.

6.2 Dupire's Equation

Assume, as usual, that the forward option price can be obtained as an expectation under some SDE. With p the Green's function, we have

$$U(F, K, t, T) = \int (F_T - K)_+ p(T, F_T; t, F) dF_T$$

and

$$\frac{d^2}{dK^2} U(F, K, t, T) = p(T, K; t, F)$$

The right-hand side is non-negative as it is a probability. We conclude that the left-hand side is non-negative as well, which is equivalent with the no-arbitrage relation derived in Sect. 2.4. To investigate whether the no-arbitrage relation $\frac{d}{dT} V \geq 0$ is satisfied, consider the relation

$$\begin{aligned} & U(F, K, t, T + \Delta T) - U(F, K, t, T) \\ = & \int \int (F_{T+\Delta T} - K)_+ p(T + \Delta T, F_{T+\Delta T}; T, F_T) p(T, F_T; t, F) dF_T dF_{T+\Delta T} \\ & - \int (F_T - K)_+ p(T, F_T; t, F) dF_T \end{aligned}$$

which is positive since

$$\int (F_{T+\Delta T} - K)_+ p(T + \Delta T, F_{T+\Delta T}; T, F_T) dF_{T+\Delta T} \geq (F_T - K)_+$$

according to Jensen's inequality. We obtain the relation $\frac{d}{dT} U \geq 0$ which is slightly weaker than the no arbitrage relation $\frac{d}{dT} V \geq 0$.

Using the forward Kolmogorov equation gives

$$\begin{aligned} \frac{d}{dT} U &= \int (F_T - K)_+ \frac{d}{dT} p(T, F_T; t, F) dF_T \\ &= \int (F_T - K)_+ \frac{1}{2} \frac{d^2}{dF_T^2} \sigma(T, F_T)^2 F_T^2 p(T, F_T; t, F) dF_T \end{aligned}$$

$$
\begin{aligned}
&= \int \frac{1}{2}\sigma(T, F_T)^2 F_T^2 p(T, F_T; t, F) \frac{d^2}{dF_T^2}(F_T - K)_+ dF_T \\
&= \frac{1}{2}\sigma(T, K)^2 K^2 p(T, K; t, F) = \frac{1}{2}\sigma(T, K)^2 K^2 \frac{d^2}{dK^2} U(F, K, t, T)
\end{aligned}
$$

$$
\Leftrightarrow \sigma(T, K) = \sqrt{\frac{\frac{d}{dT}U}{\frac{1}{2}K^2 \frac{d^2}{dK^2}U}}
$$

This interesting relation by Dupire (1994) shows how the local volatility can be computed from European call option prices. According to the above discussion, the factors inside the square root are positive if the model is arbitrage-free. Expressing the price in terms of implied volatility gives after some computations the relation between the local volatility and the implied volatility:

$$
\begin{aligned}
&\sigma(T, K) = \\
&\sigma_{\text{imp}} \sqrt{\frac{1+2\frac{\sigma_{\text{imp},T}}{\sigma_{\text{imp}}}(T-t)}{(1+Kd_+\sigma_{\text{imp},K}\sqrt{T-t})(1+Kd_-\sigma_{\text{imp},K}\sqrt{T-t})+(K\sigma_{\text{imp},K}+K^2\sigma_{\text{imp},KK})\sigma_{\text{imp}}(T-t)}}
\end{aligned}
$$

This equation can be used for calibration since the local volatility surface can be computed from a given implied volatility surface. In practice, the implied volatilities are found in the market only for a discrete set of maturities and strikes. The implied volatility surface can be derived from the market quotes through a 2-dimensional interpolation. As Dupire's formula contains derivations of the implied volatility, the resulting local volatilities are highly dependent on the choice of interpolation scheme.

6.3 Short Maturity Expansion

With the variables

$$
\begin{cases} \tau = T - t \\ x = \ln(K/F) \end{cases}
$$

and the functions

$$
\begin{cases} \tilde{\sigma}(\tau, x) = \sigma(T, K) \\ I(\tau, x) = \sigma_{\text{imp}}(T, K) \end{cases}
$$

Dupire's equation takes the form

$$
I^2 + 2II_\tau \tau = \tilde{\sigma}^2 \left(\left(1 - x\frac{I_x}{I} \right)^2 - \frac{1}{4} I^2 I_x^2 \tau^2 + II_{xx}\tau \right)
$$

This non-trivial PDE can be solved by Taylor expanding I around $\tau = 0$:

$$I(\tau, x) = I_0(x) + I_1(x)\tau + I_2(x)\tau^2 + \dots$$

To the lowest order τ^0, we obtain

$$\begin{aligned} I &= \tilde{\sigma}_0 \left(1 - x\frac{I_0'}{I_0}\right) \\ \Rightarrow 1 &= \tilde{\sigma}_0 (I_0^{-1} + x\frac{d}{dx} I_0^{-1}) = \tilde{\sigma}_0 \frac{d}{dx}(x I_0^{-1}) \\ \Rightarrow I_0(x) &= \left(\frac{1}{x}\int_0^x \frac{dx'}{\tilde{\sigma}_0(x')}\right)^{-1} = \left(\int_0^1 \frac{ds}{\tilde{\sigma}_0(sx)}\right)^{-1} \end{aligned}$$

where $\tilde{\sigma}_0$ is the lowest order term in the Taylor expansion

$$\tilde{\sigma}(\tau, x) = \tilde{\sigma}_0(x) + \tilde{\sigma}_1(x)\tau + \tilde{\sigma}_2(x)\tau^2 + \dots$$

We conclude that the local volatility and the implied volatility are to the lowest order related by

$$\begin{cases} \sigma(K) &= \sigma_{\text{imp}}(K)\left(1 - K\ln(K/F)\frac{\sigma_{\text{imp},K}(K)}{\sigma_{\text{imp}}(K)}\right)^{-1} \\ \sigma_{\text{imp}}(K) &= \left(\int_0^1 \frac{ds}{\sigma(F^{1-s}K^s)}\right)^{-1} \end{cases}$$

It is straightforward to verify that this formula reduces to the expression

$$\sigma_{\text{imp}}(K) \approx \sigma((F + K)/2)$$

of Sect. 6.1 when $|K - F|$ is small. The formula in this section is more natural as it depends on the local volatilities in the whole range between K and F and not only on a single point $(F + K)/2$. For instance, we do not expect to see any significant change in the implied volatility if there is a spike (or dip) of the local volatility at this single point.

If $\sigma_{\text{imp}}'(F) = 0$, which is equivalent with $\sigma'(F) = 0$, the second derivative of the above expression for $\sigma_{\text{imp}}(K)$ gives

$$\sigma_{\text{imp}}''(F) = \frac{1}{3}\sigma''(F)$$

or equivalently

$$\sigma_{\text{imp}}(K) \approx \sigma\left(F + \frac{1}{\sqrt{3}}(K - F)\right)$$

for $|K - F|$ small, which was previously derived in Sect. 6.1.

I can be computed via induction by assuming that $I_0(x), \ldots, I_{n-1}(x)$ are known and deriving an expression for $I_n(x)$. We do the computation by inserting the Taylor expansion of I into the PDE and equating terms of order τ^n. As the terms that contain $I_k(x)$ with $k > n$ lead to higher order powers, the Taylor expansion can be truncated at $I_n(x)\tau^n$. The PDE then consists of two parts: terms containing I_n and terms containing $I_k(x)$, $k < n$. By induction, the latter part is known and can be summarized into a function $f(x)$. It remains to compute the terms that contain I_n and is of order τ^n. For instance,

$$\begin{aligned}
\left(1 - x\frac{I_x}{I}\right)^2\Bigg|_{\tau^n, I_n} &= \left(1 - x\frac{I_0' + \ldots + I_n'\tau^n}{I_0 + \ldots + I_n\tau^n}\right)^2\Bigg|_{\tau^n, I_n} \\
&= \left(1 - x\frac{I_0' + I_n'\tau^n}{I_0 + I_n\tau^n}\right)^2\Bigg|_{\tau^n, I_n} \\
&= \left(1 - \frac{x}{I_0}\left(I_0' + I_n'\tau^n\right)\left(1 - \frac{I_n}{I_0}\tau^n\right)\right)^2\Bigg|_{\tau^n, I_n} \\
&= -2\left(1 - x\frac{I_0'}{I_0}\right)x\frac{\tau^n}{I_0}\left(I_n' - \frac{I_n}{I_0}I_0'\right)
\end{aligned}$$

Equating the τ^n terms in the PDE gives then

$$2I_0I_n + 2nI_0I_n = \tilde{\sigma}_0^2\left(-2\left(1 - x\frac{I_0'}{I_0}\right)x\frac{\tau^n}{I_0}\left(I_n' - \frac{I_n}{I_0}I_0'\right)\right) + f(x)$$

which can be written as

$$I_n'(x) + g(x)I_n(x) = h(x)$$

for known functions $g(x)$ and $h(x)$. This ordinary differential equation can be solved by multiplying both sides with $\exp\left(\int_0^x g(x')dx'\right)$. The solution is

$$I_n(x) = e^{-\int_0^x g(x')dx'}\left(\int_0^x h(s)e^{\int_0^s g(x')dx'} + I_n(0)\right)$$

We leave it as an exercise for the reader to derive the expressions for the lowest order terms.

6.4 Dynamics

In the discussion of perturbative ATM expansion of local volatility models we showed that the implied volatility and the local volatility are to lowest order related by

$$\sigma_{\text{imp}}(K) = \sigma((F + K)/2)$$

where F is today's forward. Assume that market prices $\sigma_{\text{imp}}(K)$ are given and that we would like to calibrate the local volatility model to match these prices. This is done by setting

$$\sigma(z) = \sigma_{\text{imp}}(2z - F)$$

Assume then that we revisit the model at a later time, for example, tomorrow, when the forward has changed from F to $\tilde{F}$. The local volatility model predicts the market prices

$$\tilde{\sigma}_{\text{imp}}(K) = \sigma((\tilde{F} + K)/2) = \sigma_{\text{imp}}(K + \tilde{F} - F)$$

We see that the curve has been shifted sideways by an amount $\tilde{F} - F$. It means, for example, that when the forward increases, i.e. $\tilde{F} > F$, the implied volatility curve moves to the left. This is opposite to the expected market behavior. Observe that the argument is only true close to the ATM point and for small moves in the forward. In fact, it is easy to understand that it cannot be true for arbitrary points on the volatility curve as an arithmetic shift would lead to the existence of volatilities for negative values of the underlying. Nevertheless, the argument was not made to show the exact dependence of the implied volatility on the forward, but rather to show its qualitative behavior.

Local volatility models are useful in finance because of the easy fit to market data. Their volatility dynamics, however, are in the opposite direction to what is observed in the market. Local volatility models are for this reason often used in combination with other models. To illustrate the benefits of combining models, consider a model that has the opposite properties compared to local volatility models, i.e. it has good dynamics but is hard to fit to the market. Assume that we use this model and fit it to the market as well as we can. On top of this model we can then add a local volatility model that can be fitted to the remaining difference between market data and model data. By adding a local volatility model on top of a sticky-delta model, the result is somewhere between sticky-strike and sticky-delta behavior. As the market often has dynamics in this region, it comes down to choosing the appropriate portions of local volatility model and sticky-delta model.

Bibliography

Dupire B (1994) Pricing with a smile. Risk January:18–20

Chapter 7
Stochastic Volatility Models

A natural generalization of the Black–Scholes model is to allow the volatility to be stochastic. This is motivated by the fact that a historical analysis shows that the volatility indeed behaves as if it was stochastic. In this chapter we consider various techniques for solving stochastic volatility models. For optimal transparency, we focus on a particularly simple model

$$\begin{aligned} dF_t &= \sigma_t F_t dW_t \\ d\sigma_t &= \epsilon_t \sigma_t dZ_t \end{aligned}$$

where the Brownian motion W is of the form

$$W = \rho Z + \sqrt{1-\rho^2} Z^{\perp}$$

and $Z^{\perp}$ is a Brownian motion independent of Z. It is straightforward to generalize the solution techniques that we present to more general SDEs. For instance, instead of using a lognormal process, it is possible to use a shifted lognormal or a CEV process for the underlying. Furthermore, when pricing products that have payoffs at more than one instance in time, it is useful to add a mean-reverting drift term to the SDE for the volatility. The reason is that for most product types the time dependence of the variance of the volatility can be better matched with a mean-reverting process than with a lognormal process. Recall that as the forward can be written as a quotient of a tradable and a numeraire, the fundamental theorem of asset pricing implies that the forward process cannot have a drift term in the pricing measure. For the volatility, on the other hand, there is no such restriction.

The above equations are assumed to be formulated in the forward measure. It means that we need to take the expectation of $(F - K)_+$ to compute the European call option price. According to the Feynman-Kac theorem, the price can also be calculated with the corresponding backward Kolmogorov equation. The derivation of this PDE is as in Sect. 3.7. The result is

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_7,

$$U_t + \frac{1}{2}\sigma^2 F^2 U_{FF} + \rho\epsilon\sigma^2 F U_{F\sigma} + \frac{1}{2}\epsilon^2\sigma^2 U_{\sigma\sigma} = 0$$

We argued in Sect. 3.5 that $\sigma^2(T - t)$ is often much smaller than 1, which motivates a perturbative treatment. Similarly, restricting ourselves to contracts with short maturities, we can assume that $\epsilon^2(T - t) \ll 1$. Several of our arguments and methods in this chapter are therefore based on the assumption that $\epsilon^2(T-t)$ is small.

We would like to point out that the priority of the models in this book is not to explain reality. Instead, the aim is to obtain models that are flexible enough to match market data. Despite this, our choice of models is often based on how we believe that the reality behaves. This is because the reality can give us clues about how to model certain phenomena. A good example of this practice can be seen by the choice of stochastic volatility models, based on empirical properties of the underlying that implies an option skew and smile. In reality, the estimated historical volatility of volatility often only makes up a fraction (perhaps a quarter or half) of the implied volatility of the volatility. Other contributing factors to the skew and the smile include supply and demand, and fat tails coming from underestimated extreme events. Despite the fact that a large part of the smile is coming from other effects, it is common to use stochastic volatility models because they are simple to work with. However, as we pointed out in Sect. 4.2, although it is sometimes possible to use unrealistic processes for vanilla pricing, it can lead to serious mispricing of path-dependent derivatives.

It is also possible to build models based on the assumption that it is the implied volatility that is stochastic rather than the local volatility, see Hafner (2004) and references therein. We have chosen to omit any discussion of this model type as it is not as popular as the models in this chapter.

The focus of this chapter is on perturbative methods. We also treat the semi-analytic method of Fourier transforms, applicable to special types of stochastic volatility models. We compare the various techniques, discuss the dynamics and show a relation to local volatility models.

7.1 Skew and Smile

The origin of the skew and smile for stochastic volatility models can be understood by relatively simple arguments. We first argue for the existence of the skew and then for the smile. We also derive the form of the skew and smile for contracts with short maturity and with small volatility of volatility.

To explain the skew in stochastic volatility models, assume for concreteness that the correlation between the underlying and the volatility is negative and consider an ITM call option. Because of the vega profile, the option has its strongest dependence on the volatility around the ATM point. For the ITM option to end up in the ATM region, the underlying needs to decrease, which means an increase in the local volatility due to the anti-correlation. Thus, if the market moves in such a way that the option ends up where it has a strong dependence on the volatility, then the volatility has increased from its original value. This clearly implies a high implied volatility for ITM options. A similar argument proves that OTM options have low implied volatility. This argument explains the skew of stochastic volatility models. We also see that the larger the correlation, the steeper the skew.

To understand the smile effect, assume first that the correlation is zero. As the processes of the underlying and the volatility are independent, the European option price can be computed as

$$\begin{aligned} U &= E[(F-K)_+ | dF_t = \sigma F_t dW_t, d\sigma_t = \epsilon_t \sigma_t dZ_t] \\ &= E\left[U^0 \left(\int_0^T \sigma^2 dt \right) \middle| d\sigma_t = \epsilon_t \sigma_t dZ_t \right] \end{aligned}$$

where U^0 is the non-stochastic volatility price obtained from the Black–Scholes formula. According to Sect. 3.5, the Black–Scholes formula is a concave function of the volatility for strikes close to the ATM point. It means that the expectation in the above expression is lower than the non-stochastic volatility result, see Sect. 3.9. The same type of argument implies that the expectation is higher than the non-stochastic volatility result in the convex regions away from the ATM point. This argument shows that stochastic volatility produces volatility smiles. It also follows that the smile is more pronounced for larger volatility of volatility.

We now assume that $\epsilon^2(T-t) \ll 1$ and $\sigma^2(T-t) \ll 1$ to mathematically argue that stochastic volatility models imply both skew and smile. Observe that

$$\begin{aligned} &U(T) \approx U(t) + (T-t)U_t \\ &\qquad \Leftrightarrow U(t) \approx U(T) + \left(\frac{1}{2}\sigma^2 F^2 U_{FF} + \rho\epsilon\sigma^2 F U_{F\sigma} + \frac{1}{2}\epsilon^2\sigma^2 U_{\sigma\sigma} \right)(T-t) \end{aligned}$$

Viewing this equation to the lowest order in ϵ and $\sigma^2(T-t)$ makes it possible to replace the U on the right-hand side with the price U^0 of the corresponding non-stochastic volatility model, i.e. the Black–Scholes formula. Indeed, the contribution from the stochastic volatility only leads to higher-order terms. We then see that the effect of the stochastic volatility on the option price is through the option vanna $U^0_{F\sigma}$ and volga $U^0_{\sigma\sigma}$, while the effect from the non-stochastic volatility is through the gamma U^0_{FF}.

From Sect. 3.5 it follows that the second-order greeks are given by

$$\begin{cases} U^0_{FF} &= \dfrac{1}{F\sigma\sqrt{\tau}} n(d_+) \\ U^0_{F\sigma} &= -\dfrac{d_-}{\sigma} n(d_+) \\ U^0_{\sigma\sigma} &= \dfrac{F\sqrt{\tau}}{\sigma} d_+ d_- n(d_+) \end{cases} \qquad \tau = T-t, \quad d_\pm = \frac{\ln(F/K)}{\sigma\sqrt{\tau}} \pm \frac{1}{2}\sigma\sqrt{\tau}$$

with zeros at

$$\begin{cases} U^0_{F\sigma} = 0 \Leftrightarrow d_- = 0 \Leftrightarrow K = Fe^{-\sigma^2\tau/2} \\ U^0_{\sigma\sigma} = 0 \Leftrightarrow d_\pm = 0 \Leftrightarrow K = Fe^{\pm\sigma^2\tau/2} \end{cases}$$

The maxima of the greeks with respect to the strike can be computed by taking the derivative with respect to K:

$$\begin{cases} U^0_{FFK} = \dfrac{1}{KF\sigma^2\tau} d_+ n(d_+) \\ U^0_{F\sigma K} = \dfrac{n(d_+)}{K\sigma^2\sqrt{\tau}}(1 - d_+ d_-) \\ U^0_{\sigma\sigma K} = \dfrac{n(d_-)}{\sigma^2}(d_+^2 d_- - d_- - d_+) \end{cases}$$

with zeros at

$$\begin{cases} U^0_{FFK} = 0 \Leftrightarrow d_+ & = 0 \Leftrightarrow K = Fe^{\sigma^2\tau/2} \\ U^0_{F\sigma K} = 0 \Leftrightarrow 1 - d_+ d_- & = 0 \Leftrightarrow K = Fe^{\pm\sigma\sqrt{\tau}\sqrt{1+\frac{1}{4}\sigma^2\tau}} \\ U^0_{\sigma\sigma K} = 0 \Leftrightarrow d_+^2 d_- - d_- - d_+ = 0 & \end{cases}$$

The last equation can be solved by using that $\sigma^2(T-t)$ is small:

$$\begin{aligned} U^0_{\sigma\sigma K} = 0 &\Leftrightarrow (\ln(F/K) + \sigma^2\tau/2)^2(\ln(F/K) - \sigma^2\tau/2) - 2\ln(F/K)\sigma^2\tau = 0 \\ &\Leftrightarrow (\ln(F/K))^3 + \frac{1}{2}\sigma^2\tau(\ln(F/K))^2 - 2\sigma^2\tau\ln(F/K) \approx 0 \\ &\Leftrightarrow K \approx F \text{ or } Fe^{\pm\sigma\sqrt{2\tau}} \end{aligned}$$

From which we get the extrema

$$\begin{cases} U^0_{FF}|\text{extr.} = U^0_{FF}\big|_{K=Fe^{\sigma^2\tau/2}} = \dfrac{1}{\sqrt{2\pi\sigma^2\tau}F} \\ U^0_{F\sigma}|\text{extr.} = U^0_{F\sigma}\big|_{K=Fe^{\pm\sigma\sqrt{\tau}\sqrt{1+\frac{1}{4}\sigma^2\tau}}} \\ \qquad = -\dfrac{1}{\sigma}\left(\pm\sqrt{1+\tfrac{1}{4}\sigma^2\tau} - \sigma\sqrt{\tau}/2\right) n\left(\pm\sqrt{1+\tfrac{1}{4}\sigma^2\tau} + \sigma\sqrt{\tau}/2\right) \\ \qquad \approx \mp\dfrac{1}{\sqrt{2\pi\sigma^2 e}} \\ U^0_{\sigma\sigma}|\text{extr.} \approx U^0_{\sigma\sigma}\big|_{K\approx F} \quad \text{or} \quad {}_{K\approx K\approx Fe^{\pm\sigma\sqrt{2\tau}}} \\ \qquad = -\dfrac{F\sigma\tau^{3/2}}{4\sqrt{2\pi}}e^{-\sigma^2\tau/8} \quad \text{or} \quad \dfrac{F\sqrt{\tau}}{\sigma}\left(2 - \dfrac{1}{4}\sigma^2\tau\right) n\left(\pm\sqrt{2} + \dfrac{1}{2}\sigma\sqrt{\tau}\right) \\ \qquad \approx -\dfrac{F\sigma\tau^{3/2}}{4\sqrt{2\pi}} \quad \text{or} \quad 2\dfrac{F\sqrt{\tau}}{\sqrt{2\pi\sigma^2}}e^{-1} \end{cases}$$

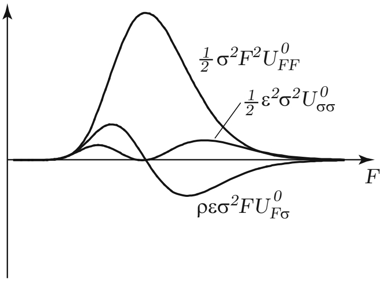

Fig. 7.1 Contribution to stochastic volatility from second-order greeks

In Fig. 7.1 we plot the three terms that are used in expansion of $U(t)$. The term proportional to the vanna gives skew around the ATM while the term proportional to the volga gives a smile effect around the ATM. We also see that the non-stochastic volatility effect from gamma is much larger than the stochastic volatility effect. Furthermore, as long as the correlation is not too close to zero, the skew effect from the vanna is much larger than the smile effect from the volga. Indeed, the gamma term is of order ϵ^0, the vanna term of order ϵ^1 and the volga term of order ϵ^2. Thus, there is a second-order contribution from the vanna comparable to the volga term which we neglected by only looking at the lowest-order effect. This effect and other higher-order contributions are covered in the next section.

Recall from Sect. 4.3 that the time decay in the option price was compensated by the gamma gain from the delta hedge. For stochastic volatility models we see from the differential equation that the time decay is different and is now compensated by three terms: the gamma gain, the vanna gain and the volga gain. The gamma gain comes from the non-stochastic volatility effect that the underlying is sold at high values and bought at low values in the delta hedge. The vanna and volga gain comes from the volatility hedge.

7.2 Perturbation for Small Volatility of Volatility

The European call option problem

$$\begin{cases} U_t + \dfrac{1}{2}\sigma^2 F^2 U_{FF} + \rho\epsilon\sigma^2 F U_{F\sigma} + \dfrac{1}{2}\epsilon^2\sigma^2 U_{\sigma\sigma} = 0 \\ U(t = T) = (F - K)_+ \end{cases}$$

can through a change of coordinates

$$\begin{cases} \Phi = U/K \\ x \ = \ln(F/K) \\ \tau \ = T - t \end{cases}$$

be made dimensionless

$$\begin{cases} \mathcal{D}\Phi = \epsilon\rho\sigma^2\partial_{x\sigma}\Phi + \frac{1}{2}\epsilon^2\sigma^2\partial_\sigma^2\Phi, \quad \mathcal{D} = \partial_\tau - \frac{1}{2}\sigma^2(\partial_x^2 - \partial_x) \\ \Phi(\tau = 0, x) = (e^x - 1)_+ \end{cases}$$

A perturbative expansion in ϵ,

$$\Phi = \Phi^0 + \epsilon\Phi^1 + \epsilon^2\Phi^2 + ...$$

gives a chain of PDEs by equating equal powers of ϵ:

$$\mathcal{D}\Phi^n = \rho\sigma^2\partial_{x\sigma}\Phi^{n-1} + \frac{1}{2}\sigma^2\partial_\sigma^2\Phi^{n-2}, \Phi^{-1} = 0 = \Phi^{-2}$$

with boundary conditions given by

$$\Phi^n(\tau = 0, x) = \begin{cases} (e^x - 1)_+ & n = 0 \\ 0 & n > 0 \end{cases}$$

We conclude that Φ^0 depends on σ and on τ only through the combination $\sigma^2\tau$. Indeed, Φ^0 is the standard Black–Scholes solution. As $\Phi^0 = \Phi^0(\sigma^2\tau)$, we obtain

$$\partial_\sigma\Phi^0 = \frac{2\tau}{\sigma}\partial_\tau\Phi^0 = \sigma\tau(\partial_x^2 - \partial_x)\Phi^0$$

where we used the PDE for Φ^0 in the last step.

We now show that the solution can be written of the form

$$\Phi^n = \sum_{i,j,k\geq 0} c_{i,j,k}^n \tau^i \sigma^j \partial_x^k \Phi^0$$

Clearly, this is true for $n = 0$. We now assume it to be true for $n - 2$ and $n - 1$ and prove it for n. The general case follows by induction. Consider first the situation when the right-hand side of the PDE is of this form. We are then interested in solving

$$\mathcal{D}\Psi = \sum_{ijk} c_{i,j,k}\tau^i\sigma^j\partial_x^k\Phi^0$$

A solution to this equation that satisfies $\Psi(\tau = 0) = 0$ is given by

$$\Psi = \sum_{ijk} \frac{1}{i} c_{i-1,j,k}\tau^i\sigma^j\partial_x^k\Phi^0$$

Indeed,

$$\begin{aligned}\mathcal{D}\Psi &= \sum_{ijk} \frac{1}{i} c_{i-1,j,k} \mathcal{L} \tau^i \sigma^j \partial_x^k \Phi^0 \\ &= \sum_{ijk} \frac{1}{i} c_{i-1,j,k} [\mathcal{L}, \tau^i] \sigma^j \partial_x^k \Phi^0 + \tau^i \sigma^j \partial_x^k \mathcal{L} \Phi^0 = \sum_{ijk} c_{i,j,k} \tau^i \sigma^j \partial_x^k \Phi^0\end{aligned}$$

We now assume that

$$\Phi^m = \sum_{ijk} c^m_{i,j,k} \tau^i \sigma^j \partial_x^k \Phi^0, \quad m = 0, 1, ..., n-1$$

and show that Φ^n is of this form. The first thing to be done is to compute the terms on the right-hand side of the PDE:

$$\begin{aligned}\rho\sigma^2 \partial_{x\sigma} \Phi^{n-1} &= \sum_{ijk} \rho c^{n-1}_{i,j,k} \tau^i \sigma^2 \partial_\sigma \sigma^j \partial_x^{k+1} \Phi^0 \\ &= \sum_{ijk} \rho c^{n-1}_{i,j,k} \tau^i (j\sigma^{j+1} \partial_x^{k+1} + \sigma^{j+2} \partial_x^{k+1} \sigma\tau(\partial_x^2 - \partial_x)) \Phi^0 \\ &= \sum_{ijk} \rho c^{n-1}_{i,j,k} (j\tau^i \sigma^{j+1} \partial_x^{k+1} + \tau^{i+1} \sigma^{j+3} \partial_x^{k+3} - \tau^{i+1} \sigma^{j+3} \partial_x^{k+2}) \Phi^0 \\ &= \sum_{ijk} \rho \left((j-1) c^{n-1}_{i,j-1,k-1} + c^{n-1}_{i-1,j-3,k-3} - c^{n-1}_{i-1,j-3,k-2} \right) \tau^i \sigma^j \partial_x^k \Phi^0 \\ \frac{1}{2}\sigma^2 \partial_\sigma^2 \Phi^{n-2} &= \sum_{ijk} \frac{1}{2} c^{n-2}_{i,j,k} \tau^i \sigma^2 \partial_\sigma^2 \sigma^j \partial_x^k \Phi^0 \\ &= \sum_{ijk} \frac{1}{2} c^{n-2}_{i,j,k} \tau^i \left(j(j-1)\sigma^j \partial_x^k + 2j\sigma^{j+1} \partial_x^k \sigma\tau(\partial_x^2 - \partial_x) \right. \\ &\quad \left. + \sigma^{j+2} \partial_x^k \partial_\sigma \sigma\tau(\partial_x^2 - \partial_x) \right) \Phi^0 \\ &= \sum_{ijk} \frac{1}{2} c^{n-2}_{i,j,k} \left(j(j-1)\tau^i \sigma^j \partial_x^k + (2j+1)\tau^{i+1} \sigma^{j+2} (\partial_x^{k+2} - \partial_x^{k+1}) \right. \\ &\quad \left. + \tau^{i+2} \sigma^{j+4} (\partial_x^{k+4} - 2\partial_x^{k+3} + \partial_x^{k+2}) \right) \Phi^0 \\ &= \sum_{ijk} \left(\frac{1}{2} j(j-1) c^{n-2}_{i,j,k} + (j-3/2) c^{n-2}_{i-1,j-2,k-2} \right. \\ &\quad - (j-3/2) c^{n-2}_{i-1,j-2,k-1} + \frac{1}{2} c^{n-2}_{i-2,j-4,k-4} \\ &\quad \left. - c^{n-2}_{i-2,j-4,k-3} + \frac{1}{2} c^{n-2}_{i-2,j-4,k-2} \right) \tau^i \sigma^j \partial_x^k \Phi^0\end{aligned}$$

which gives

$$\Phi^n = \sum_{ijk} c^n_{i,j,k} \tau^i \sigma^j \partial_x^k \Phi^0$$

with

$$\begin{aligned} i c^n_{i,j,k} = \rho \Big((j-1) c^{n-1}_{i-1,j-1,k-1} + c^{n-1}_{i-2,j-3,k-3} - c^{n-1}_{i-2,j-3,k-2} \Big) \\ \frac{1}{2} j(j-1) c^{n-2}_{i-1,j,k} + (j-3/2) c^{n-2}_{i-2,j-2,k-2} - (j-3/2) c^{n-2}_{i-2,j-2,k-1} \\ + \frac{1}{2} c^{n-2}_{i-3,j-4,k-4} - c^{n-2}_{i-3,j-4,k-3} + \frac{1}{2} c^{n-2}_{i-3,j-4,k-2} \end{aligned}$$

We now make explicit use of the formula by computing the lowest order terms Φ^1 and Φ^2. As $c^0_{0,0,0} = 1$ is the only non-zero coefficient for $n = 0$, we conclude that there are only two non-zero coefficients for $n = 1$: $c^1_{2,3,3} = \rho/2$ and $c^1_{2,3,2} = -\rho/2$. It means that

$$\Phi^1 = \frac{1}{2} \rho \tau^2 \sigma^3 (\partial_x^3 - \partial_x^2) \Phi^0$$

and in the same way we see that

$$\begin{aligned} \Phi^2 &= \Big(\frac{1}{3} \rho 3 \frac{1}{2} \rho \tau^3 \sigma^4 \partial_x^4 + \frac{1}{4} \rho \frac{1}{2} \rho \tau^4 \sigma^6 \partial_x^6 - \frac{1}{4} \rho \frac{1}{2} \rho \tau^4 \sigma^6 \partial_x^5 \\ &\quad - \frac{1}{3} \rho 3 \frac{1}{2} \rho \tau^3 \sigma^4 \partial_x^3 - \frac{1}{4} \rho \frac{1}{2} \rho \tau^4 \sigma^6 \partial_x^5 + \frac{1}{4} \rho \frac{1}{2} \rho \tau^4 \sigma^6 \partial_x^4 \\ &\quad + \frac{1}{2} \frac{1}{2} \tau^2 \sigma^2 \partial_x^2 - \frac{1}{2} \frac{1}{2} \tau^2 \sigma^2 \partial_x \\ &\quad + \frac{1}{3} \frac{1}{2} \tau^3 \sigma^4 \partial_x^4 - \frac{1}{3} \tau^3 \sigma^4 \partial_x^3 + \frac{1}{3} \frac{1}{2} \tau^3 \sigma^4 \partial_x^2 \Big) \Phi^0 \\ &= \Big(\frac{1}{2} \rho^2 \tau^3 \sigma^4 (\partial_x^4 - \partial_x^3) + \frac{1}{8} \rho^2 \tau^4 \sigma^6 (\partial_x^6 - 2\partial_x^5 + \partial_x^4) \\ &\quad + \frac{1}{4} \tau^2 \sigma^2 (\partial_x^2 - \partial_x) + \frac{1}{6} \tau^3 \sigma^4 (\partial_x^4 - 2\partial_x^3 + \partial_x^2) \Big) \Phi^0 \end{aligned}$$

Using the recursive formula for c, Φ^n can be computed for arbitrary n. In fact, it is straightforward to implement the formula in a computer to calculate Φ up to arbitrary powers.

The implied volatility from these lowest-order terms can be computed through a Taylor expansion

$$\sigma_{\text{imp}} \approx \sigma_{\text{imp},0} + \epsilon \sigma_{\text{imp},1}$$

We equate $\Phi(\sigma)$ and $\Phi^0(\sigma_{\text{imp}})$:

$$\Phi = \Phi^0(\sigma_{\text{imp}}) \approx \Phi^0(\sigma_{\text{imp},0}) + \Phi^0_\sigma(\sigma_{\text{imp},0})\epsilon\sigma_{\text{imp},1}$$

and

$$\Phi \approx \Phi^0(\sigma) + \epsilon\Phi^1(\sigma) = \Phi^0(\sigma) + \frac{1}{2}\epsilon\rho\tau^2\sigma^3(\partial_x^3 - \partial_x^2)\Phi^0 = \Phi^0(\sigma) + \frac{1}{2}\epsilon\rho\tau\sigma^2\Phi^0_{x\sigma}$$

up to order ϵ^1. Equating the terms with and without epsilon, respectively, gives

$$\begin{cases} \sigma_{\text{imp},0} = \sigma \\ \sigma_{\text{imp},1} = \dfrac{1}{2}\rho\sigma^2\tau\dfrac{\Phi^0_{x\sigma}}{\Phi^0_\sigma} \end{cases}$$

Using

$$\begin{aligned} \Phi^0 &= e^x N(d_+) - N(d-), \quad d_\pm = \frac{x}{\sigma\sqrt{\tau}} \pm \frac{1}{2}\sigma\sqrt{\tau} \\ \Rightarrow \Phi^0_\sigma &= n(d_-)\sqrt{\tau} \\ \Rightarrow \Phi^0_{x\sigma} &= -\frac{d_-}{\sigma}n(d_-) \end{aligned}$$

we obtain

$$\sigma_{\text{imp},1} = -\frac{1}{2}\rho\sigma\sqrt{\tau}d_- = -\frac{1}{2}\rho(x - \sigma^2\tau/2)$$

which means that the implied volatility has the lowest-order expression

$$\sigma_{\text{imp}} \approx \sigma - \frac{1}{2}\epsilon\rho(x - \sigma^2\tau/2)$$

Stochastic volatility models are often used with $\rho = 0$. The skew from the correlation is then lost but can be regained by using a non-lognormal process for the underlying, with the consequence of different dynamics. For the zero correlation case, Φ^2 is the lowest-order contribution. We then have

$$\begin{aligned} \Phi^2 &= \left(\frac{1}{4}\tau^2\sigma^2(\partial_x^2 - \partial_x) + \frac{1}{6}\tau^3\sigma^4(\partial_x^2 - \partial_x)^2\right)\Phi^0 \\ &= \frac{1}{4}\tau\sigma\Phi^0_\sigma + \frac{1}{6}\tau^2\sigma^3(\partial_x^2 - \partial_x)\partial_\sigma\Phi^0 \\ &= \frac{1}{4}\tau\sigma\Phi^0_\sigma + \frac{1}{6}\tau^2\partial_\sigma\left(\sigma^3(\partial_x^2 - \partial_x)\Phi^0\right) - \frac{1}{2}\tau^2\sigma^2\left(\partial_x^2 - \partial_x\right)\Phi^0 \\ &= \frac{1}{4}\tau\sigma\Phi^0_\sigma + \frac{1}{6}\tau\partial_\sigma\left(\sigma^2\Phi^0\right) - \frac{1}{2}\tau\sigma\Phi^0_\sigma \\ &= \frac{1}{12}\tau\sigma\Phi^0_\sigma + \frac{1}{6}\tau\sigma^2\Phi^0_{\sigma\sigma} \end{aligned}$$

$$\Rightarrow \Phi \approx \Phi^0 + \frac{1}{12}\epsilon^2\tau\sigma\Phi^0_\sigma + \frac{1}{6}\epsilon^2\tau\sigma^2\Phi^0_{\sigma\sigma}$$

To find the implied volatility we make a Taylor expansion

$$\begin{aligned}
\sigma_{\text{imp}} &\approx \sigma_{\text{imp},0} + \epsilon\sigma_{\text{imp},1} + \epsilon^2\sigma_{\text{imp},2}\\
\Rightarrow \Phi^0(\sigma_{\text{imp}}) &\approx \Phi^0(\sigma_{\text{imp},0}) + \Phi^0_\sigma(\sigma_{\text{imp},0})\epsilon\sigma_{\text{imp},1} + \Phi^0_\sigma(\sigma_{\text{imp},0})\epsilon^2\sigma_{\text{imp},2}\\
&\quad + \frac{1}{2}\Phi^0_{\sigma\sigma}(\sigma_{\text{imp},0})\epsilon^2\sigma^2_{\text{imp},1}
\end{aligned}$$

Equating equal powers of ϵ in the equation $\Phi(\sigma) = \Phi^0(\sigma_{\text{imp}})$ gives

$$\begin{cases}
\sigma_{\text{imp},0} = \sigma\\
\sigma_{\text{imp},1} = 0\\
\sigma_{\text{imp},2} = \dfrac{\frac{1}{12}\tau\sigma\Phi^0_\sigma + \frac{1}{6}\tau\sigma^2\Phi^0_{\sigma\sigma}}{\Phi^0_\sigma} = \dfrac{1}{6}\sigma\tau\left(\sigma\dfrac{\Phi^0_{\sigma\sigma}}{\Phi^0_\sigma} + \dfrac{1}{2}\right)
\end{cases}$$

Using

$$\begin{aligned}
\Phi^0_\sigma &= n(d_-)\sqrt{\tau}\\
\Rightarrow \Phi^0_{\sigma\sigma} &= \frac{\sqrt{\tau}}{\sigma}d_+d_-n(d_-) = \frac{\sqrt{\tau}}{\sigma}\left(\frac{x^2}{\sigma^2\tau} - (\sigma^2\tau/4)\right)n(d_-)
\end{aligned}$$

we obtain

$$\begin{aligned}
\sigma_{\text{imp},2} &= \frac{1}{6\sigma}\left(x^2 + \sigma^2\tau/2 - (\sigma^2\tau/2)^2\right)\\
\Rightarrow \sigma_{\text{imp}} &\approx \sigma\left(1 + \frac{\epsilon^2}{6\sigma^2}\left(x^2 + \sigma^2\tau/2 - (\sigma^2\tau/2)^2\right)\right)
\end{aligned}$$

Observe that the term containing ϵ^2 can be viewed as a correction term only if ϵ is smaller than σ or if the option is close to ATM. Unfortunately, this is far from always the case.

The general second-order expression, with non-zero correlation, can be computed in the same way. The result is

$$\sigma_{\text{imp}} \approx \sigma - \frac{1}{2}\epsilon\rho(x-\omega) + \frac{\epsilon^2}{6\sigma}\left(x^2 + \omega - \omega^2\right) + \frac{\epsilon^2}{2\sigma}\rho^2\left(-\frac{1}{2}x^2 - \frac{1}{2}x\omega + \omega^2\right)$$

where $\omega = \sigma^2\tau/2$. The method we have presented here is easily extendable to other processes such as the shifted lognormal process or the CEV process. The CEV process becomes particularly simple to include with the interpolation approach described at the end of Sect. 5.8.

7.3 Conditional Expectation Approach

The method described here works only for zero correlation, which means that the processes of the underlying and the volatility are independent. The expectation of the payoff can then be done iteratively: the expectation is first taken over the underlying and then over the volatility. As we know that the expectation with respect to the underlying is given by the Black–Scholes formula, it remains to compute the expectation of the Black–Scholes formula with respect to the volatility. This is usually done by only considering the lowest-order terms in ϵ. Before doing the actual computation, note that

$$\int_0^T Z_t dt = \int_0^T (d(Z_t t) - t dZ_t) = Z_T T - \int_0^T t dZ_t = \int_0^T (T-t) dZ_t$$

is a normal distribution with mean 0 and variance

$$\int_0^T (T-t)^2 dt = \frac{1}{3} T^3$$

which implies that

$$E\left[\left(\int_0^T Z_t dt\right)^2\right] = \frac{1}{3} T^3$$

Let $U^0(\cdot)$ denote the dependence of the Black–Scholes call formula on the volatility. We then have

$$\begin{aligned}
U &= E[(F-K)_+ | dF_t = \sigma_t F_t dW_t, d\sigma_t = \epsilon \sigma_t dZ_t, dW_t dZ_t = 0] \\
&= E\left[U^0\left(\sqrt{\frac{1}{T}\int_0^T \sigma_t^2 dt}\right) \middle| d\sigma_t = \epsilon\sigma_t dZ_t \right] \\
&= E\left[U^0\left(\sigma\sqrt{\frac{1}{T}\int_0^T e^{2\epsilon Z_t - \epsilon^2 t} dt}\right)\right] \\
&\approx E\left[U^0\left(\sigma\sqrt{\frac{1}{T}\int_0^T (1 + 2\epsilon Z_t - \epsilon^2 t + 2\epsilon^2 Z_t^2) dt}\right)\right] \\
&\approx E\left[U^0\left(\sigma + \frac{\epsilon\sigma}{T}\int_0^T Z_t dt + \frac{\epsilon^2\sigma}{T}\int_0^T (Z_t^2 - t) dt \right.\right. \\
&\quad \left.\left. + \frac{1}{4}\epsilon^2 \sigma T - \frac{\epsilon^2 \sigma}{2T^2}\left(\int_0^T Z_t dt\right)^2\right)\right]
\end{aligned}$$

$$\approx E\left[U^0 + \frac{\epsilon\sigma}{T}\int_0^T Z_t dt U_\sigma^0 + \frac{\epsilon^2\sigma}{T}\int_0^T (Z_t^2 - t) dt U_\sigma^0 + \frac{1}{4}\epsilon^2\sigma T U_\sigma^0 \right.$$
$$\left. -\frac{\epsilon^2\sigma}{2T^2}\left(\int_0^T Z_t dt\right)^2 U_\sigma^0 + \frac{\epsilon^2\sigma^2}{2T^2}\left(\int_0^T Z_t dt\right)^2 U_{\sigma\sigma}^0\right]$$
$$= U^0 + \frac{1}{12}\epsilon^2\sigma T U_\sigma^0 + \frac{1}{6}\epsilon^2\sigma^2 T U_{\sigma\sigma}^0$$

The price at an arbitrary time can be obtained by replacing T with $\tau = T - t$. As expected, this is identical to the zero-correlation result of the previous section.

7.4 Fourier Transform Approach

With this approach, we obtain an expression for the option price that is valid for arbitrary values of the parameters in the model. Thus, the method is not based on that any of the parameters are small. The result is an expression for the price as a 1-dimensional integral in the complex plane over a closed-form function.

The European call option price can be written as

$$U = E[(F - K)_+] = E[F_T \mathbb{1}_{F_T > K}] - KE[\mathbb{1}_{F_T > K}]$$

The first term can be computed by changing the numeraire from P_{tT} to S_t. Using the martingale measure P^* corresponding to the numeraire S_t gives that

$$E[F_T \mathbb{1}_{F_T > K}] = \frac{S_0}{P_{0T}} E^*\left[\frac{P_{TT}}{S_T} F_T \mathbb{1}_{F_T > K}\right] = F_0 E^*\left[\mathbb{1}_{F_T > K}\right]$$

according to the rule of measure change, see the Appendix. It then follows that

$$U = F_0 P^*(x_T > 0) - KP(x_T > 0)$$

where $x_t = \ln(F_t/K)$.

Using calculus of residues, the Heaviside function can be expressed as

$$\theta(x') = \frac{1}{2} + \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{ikx'}}{ik} dk$$

from which we obtain

$$
\begin{aligned}
P(x_T > 0) &= \int_{-\infty}^{\infty} p(x')\theta(x')dx' \\
&= \frac{1}{2} + \frac{1}{2\pi}\int_{-\infty}^{\infty} p(x')\left(\int_0^{\infty}\frac{e^{ikx'}}{ik}dk - \int_0^{\infty}\frac{e^{-ikx'}}{ik}dk\right)dx' \\
&= \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty}\operatorname{Re}\left(\frac{\hat{p}(k)}{ik}\right)dk
\end{aligned}
$$

where $\hat{p}$ is the Fourier transform of the PDF:

$$
\hat{p}(k) = \int_{-\infty}^{\infty} p(x')e^{ikx'}dx' = E[e^{ikx}]
$$

The computation of the first term in the expression for U is similar, but a measure change is needed:

$$
\hat{p}^*(k) = E^*\left[e^{ikx_T}\right] = \frac{P_{0T}}{S_0}E\left[\frac{S_T}{P_{TT}}e^{ikx_T}\right] = \frac{1}{F_0}E\left[F_T e^{ikx_T}\right] = \frac{K}{F_0}E\left[e^{(1+ik)x_T}\right]
$$

The results so far can be summarized as:

The European option price can be computed by

$$
\begin{aligned}
U &= F_0 P^1 - K P^0 \\
P^m &= \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty}\operatorname{Re}\left(\frac{\hat{p}_m(k)}{ik}\right)dk \\
\hat{p}_m(k) &= M_m E\left[e^{(m+ik)x_T}\right], \quad M_0 = 1, M_1 = K/F_0
\end{aligned}
$$

As we see below, $\hat{p}_m(k)$ can often be written as a closed-form expression. Therefore, all that remains to price options is to compute the above integrals. Note that at this stage we have not made any assumptions on the model, i.e. the above result is valid for the lognormal model, the CEV model, stochastic volatility models, etc. We now compute $\hat{p}_m$ when the underlying follows a lognormal process.

$$
dx_t = \sigma dW_t - \frac{1}{2}\sigma^2 dt \Leftrightarrow x_T = x_0 + \int_0^T \sigma dW_t - \frac{1}{2}\int_0^T \sigma^2 dt
$$

gives

$$
\begin{aligned}
\hat{p}_m(k) &= M_m e^{(m+ik)x_0} E\left[e^{(m+ik)\left(\int_0^T \sigma dW_t - \frac{1}{2}\int_0^T \sigma^2 dt\right)}\right] \\
&= M_m e^{(m+ik)x_0} E\left[e^{(m+ik)\left(\rho\int_0^T \sigma dZ_t + \sqrt{1-\rho^2}\int_0^T \sigma dZ_t^{\perp} - \frac{1}{2}\int_0^T \sigma^2 dt\right)}\right]
\end{aligned}
$$

As σ is independent of $Z^{\perp}$, it is possible to compute the expectation with respect to this variable. It is done by using the fact that

$$E\left[\exp\left(\int_0^T \sigma_t dZ_t^{\perp}\right)\right] = \exp\left(\frac{1}{2}\int_0^T \sigma_t^2 dt\right)$$

for σ_t any process independent of $Z_t^{\perp}$. The equality follows from

$$\begin{aligned} E\left[\left(\int_0^T \sigma_t dZ_t^{\perp}\right)^{2n}\right] &= E\left[\left(\sqrt{\int_0^T \sigma_t^2 dt X}\right)^{2n}\right] \\ &= \left(\int_0^T \sigma_t^2 dt\right)^n E\left[X^{2n}\right] = \frac{(2n)!}{n!2^n}\left(\int_0^T \sigma_t^2 dt\right)^n, \quad X \sim \mathcal{N}(0,1) \\ \Rightarrow \sum_{n=0}^{\infty} \frac{1}{(2n)!} E\left[\left(\int_0^T \sigma_t dZ_t^{\perp}\right)^{2n}\right] &= \sum_{n=0}^{\infty} \frac{1}{n!}\frac{1}{2^n}\left(\int_0^T \sigma_t^2 dt\right)^n \end{aligned}$$

and because the expectation of an odd number of factors on the left-hand side is equal to zero. We arrive at

$$\begin{aligned} \hat{p}_m(k) &= M_m e^{(m+ik)x_0} E\left[e^{(m+ik)\left(\rho\int_0^T \sigma dZ_t + \frac{1}{2}(m+ik)(1-\rho^2)\int_0^T \sigma^2 dt - \frac{1}{2}\int_0^T \sigma^2 dt\right)}\right] \\ &= M_m e^{(m+ik)x_0} E\left[e^{(m+ik)\rho\int_0^T \sigma dZ_t + \frac{1}{2}(m+ik)((m+ik)(1-\rho^2)-1)\int_0^T \sigma^2 dt}\right] \end{aligned}$$

The computations up to this stage are rather general. A lognormal process has been used for the underlying but no assumptions have been made for the volatility process. For simplicity, we assume that the volatility satisfies a particularly simple SDE:

$$\begin{aligned} d\sigma_t &= \epsilon dZ_t \\ \Rightarrow \int_0^T \sigma_t dZ_t &= \epsilon^{-1}\int_0^T \sigma_t d\sigma_t = \frac{1}{2}\epsilon^{-1}\int_0^T d\sigma_t^2 - \frac{1}{2}\epsilon\int_0^T dt \\ &= \frac{1}{2}\epsilon^{-1}(\sigma_T^2 - \sigma_0^2) - \frac{1}{2}\epsilon T \end{aligned}$$

which gives that

$$\begin{aligned} \hat{p}_m(k) &= M_m e^{(m+ik)(x_0 - \rho\epsilon^{-1}\sigma_0^2/2 - \rho\epsilon T/2)} E\left[e^{a\sigma_T^2 + b\int_0^T \sigma^2 dt}\right], \\ a &= \frac{1}{2}(m+ik)\rho\epsilon^{-1}, b = \frac{1}{2}(m+ik)((m+ik)(1-\rho^2)-1) \end{aligned}$$

The expectation can be computed by using the Feynman-Kac theorem and solving the resulting PDE:

$$\begin{cases} g_\tau &= \frac{1}{2}\epsilon^2 g_{\sigma\sigma} + b\sigma^2 g, \quad \tau = T - t \\ g(\tau = 0, \sigma) &= e^{a\sigma^2} \end{cases}$$

Based on the form of the initial condition, we make the substitution $g = e^f$ to transform the problem into

$$\begin{cases} f_\tau &= \frac{1}{2}\epsilon^2 f_{\sigma\sigma} + \frac{1}{2}\epsilon^2 f_\sigma^2 + b\sigma^2 \\ f(\tau = 0, \sigma) &= a\sigma^2 \end{cases}$$

We see immediately that the solution has the form

$$f = A(\tau) + B(\tau)\sigma + C(\tau)\sigma^2$$

Inserting this expression into the PDE gives three ordinary differential equations, for which the solutions are straightforward. This gives us an analytical expression for g and therefore also for $\hat{p}_m(k)$.

The solution method succeeded because the integral $\int_0^T \sigma dZ_t$ could be converted to powers of σ_T. This, in turn, was possible because of the simple form of the volatility SDE. If, for instance, the SDE instead had been a CEV process with an arbitrary skew parameter β, the resulting PDE would have been too complex to solve with a simple ansatz. Even for basic models as the lognormal process, the resulting PDE becomes rather complex. Despite this, there are a couple of interesting SDEs for which the method works. For example, adding a mean-reverting drift to the above volatility SDE does not make the solution procedure much more complicated. Furthermore, the method works fine on well-known models such as the *Heston model* where $\eta = \sqrt{\sigma}$ follows a mean-reverting square root process

$$d\eta_t = \lambda(\bar{\eta} - \eta_t)dt + \epsilon\sqrt{\eta_t}dZ_t$$

The method of solution is then almost identical to the above.

The method we have reviewed is only one of several types of Fourier transform techniques that can be used for option pricing. A more direct approach is to first make a substitution of variables in the stochastic volatility SDE so that the coefficients are independent of the underlying F. Once this is done, a Fourier transform gives a PDE of one dimension lower. The resulting volatility PDE for call option pricing has an initial condition that is independent of the volatility. The PDE is therefore fundamental because it can be used for any type of payoff, as long as the payoff only depends on the underlying F and not on the volatility σ. If an analytical solution can be found to the volatility PDE, all that remains is to perform the inverse Fourier transform. The reason for not devoting more time to this technique is that it is similar to the above approach. Indeed, both techniques only work for special forms of the volatility SDE and the remaining complex integral needs to be solved numerically. It should also be pointed out that there is a complication in the convergence of the Fourier transform of the payoff. This problem has been solved in Lewis (2000) by using a Fourier transform variable with an imaginary part and in Carr and Madan (1999) by using a damping function.

7.5 Comparison of Methods

In this section we compare the methods that have been discussed. But first we would like to point out that there are several other ways to solve stochastic volatility models apart from the ones mentioned in this chapter. We have chosen to focus on the techniques that we find to be the most theoretically appealing as well as useful from a practical perspective.

One of the methods that has been neglected in this chapter is the *SABR model*. It has a CEV SDE for the underlying and a correlated lognormal SDE for the volatility. The processes are therefore identical to the ones used in this chapter with the exception that SABR allows a CEV type of process in the underlying instead of limiting itself to a lognormal process. There are two components giving rise to a skew: the CEV parameter and the correlation between the Brownian drivers. The traditional wisdom is that an appropriate combination of these parameters can match the skew as well as the dynamics. The reason is that the CEV parameter implies sticky-strike dynamics while the correlation gives sticky-delta dynamics. However, as we argue in Sect. 7.7, this statement is not completely sound. Observe that this does not invalidate the results of the original paper by Hagan et al. (2002) as the SABR model was intended to be used for single-maturity derivatives and in Sect. 4.3 we showed that it is then possible to impose arbitrary dynamics, including the choice made by the inventors of the model.

The SABR model was originally solved using perturbation around the ATM point. Somewhat confusingly, when referring to the SABR model it is often not the processes that are meant but rather the processes together with the solution technique. The perturbative method used in the original paper contains rather complicated computations and it is hard to obtain anything but the lowest-order contribution. Even the lowest order term is complicated and we have therefore chosen to omit a detailed discussion of the SABR model. Furthermore, the SABR model is known to produce implied volatilities in disagreement with market data when far away from ATM and the model assumes both a small volatility and a small volatility of the volatility. Despite the shortcomings, the SABR model is popular among practitioners and has become the industry standard for pricing many types of products that have a single payment date (it works less well for multiple payment dates because of the lack of mean-reversion in the volatility process). In fact, it has become so popular that the quotes in many markets, e.g. caps and swaptions, follow the SABR model almost religiously.

Table 7.1 Stochastic volatility models comparison.

Method	General SDEs	Exact	High performance	$\rho \neq 0$
Vol of vol perturbation	Yes	No	Yes	Yes
Conditional expectation	Yes	No	Yes	No
Fourier transforms	No	Yes	No	Yes
SABR	No	No	Yes	Yes

The choice of method for solving stochastic volatility models depends on the purpose. The perturbative expansion technique in Sect. 7.2 is often to prefer because of the resulting simple expressions which allow for extensions, e.g. to include higher order terms, to do the perturbation around a different base function or to use a mean-reverting volatility process. The same comment applies to the conditional expectation approach, but it requires a zero correlation. The advantage of this approach is that the computations are even simpler. The Fourier transform approach is usually applied when the user is not satisfied with perturbative expansions but prefers analytical solutions. The drawback is that an integral has to be computed numerically, which can be time consuming. Furthermore, the Fourier transform technique is only efficient for special types of SDEs.

We compare the techniques of this chapter in Table 7.1. The comparison is made based on the following criteria: if the techniques can solve general forms of SDEs, if the solution is exact or perturbative, if an implementation is of high performance (i.e. if only a few algebraic operations are needed) and whether it is possible to apply the technique to models with non-zero correlation.

7.6 Relations to Implied and Local Volatility

In Sect. 6.2 we used the forward Kolmogorov equation to find a relation between the local volatility and the implied volatility. We now use the same technique to find a corresponding relation for a stochastic volatility model:

$$
\begin{aligned}
\frac{d}{dT}U &= \int (F_T - K)_+ \frac{d}{dT} p(T, F_T, \sigma_T; t, F, \sigma) dF_T d\sigma_T \\
&= \int (F_T - K)_+ \left(\frac{1}{2} \frac{d^2}{dF_T^2} \sigma_T^2 F_T^2 + \frac{d^2}{dF_T d\sigma_T} \rho\epsilon\sigma_T^2 F_T + \frac{1}{2} \frac{d^2}{d\sigma_T^2} \epsilon^2 \sigma_T^2 \right) \\
&\quad p(T, F_T, \sigma_T; t, F, \sigma) dF_T d\sigma_T \\
&= \int \frac{1}{2} \sigma_T^2 F_T^2 p(T, F_T, \sigma_T; t, F, \sigma) \frac{d^2}{dF_T^2} (F_T - K)_+ dF_T d\sigma_T \\
&= \frac{1}{2} K^2 \int \sigma_T^2 p(T, K, \sigma_T; t, F, \sigma) d\sigma_T
\end{aligned}
$$

$$= \frac{1}{2}K^2 \int \sigma_T^2 p(T, \sigma_T; t, \sigma | F_T = K) d\sigma_T p(T, K; t, F)$$

$$= \frac{1}{2}K^2 E[\sigma_T^2 | F_T = K] p(T, K; t, F)$$

$$= \frac{1}{2}K^2 E[\sigma_T^2 | F_T = K] \int \delta(F_T - K) p(T, F_T, \sigma_T; t, F, \sigma) dF_T d\sigma_T$$

$$\Leftrightarrow E[\sigma_T^2 | F_T = K] = \frac{\frac{d}{dT}U}{\frac{1}{2}K^2 \frac{d^2}{dK^2}U}$$

As expected, the formula gets reduced to Dupire's formula for non-stochastic volatility. Unlike the situation for non-stochastic volatility, this formula is not suitable for calibration. However, by recognizing that the right-hand side is the local volatility, we obtain a relation (Derman and Kani (1998)) between a stochastic volatility model and the local volatility model that gives the same option prices

$$\sigma_{\text{loc}}^2(T, K) = E[\sigma_T^2 | F_T = K]$$

7.7 Dynamics

The dynamics of stochastic volatility models can be understood from the following dimensionless formulation of the benchmark model that has been used in this chapter:

$$\begin{cases} \mathcal{D}\Phi = \epsilon\rho\sigma^2 \partial_{x\sigma}\Phi + \frac{1}{2}\epsilon^2\sigma^2\partial_\sigma^2\Phi, \quad \mathcal{D} = \partial_\tau - \frac{1}{2}\sigma^2(\partial_x^2 - \partial_x) \\ \Phi(\tau = 0, x) = (e^x - 1)_+ \end{cases}$$

The implied volatility can be computed by comparing with the corresponding non-stochastic volatility model, obtained by setting $\epsilon = 0$. We then see that F and K enter both expressions only through the combination $x = \ln(F/K)$. It means that the implied volatility must be of the form $\sigma_{\text{imp}}(K, F) = \sigma_{\text{imp}}(K/F)$, i.e. the model is of sticky-delta type. However, as pointed out in Mercurio and Morini (2008), this argument is not completely sound. The reason is that a change in F implies a change in the volatility σ as they are correlated. Indeed, assume that F changes by dF. It means that the driver of the forward changes by $dW = dF/\sigma F$ which implies a change

$$d\sigma = \epsilon\sigma(\rho dW + \sqrt{1-\rho^2}dW^\perp) = \frac{\epsilon\rho}{F}dF + \epsilon\sigma\sqrt{1-\rho^2}dW^\perp$$

As $E[dW^\perp] = 0$, the average change in the volatility is given by $\frac{\epsilon\rho}{F}dF$.

To understand the effect of the volatility change on the dynamics, we use the lowest-order perturbative expression

$$\sigma_{\mathrm{imp}}(K) \approx \sigma - \frac{1}{2}\epsilon\rho \left(\ln(F/K) - \sigma^2\tau/2\right)$$

derived earlier in this chapter. When F changes with dF and σ with $\frac{\epsilon\rho}{F}dF$, we obtain the following new expression for the implied volatility to the lowest order:

$$\begin{aligned}\sigma_{\mathrm{imp}}^{\mathrm{new}}(K) &\approx \sigma + \frac{\epsilon\rho}{F}dF - \frac{1}{2}\epsilon\rho \left(\ln((F+dF)/K) - \sigma^2\tau/2\right) \\ &= \frac{\epsilon\rho}{F}dF - \frac{1}{2}\epsilon\rho \ln\left(1 + \frac{dF}{F}\right) + \sigma_{\mathrm{imp}}(K) \approx \frac{1}{2}\frac{\epsilon\rho}{F}dF + \sigma_{\mathrm{imp}}(K)\end{aligned}$$

Consider the instance of a negative correlation which means that the implied volatility decreases with the strike. As $\rho < 0$, the implied volatility shifts downward when the underlying increases, i.e. when $dF > 0$, according to the above formula. It means that just as for local volatility models, the implied volatility surface moves in the wrong direction when the underlying is changing. In fact, we see from the above computation that the change in F gives a sticky-delta behavior while the change in the volatility, coming from the correlation with the underlying, gives a larger and opposite change. It is straightforward to verify that the same effect occurs for positive correlation. Furthermore, it can be shown that the implied volatility moves in the wrong direction not only for our benchmark model but for any stochastic volatility model, as long as the volatility of volatility is small. It seems that this behavior also occurs for finite values of the volatility of volatility, see Mercurio and Morini (2008) for a specific example.

Observe that the above dynamics effect only occurs for non-zero correlation. Indeed, when the underlying and the volatility are uncorrelated we clearly obtain a model with sticky-delta dynamics. We therefore conclude that stochastic volatility models have one type of dynamics for the skew part and another for the smile part.

When pricing vanillas, it is possible to impose dynamics different from the inherent behavior of the model, for example, a sticky-delta behavior. Unfortunately, as we discussed in Sect. 4.3, this does not help when it comes to the pricing of path-dependent derivatives.

Let us move on to the hedging in a stochastic volatility model. Because of bid-offer spreads, the vega hedge is more expensive than the delta hedge. Vega hedging is for this reason sometimes avoided. It is then possible to pick up a part of the volatility risk through the delta hedge. Indeed, using

$$\frac{dU}{dF} = \frac{\partial U}{\partial F} + \frac{\partial U}{\partial \sigma}\frac{d\sigma}{dF}$$

and the above expression for $d\sigma$, we see that the best estimate for the delta is given by

$$\frac{dU}{dF} = \frac{\partial U}{\partial F} + \frac{\epsilon\rho}{F}\frac{\partial U}{\partial \sigma}$$

We then pick up the change in volatility that is parallel to the driver of the underlying while the orthogonal part is left unhedged. This technique works best for a high absolute value of the correlation between the underlying and the volatility.

The danger of using this method is that the correlation is often calibrated to the implied volatility skew. It can then have a completely different value from the observed correlation between the underlying and the volatility. Such a mismatch can lead to a stochastic volatility model with bad dynamics and the correction term $\frac{\epsilon\rho}{F}\frac{\partial U}{\partial \sigma}$ might do more bad than good.

7.8 Local Stochastic Volatility

It is popular to combine stochastic volatility models with local volatility models. One example of how this can be done is by letting the forward follow

$$dF_t = \sigma_t A(t, F_t) F_t dW_t$$

and allowing σ_t to be stochastic. A popular approach for calibrating such models is via fixed point iterators, see Ren et al. (2007).

Bibliography

Carr P, Madan D (1999) Option valuation using the fast fourier transform. J Comput Finance 3:463–520

Derman E, Kani I (1998) Stochastic Implied Trees: Arbitrage pricing with stochastic volatility term and strike structure of volatility. Int J Theoretical App Finance 1:61–110

Hafner R (2004) Stochastic implied volatility: a factor-based model. Springer, Berlin

Hagan P, Kumar D, Lesniewski A, Woodward D (2002) Managing smile risk. WILMOTT Magazine, September:84–108

Lewis AL (2000) Option Valuation under stochastic volatility: with matematica code. Finance Press, Newport Beach, California

Mercurio F, Morini M (2008) A note on hedging with local and stochastic volatility models. Social Science Research Network. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1294284. Accessed 16 May 2011

Ren Y, Madan D, Qian M (2007) Calibrating and pricing with embedded local volatility models. Risk September:138–143

Chapter 8
Lévy Models

We have seen that modeling the logarithmic returns of the underlying as a Brownian motion does not capture the rich structure observed in the option markets. It can explain neither the skew or smile nor the dynamics of the implied volatility surface. As an attempt at improvement, we extended the Black–Scholes model in the previous two chapters to local volatility models and stochastic volatility models. Despite the success and frequent use of these model types around financial institutions, we have made it clear that they also suffer from the fact that the dynamics disagree with market behavior. For this reason we now discuss yet another model class generalizing the Black–Scholes framework.

Instead of using the Brownian motion as the fundamental process in the modeling, Lévy models use a more general class of processes that include jumps. The class of Lévy processes is large and the modeling can therefore contain several free parameters giving a flexibility in the model and the possibility to match skew and smile accurately. Furthermore, the occurrence of jumps in Lévy processes makes these models more suitable than stochastic volatility models for describing skew and smile for short maturities. Unfortunately, there are in general no closed-form solutions for option prices in Lévy models. Another disadvantage is that these models are complicated to implement efficiently on a tree structure.

Just as local volatility models and stochastic volatility models can be used together, it is also possible to combine these models with Lévy models to obtain the best of the individual models. For example, by replacing the Brownian driver with a Lévy driver in these model types, we obtain local Lévy models (Carr et al. (2004)) and stochastic volatility models with jumps (Bates (1996)). One reason for combining Lévy processes with stochastic volatility models is of practical character: both models are relatively simple to solve in Fourier space. Therefore, if Fourier transforming in order to solve a Lévy process, it is possible to add stochastic volatility without too much extra work.

Because we expect the reader to be familiar with the mathematics of Brownian motions, we chose not too include this theory in the previous chapters, but instead summarized it in the Appendix. The mathematics behind Lévy processes is, on the other hand, not as well known and we have therefore chosen to include it in the main

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_8,

text. As a result, this chapter has a more mathematical character. We spend most of the time on defining Lévy processes and on developing the corresponding stochastic calculus. We also discuss important special cases and indicate how they can be used for pricing.

8.1 Lévy Processes

To obtain a richer set of processes, we relax the defining conditions for a Brownian motion that were given in the Appendix. To understand which of the conditions that should be kept and which that can be omitted, we consider the example of a process that describes the returns of a financial product such as an equity stock. For arbitrage to be absent, the returns over two non-overlapping time periods must be independent. Furthermore, it is a plausible assumption that the returns over time periods of equal length should be identically distributed. It therefore makes sense to keep the conditions that the increments should be independent and stationary, the latter meaning that the distribution of $X_{t+\Delta t} - X_t$ is independent of t.

As usual, there are violations to every assumption underlying a model. For instance, financial returns have a tendency to be auto-correlated when viewed at time scales that are so small that they are comparable with the time for information to flow through the market. The distribution of the returns of a financial underlying can also be fundamentally different on certain days, violating the stationarity assumption. An example is given by days when close-call political elections are held or when statistical data concerning the financial market is presented. This type of external information flow into the financial market can also result in jumps of financial assets.

To include jumps in the underlying process, thereby obtaining a more realistic description of financial instruments, it is necessary to relax the condition of continuous paths that was used for Brownian motions. When relaxing this condition, it is no longer possible to require the increments to be normally distributed. We therefore see that from a financial context, it makes sense to extend the Brownian motion to *Lévy processes* X defined by:

- The increments are stationary and independent
- $X(0) = 0$
- For every $\epsilon > 0$ and $t \geq 0$, $\lim_{\Delta t \to 0} P\left(|X_{t+\Delta t} - X_t| > \epsilon\right) = 0$

The second condition has been added for normalization while the third condition turns out to be useful in the analysis of processes with independent and stationary increments.

It can be shown that for every Lévy process, there is a modification Y that is càdlàg. A function f is said to be *càdlàg* if the limits $f(t_-) = \lim_{\Delta t \to 0} f(t - \Delta t)$ and $f(t_+) = \lim_{\Delta t \to 0} f(t + \Delta t)$ exists and $f(t) = f(t_+)$. A process Y is said to be càdlàg if its paths have this property and it is said to be a modification (or version) of X if $Y_t = X_t$ a.s. for all t. Motivated by the fact that all Lévy processes have a càdlàg modification, we restrict ourselves to considering càdlàg Lévy processes.

It is easy to understand that the main defining condition of Lévy processes, namely that of independent and stationary increments, is fundamental and not specific to finance. For this reason, Lévy processes have been used in several of the mayor quantitative disciplines. Thus, there is a long history of work done on Lévy processes and many powerful theorems exist. This makes them an ideal tool for financial modeling.

8.2 Lévy-Ito Decomposition

The most well-known Lévy process is the Brownian motion. In fact, the defining conditions for Lévy processes are satisfied even after including time-independent drift and volatility in the Brownian motion. However, the defining conditions also allow for processes with discontinuous paths. To study the fundamental properties of such processes, we initially focus on pure jump processes, i.e. Lévy processes that are constant until a jump occurs.

For an arbitrary t and positive integer M, divide the time-interval $[0, t)$ into M subintervals of equal length: $[0, t) = [0, t/M) \cup [t/M, 2t/M) \cup \ldots \cup [(\text{M-1})t/M, t)$. A *Poisson process* is a pure jump process such that the probability of more than one jump occurring in any of the subinterval tends quickly to zero when $M \to \infty$. It follows from the independence and stationarity that for large M and small Δt, the probability of a jump occurring in $[0, \Delta t)$ is equal to M times the probability of a jump occurring in $[0, \Delta t/M)$. This implies that the probability of a jump in $[0, \Delta t)$ equals $\lambda \Delta t$ for Δt small. From the stationarity of increments, we conclude that the probability of a jump in $[t, t + \Delta t)$ is equal to $\lambda \Delta t$ for small Δt, where λ is independent of t.

For M large, the probability of k jumps occurring in the interval $[0, t)$ equals the probability of k of the subintervals to have one jump each and for $M - k$ subintervals to have no jumps. This gives us

$$
\begin{aligned}
P(N_t = k) &= \frac{M!}{k!(M-k)!} \left(\frac{\lambda t}{M}\right)^k \left(1 - \frac{\lambda t}{M}\right)^{M-k} \\
&\to \frac{M^k}{k!} \left(\frac{\lambda t}{M}\right)^k \left(1 - \frac{\lambda t}{M}\right)^M \to e^{-\lambda t} \frac{(\lambda t)^k}{k!}
\end{aligned}
$$

when $M \to \infty$. Thus, N_t is *Poisson distributed* with parameter λt. In particular, it follows that the expected number of jumps before t equals λt.

We observe that the defining conditions for Lévy processes allow for arbitrary jump sizes, as long as the jumps are independent. Using the relation $N_t = \sum_{i=1}^{N_t} 1$, the Poisson process can be generalized to

$$X_t = \sum_{i=1}^{N_t} Y_i$$

where $\{Y_i\}$ are independent and identically distributed (*i.i.d.*) variables that are independent of N_t. This process is called a *compound Poisson process* and is quite general as we have the freedom in the choice of the jump intensity λ as well as the jump distribution. In fact, it is intuitively clear that the only Lévy processes with piece-wise constant paths are the compound Poisson processes.

To express the compound Poisson process in a form suitable for generalization to arbitrary Lévy processes, we introduce the measure μ_X such that $\mu_X(\omega; [0,t], C)$ is the number of jumps with size in C that occurs before t for the path $\omega \in \Omega$, where $C \subset \mathbb{R}$ is a Borel set. The compound Poisson process can then be written as

$$X_t = \int_0^t \int_{\mathbb{R}} x\mu_X(ds, dx)$$

For a space E equipped with a σ-algebra $\mathcal{B}$, a function $\mu : \Omega \times \mathcal{B} \to \mathbb{R}_+$ is said to be a *random measure* if (1) for each fixed ω, $\mu(\omega, \cdot)$ is a measure on $(E, \mathcal{B})$ and (2) for each fixed $B \in \mathcal{B}$, $\mu(\cdot, B)$ is a random variable on Ω. Let $\bar{\mu}$ be a σ-finite measure on $(E, \mathcal{B})$, i.e. a measure such that E can be written as the countable union of sets with finite measure. The measure μ is then said to be a *Poisson random measure* with respect to $\bar{\mu}$ if its range is the positive integers and it satisfies

- $\mu(\cdot, B_1), \mu(\cdot, B_2), \ldots, \mu(\cdot, B_n)$ are independent for disjoint $B_1, B_2, \ldots, B_n$.
- If $\bar{\mu}(B) < \infty$, then $\mu(\cdot, B)$ is Poisson distributed with density $\bar{\mu}(B)$.

$$P\left(\mu(\cdot, B) = k\right) = \frac{\bar{\mu}(B)^k}{k!} e^{-\bar{\mu}(B)}$$

We are interested in the situation when $E = \mathbb{R}_+ \times \mathbb{R}$ contains the time $t \in \mathbb{R}_+$ and the size $x \in \mathbb{R}$ of the jump, and $\mathcal{B}$ is the Borel σ-algebra on E. We take the measure $\bar{\mu}$ on E to be the product of the Lebesgue measure on $\mathbb{R}_+$ and the measure on $\mathbb{R}$ that describes the jump-size distribution. The measure μ_X for a compounded Poisson distribution is then a Poisson random measure. The *Lévy measure* ν_X counts the expected number of jumps with a certain jump size:

$$\nu_X(C) = E\left[\mu_X(\cdot; [0,t], C)\right]/t$$

From the independence and stationarity of the increments, it follows that ν_X does not depend on t.

Consider the martingale $\tilde{N}_t = N_t - \lambda t$ obtained by subtracting the intensity from a Poisson process. This is called a *compensated Poisson process*. In the same manner, the compensated compound Poisson process can be defined by

$$\tilde{X}_t = \int_0^t \int_{\mathbb{R}} x\tilde{\mu}_X(ds, dx)$$

where $\tilde{\mu}_X$ is the compensated Poisson random measure defined by

$$\tilde{\mu}_X(\omega; [0,t], C) = \mu_X(\omega; [0,t], C) - t\nu_X(C)$$

After this reformulation of (compensated) compound Poisson processes in terms of (compensated) Poisson random measures, we are prepared for a generalization to arbitrary Lévy processes. It is well known that for every Lévy process, the measure μ_X is a Poisson random measure with Lévy measure ν_X satisfying $\int_{\mathbb{R}} \min\left(1, |x|^2\right) \nu_X(dx) < \infty$. Motivated by our discussion of compound Poisson processes, it is tempting to express the pure jump part of an arbitrary Lévy processes as $\int_0^t \int_{\mathbb{R}} x\mu_X(ds, dx)$. Unfortunately, there is a problem with convergence as a càdlàg function can have an infinite, but countable, number of jumps in any time interval. The rescue lies in the fact that for any $\epsilon > 0$ and $t > 0$, there is only a finite number of jumps in $[0, t)$ with size larger than ϵ. It means that the integral $\int_0^t \int_{|x|>\epsilon} x\mu_X(ds, dx)$ contains only a finite number of jumps and is therefore well defined. Because of the possibility of an infinite number of small jumps, however, the limit when $\epsilon \to 0$ might not exist. One approach to reach convergence is to replace μ_X with the compensated measure $\tilde{\mu}_X$ and apply convergence results for martingales. It can be shown that the integral then converges in the limit of small ϵ, but there are instead problems with convergence for large values of $|x|$. The standard way to solve this dilemma is to separate the integral into two pieces where μ_X is used for $|x| > 1$ and $\tilde{\mu}_X$ for $|x| \leq 1$. We then obtain the convergent result

$$\int_0^t \int_{|x|>1} x\mu_X(ds, dx) + \int_0^t \int_{|x|<1} x\tilde{\mu}_X(ds, dx)$$

The above expression describes the jump part of the Lévy process. For an expression for the most general Lévy process, it turns out to be sufficient to add an independent Brownian motion with drift. This is the idea of the *Lévy-Ito decomposition* that states that for an arbitrary Lévy process there exist constants γ and σ such that X_t is a.s. equal to

$$\gamma t + \sigma W_t + \int_0^t \int_{|x|>1} x\mu(ds, dx) + \int_0^t \int_{|x|<1} x\tilde{\mu}(ds, dx)$$

where μ is a Poisson random measure, $\tilde{\mu} = \mu - \nu$ is the corresponding compensated measure and the Lévy measure satisfies $\int_{\mathbb{R}} \min\left(1, |x|^2\right) \nu(dx) < \infty$. W_t is a standard Brownian motion that is independent of μ.

It can be shown that the jump part of a Lévy process is of finite variation if and only if $\int_{\mathbb{R}} \min(1, |x|) \nu(dx) < \infty$. The truncation of small jumps is therefore not needed and the Lévy-Ito decomposition can be written as

$$\gamma t + \sigma W_t + \int_0^t \int_{\mathbb{R}} x\mu(ds, dx)$$

As we see later, the distribution of a Lévy process is in general easier to analyze in Fourier space. For this purpose, we use Lévy-Ito decomposition to derive the characteristic function for the most general Lévy process. As usual, we start by considering a compound Poisson process $X_t = \sum_{i=1}^{N_t} Y_i$, which has the characteristic function

$$\begin{aligned}\Phi_{X_t}(k) &= E\left[e^{ik\sum_{i=1}^{N_t} Y_i}\right] = \sum_{j=0}^{\infty} P(N_t = j) E\left[e^{ik\sum_{i=1}^{j} Y_i}\right] \\ &= \sum_{j=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} E\left[e^{ikY}\right]^j = \exp\left(\lambda t \left(E\left[e^{ikY}\right] - 1\right)\right) \\ &= \exp\left(t \int_{\mathbb{R}} \left(e^{ikx} - 1\right) \lambda p_Y(dx)\right) = \exp\left(t \int_{\mathbb{R}} \left(e^{ikx} - 1\right) \nu_X(dx)\right)\end{aligned}$$

As the terms in the Lévy-Ito decomposition are independent, it can be shown that for a general Lévy process we have

$$\Phi_{X_t}(k) = e^{t\Psi(k)}$$

where the characteristic exponent is given by

$$\Psi(k) = i\gamma k - \frac{1}{2}\sigma^2 k^2 + \int_{\mathbb{R}} \left(e^{ikx} - 1 - ikx\mathbb{1}_{|x|\le 1}\right) \nu_X(dx)$$

This expression for the characteristic function is called the *Lévy-Khinchin representation*. It is defined by the so-called *characteristic triplet* (σ^2, ν, γ).

8.3 Stochastic Calculus

Before discussing stochastic calculus for Lévy processes, we recall that the corresponding calculus for Brownian motion was in fact developed in the Appendix for the more general class of continuous semimartingales. The reason was that, as opposed to Brownian motions and Lévy processes, this class is closed under the operations of interest to us, for instance, the application of a second-order differentiable function. It would therefore make sense to develop stochastic calculus for non-continuous semimartigales, which is the corresponding generalization of Lévy processes. Indeed, the stochastic calculus for this class is similar to that for continuous semimartingales. However, the class of non-continuous semimartingales is a bit too general for our purposes (the same can also be said for the continuous semimartingale generalization of Brownian motions but we have anyway chosen to include it because we assume that most readers are somewhat familiar with this topic) and we therefore only develop stochastic calculus for processes of the form

$$X_t = \int_0^t \gamma_s ds + \int_0^t \sigma_s dW_s + \int_0^t \int_{|x|>1} \delta_{sx}\mu(ds,dx) + \int_0^t \int_{|x|\le 1} \delta_{sx}\tilde{\mu}(ds,dx)$$

which clearly generalizes Lévy processes.

For X_t of the above form and f a C^2 function, Ito's lemma reads:

$$\begin{aligned} f(X_t) = f(X_0) &+ \int_0^t f'(X_{s-})dX_s + \frac{1}{2}\int_0^t \sigma_s^2 f''(X_s)ds \\ &+ \int_0^t \int_{\mathbb{R}} \big(f(X_{s-}+\delta_{sx}) - f(X_s) - \delta_{sx} f'(X_{s-})\big)\,\mu(ds,dx) \end{aligned}$$

This formula can be understood by first considering the situation when X does not have any jumps. The last term then vanishes and the first three terms are exactly what we had expected from the continuous version of Ito's lemma. On the other hand, if X is a pure jump process, i.e. $X_t = \int_0^t \int_{\mathbb{R}} \delta_{sx}\mu(ds,dx)$, then clearly

$$f(X_t) = f(X_0) + \int_0^t \int_{\mathbb{R}} \left(f(X_{s-}+\delta_{sx}) - f(X_s)\right)\mu(ds,dx)$$

This also agrees with the above version of Ito's lemma as the second term cancels the last term in the double integral. The general formula is obtained by combining the special cases when X has no jumps and when it is a pure jump process. Inserting the expression for dX_t gives

$$\begin{aligned} f(X_t) = f(X_0) &+ \int_0^t \sigma_s f'(X_{s-})dW_s + \int_0^t \int_{\mathbb{R}} \left(f(X_{s-}+\delta_{sx}) - f(X_s)\right)\tilde{\mu}(ds,dx) \\ &+ \int_0^t \gamma_s f'(X_{s-})ds + \frac{1}{2}\int_0^t \sigma_s^2 f''(X_s)ds \\ &+ \int_0^t \int_{\mathbb{R}} \big(f(X_{s-}+\delta_{sx}) - f(X_s) - \delta_{sx} f'(X_{s-})\mathbb{1}_{|x|\le 1}\big)\,\nu(ds,dx) \end{aligned}$$

where the first line is the martingale part of $f(X_t)$ and the second and third line are the bounded-variation part.

We now consider two useful applications of Ito's lemma. The first regards the instance when $Y_t = f(X_t) = e^{X_t}$, which gives

$$\begin{aligned} \frac{dY_t}{Y_{t-}} &= \sigma_t dW_t + \int_{\mathbb{R}} \left(e^{\delta_{tx}} - 1\right)\tilde{\mu}(dt,dx) + \gamma_t dt + \frac{1}{2}\sigma_t^2 dt \\ &\quad + \int_{\mathbb{R}} \left(e^{\delta_{tx}} - 1 - \delta_{tx}\mathbb{1}_{|x|\le 1}\right)\nu(dt,dx) \end{aligned}$$

As an example of Ito's lemma in higher dimensions, we consider the product XY where X is as above and Y is given by a corresponding expression:

$$Y_t = \int_0^t \gamma_s' ds + \int_0^t \sigma_s' dW_s + \int_0^t \int_{|x|>1} \delta_{sx}' \mu(ds, dx) + \int_0^t \int_{|x|\le 1} \delta_{sx}' \tilde{\mu}(ds, dx)$$

In the same way as in the 1-dimensional case, we obtain

$$\begin{aligned} X_t Y_t &= X_0 Y_0 + \int_0^t X_{s-} dY_s + \int_0^t Y_{s-} dX_s + \int_0^t \sigma_s \sigma_s' ds \\ &\quad + \int_0^t \int_{\mathbb{R}} \big((X_{s-} + \delta_{sx})(Y_{s-} + \delta_{sx}') - X_{s-} Y_{s-} \\ &\quad - \delta_{sx}' X_{s-} - \delta_{sx} Y_{s-}\big)\, \mu(ds, dx) \\ &= X_0 Y_0 + \int_0^t X_{s-} dY_s + \int_0^t Y_{s-} dX_s + \int_0^t \sigma_s \sigma_s' ds + \int_0^t \int_{\mathbb{R}} \delta_{sx} \delta_{sx}' \mu(ds, dx) \end{aligned}$$

For a useful application of the product formula, let

$$X_t = \int_0^t \int_{\mathbb{R}} \delta_{tx} \left(\mu(dt, dx) - y(t, x)\nu(dt, dx)\right)$$

for a positive function y and assume that M_t satisfies

$$\frac{dM_t}{M_{t-}} = \psi_t \sigma_t dW_t + \int_{\mathbb{R}} \left(y(t, x) - 1\right) \tilde{\mu}(dt, dx)$$

It follows that

$$\begin{aligned} X_t M_t &= X_0 M_0 + \int_0^t X_{s-} dM_s + \int_0^t M_{s-} \int_{\mathbb{R}} \delta_{sx} \left(\mu(ds, dx) - y(s, x)\nu(ds, dx)\right) \\ &\quad + \int_0^t M_{s-} \int_{\mathbb{R}} \delta_{sx}(y(s, x) - 1)\mu(ds, dx) \\ &= X_0 M_0 + \int_0^t X_{s-} dM_s + \int_0^t M_{s-} \int_{\mathbb{R}} \delta_{sx} y(s, x) \tilde{\mu}(ds, dx) \end{aligned}$$

which means that $X_t M_t$ is a martingale. Therefore, if M describes a measure change $M_t = \frac{dQ_t}{dP_t}$, then X_t is a Q-martingale. It implies that

$$\nu^Q(dt, dx) = y(t, x)\nu^P(dt, dx)$$

Consider the relation

$$
\begin{aligned}
X_t &= \int_0^t \gamma_s ds + \int_0^t \sigma_s dW_s + \int_0^t \int_{|x|>1} \delta_{sx}\mu(ds,dx) + \int_0^t \int_{|x|\le 1} \delta_{sx}\tilde{\mu}(ds,dx) \\
&= \int_0^t \gamma_s ds + \int_0^t \sigma_s d\left(W_s - \int_0^s \psi_u\sigma_u du\right) + \int_0^t \sigma_s^2\psi_s ds + \int_0^t \int_{|x|>1} \delta_{sx}\mu(ds,dx) \\
&\quad + \int_0^t \int_{|x|\le 1} \delta_{sx}\left(\mu(ds,dx) - y(s,x)\nu(ds,dx)\right) \\
&\quad + \int_0^t \int_{|x|\le 1} \delta_{sx}\left(y(s,x) - 1\right)\nu(ds,dx)
\end{aligned}
$$

As $W_t - \int_0^t \psi_u\sigma_u du$ is a Q-Brownian motion, X_t has the Q-triple

$$
\begin{cases}
\gamma_t^Q &= \gamma_t + \sigma_t^2\psi_t + \int_{|x|\le 1} \delta_{tx}(y(t,x)-1)\nu(dt,dx)/dt \\
\sigma_t^Q &= \sigma_t \\
\nu^Q(dt,dx) &= y(t,x)\nu(dt,dx)
\end{cases}
$$

8.4 Examples of Lévy Processes

In Sect. 8.2 we gave some important examples of Lévy processes in terms of Brownian motions with drift and compound Poisson processes. In this section we consider some other examples of Lévy processes that are useful in financial modeling.

A random variable that for every positive integer M can be written as a sum of M i.i.d. variables is said to be *infinitely divisible*. This property is satisfied for a Lévy process X at arbitrary times t since

$$
X_t = \sum_{i=1}^{M} \left(X_{it/M} - X_{(i-1)t/M}\right)
$$

Furthermore, it can be shown that for every infinitely divisible distribution and $t > 0$, there exists a Lévy process X such that X_t has this distribution. The existence of a Lévy-Khinchin representation for infinitely divisible distributions is therefore a consequence of the results in Sect. 8.2.

In a decomposition $X = \sum_{i=1}^M X^{(i)}$ of an infinitely divisible distribution, the distribution of the i.i.d. components $X^{(i)}$ can be fundamentally different from that of X. In certain quantitative branches, there is no reference time scale which means that it should not be possible to determine the value of the time t from the distribution of X_t. This motivates us to consider distributions with the property that for any M, $\sum_{i=1}^M X^{(i)}$ has the same distribution as $bX^{(0)} + c$ if $\{X^{(i)}\}_{i\ge 1}$ are i.i.d. variables with the same distribution as $X^{(0)}$. Thus, summing up i.i.d. distributions leads to nothing but a scaled and translated version of the distribution. We conclude that the characteristic function of $X = X^{(0)}$ must be of the form $\Phi_X(k)^M = \Phi_X(kb)e^{ikc}$, where b and c depend on M. This property can be generalized to a continuous-time process by replacing the integer M with an arbitrary positive number a, giving

$$\Phi_X(k)^a = \Phi_X(kb)e^{ikc}$$

Distributions with this property are said to be *stable*. It can be shown that b must be of the form $a^{1/\alpha}$ for $\alpha \in (0, 2]$.

For $\alpha = 2$ it can be proven that $\nu = 0$ in the Lévy-Khinchin representation, which means that the distribution is given by a normal random variable. For $\alpha < 2$, it follows that $\sigma = 0$ in the characteristic triple and

$$\nu(dx) = \left(\frac{c_1}{x^{\alpha+1}} \mathbb{1}_{x>0} + \frac{c_2}{|x|^{\alpha+1}} \mathbb{1}_{x<0} \right) dx$$

for some positive constants c_1 and c_2, i.e. the tail decay is that of a *Pareto distribution*. The characteristic function of a stable distribution can be written as

$$\Phi_X(k) = e^{i\gamma k + \sigma^\alpha (ik\omega - |k|^\alpha)}, \quad \omega = \begin{cases} -\beta \dfrac{2}{\pi} \ln |k|, & \alpha = 1 \\ \beta |k|^{\alpha-1} \tan \dfrac{\pi\alpha}{2}, & \alpha \neq 1 \end{cases}$$

for $\sigma \geq 0$, $\beta \in [-1, 1]$ and $\gamma \in \mathbb{R}$. It can be shown that only for $1 < \alpha \leq 2$ it holds that $E[|X|] < \infty$ and only for $\alpha = 2$ we have $E[|X|^2] < \infty$. Closed-form expressions for the PDFs of the stable distributions are known in three cases:

$$\begin{cases} \textit{Gaussian distribution}(\alpha = 2) & : \dfrac{1}{2\sigma\sqrt{\pi}} e^{-(x-\gamma)^2/4\sigma^2} \\ \textit{Cachy distribution } (\alpha = 1, \beta = 0) & : \dfrac{\sigma}{\pi\left((x-\gamma)^2 + \sigma^2\right)} \\ \textit{Lévy distribution } (\alpha = 1/2, \beta = 1) & : \left(\dfrac{\sigma}{2\pi}\right)^{\frac{1}{2}} \dfrac{1}{(x-\gamma)^{3/2}} e^{-\sigma/2(x-\gamma)} \mathbb{1}_{x>\gamma} \end{cases}$$

A Lévy process is said to be *stable* if X_{at} equals $a^{1/\alpha} X_t + ct$ in distribution. Stable Lévy processes are often referred to as *Lévy flights*. The distribution of X_t is then stable for any $t > 0$. Conversely, for every stable distribution and $t > 0$, there exists a Lévy flight X such that X_t has this distribution.

Recall that there exist fundamental time scales in finance, for example, given by the time it takes for information to flow through the market. It should therefore come as no surprise that the distributions of the returns from financial instruments have been found to depend on the observing time period. An example of this is that the tails of the distributions are typically much fatter for short time intervals. Therefore, despite their attractive theoretical properties and application in various scientific areas, Lévy flights are not ideal for mathematical finance. Despite this, they have anyway been used as modeling tools in certain branches of mathematical finance, e.g. risk measurement. Unfortunately, their applicability to derivatives pricing is limited because of divergence problems originating in the infinite second moment. We therefore spend the remaining part of this section on finding alternative Lévy processes useful for financial modeling.

A popular technique for constructing interesting processes in mathematical finance is to take a familiar process X_t and then change the flow of time from t to $\tau(t)$ for some function τ. We have given several examples of this technique, see, for example, Chaps. 5 and 13. We now generalize the method by allowing τ to be a stochastic process. In fact, we let τ be a Lévy process as it is then obviously true that $X_{\tau(t)}$ is a Lévy process if X_t is a Lévy process. To not deviate too much from standard Gaussian modeling, we restrict the analysis to X being a Brownian motion with drift: $\gamma t + \sigma W_t$.

As time is an increasing process, we must restrict ourselves to τ being a *subordinator*, i.e. a Lévy process that is increasing: $\tau(t_1) \leq \tau(t_2)$ a.s. when $t_1 \leq t_2$. Clearly, such a process cannot have a diffusion part, $\sigma = 0$, and must only allow positive jumps, $\nu\left((-\infty, 0)\right) = 0$. It can be proven that the positive jumps must be of finite variation, which implies that the characteristic exponent can be written as

$$\Psi(k) = i\gamma k + \int_0^\infty \left(e^{ikx} - 1\right)\nu(dx)$$

for $\gamma \geq 0$. The integral can be analytically continued to the positive part of the real axis, which gives a well-defined moment generating function

$$M_\tau(k) = E\left[e^{k\tau}\right] = \exp\left(\gamma k - i\int_0^\infty \left(e^{ikx} - 1\right)\nu(dx)\right)$$

Clearly, a Lévy process that is positive, $X_t > 0$ a.s. for any fixed t, is a subordinator. Therefore, one way to construct subordinators is to find positive and infinitely divisible distributions. For this reason, we consider the *inverse Gaussian distribution* X with the PDF

$$p(x) = \frac{\delta}{\sqrt{2\pi x^3}} e^{-(\delta - \gamma x)^2/2x}, \quad x > 0$$

The name can be a bit misleading as the distribution is not the inverse of the Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-(x-\mu t)^2/2\sigma^2 t}$$

in the usual sense. The explanation for the name is that the Gaussian distribution describes the location in space of a drifted Brownian motion at a fixed time while the inverse Gaussian distribution describes the location in time when a positive level in space is first hit. The above formula can be obtained from the Appendix after the transformation $\delta = m/\sigma, \gamma = \mu/\sigma$.

Rewriting the inverse Gaussian distribution as

$$p(x) = \frac{\delta e^{\delta\gamma}}{\sqrt{2\pi}} x^{-3/2} e^{-(\gamma^2 x + \delta^2 x^{-1})/2}$$

and noting the similarity with the modified Bessel function

$$K_\lambda(z) = \frac{1}{2}\int_0^\infty y^{\lambda-1} e^{-z(y+y^{-1})/2} dy$$

see Watson (1995), for example, the inverse Gaussian distribution can be generalized to

$$p(x) = \frac{(\gamma/\delta)^\lambda}{2K_\lambda(\delta\gamma)} x^{\lambda-1} e^{-(\gamma^2 x + \delta^2 x^{-1})/2}, \quad x > 0$$

This is the *generalized inverse Gaussian distribution* and it does not only generalize the inverse Gaussian distribution but also the gamma distribution. This statement can be proven by letting $\delta \to 0$ and using the small z asymptotic expression

$$K_\lambda(z) \sim \frac{1}{2}\Gamma(\lambda)\,(z/2)^{-\lambda}, \quad \lambda > 0$$

to obtain

$$p(x) = \left(\frac{\gamma^2}{2}\right)^\lambda \frac{1}{\Gamma(\lambda)} x^{\lambda-1} e^{-\gamma^2 x/2}$$

which is the gamma distribution with parameters $(\lambda, 2/\gamma^2)$. If $X \sim \Gamma_{\lambda,\frac{1}{\theta}}$ is gamma distributed, the distribution of $1/X$ is

$$p(x) = \frac{\theta^\lambda}{\Gamma(\lambda)}(1/x)^{\lambda+1} e^{-\theta/x}, \quad x > 0$$

This *inverse gamma distribution* can also be obtained as a special case of the generalized inverse Gaussian distribution. Indeed, when $\lambda < 0$ and $\gamma \to 0$ we can use $K_\lambda(z) = K_{-\lambda}(z)$ to obtain

$$p(x) = \left(\frac{2}{\delta^2}\right)^\lambda \frac{1}{\Gamma(-\lambda)} x^{\lambda-1} e^{-\delta^2/2x}, \quad x > 0$$

It is well known, see Barndorff-Nielsen and Halgreen (1977), that the generalized inverse Gaussian distribution is infinitely divisible and can therefore be used in the construction of a subordinator. We use the subordinator for a drifted Brownian motion according to

$$X_t = \mu t + \beta\tau(t) + W_{\tau(t)}$$

where we assume that $\tau(1)$ is generalized inverse Gaussian distributed with parameters λ, δ and $\sqrt{\alpha^2-\beta^2}$ and is independent of W. The PDF at $t=1$ is given by

$$\begin{aligned}
p_{X_1}(x) &= P\left(X_1 \in [x, x+dx)\right)/dx \\
&= P\left(W_{\tau(1)} \in [x-\mu-\beta\tau(1), x-\mu-\beta\tau(1)+dx)\right)/dx \\
&= \int_0^\infty p_{\tau(1)}(y)\frac{1}{\sqrt{2\pi y}}e^{-(x-\mu-\beta y)^2/2y}dy
\end{aligned}$$

which through a straightforward computation can be seen to be equal to

$$a(\lambda,\alpha,\beta,\delta,\mu)\left(\delta^2+(x-\mu)^2\right)^{(\lambda-1/2)/2}e^{\beta(x-\mu)}K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)$$

where

$$a(\lambda,\alpha,\beta,\delta,\mu) = \frac{\left(\alpha^2-\beta^2\right)^{\lambda/2}}{\sqrt{2\pi}\alpha^{\lambda-1/2}\delta^\lambda K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)}$$

This is known as the *generalized hyperbolic distribution* with parameter values $(\lambda,\alpha,\beta,\delta,\mu)$.

The moment generating function for the generalized inverse Gaussian is obviously equal to

$$M_{\mathrm{GIG}}(k) = \left(\frac{\gamma^2}{\gamma^2-2k}\right)^{\lambda/2}\frac{K_\lambda\left(\delta\sqrt{\gamma^2-2k}\right)}{K_\lambda(\delta\gamma)}$$

from which we obtain the characteristic function

$$\begin{aligned}
\Phi_{\mathrm{GH}(\lambda,\alpha,\beta,\delta,\mu)}(k) &= e^{i\mu k}M_{\mathrm{GIG}(\lambda,\delta,\sqrt{\alpha^2-\beta^2})}(k^2/2+i\beta k) \\
&= e^{i\mu k}\left(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+ik)^2}\right)^{\lambda/2}\frac{K_\lambda\left(\delta\sqrt{\alpha^2-(\beta+ik)^2}\right)}{K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)}
\end{aligned}$$

for the generalized hyperbolic distribution. We conclude that the generalized inverse Gaussian distribution and the generalized hyperbolic distribution have finite moments of arbitrary orders.

Consider the situation when the subordinator $\tau(t)$ is of the form $\sigma_t^2 t$ for some stochastic process σ_t for which σ_1 is generalized inverse Gaussian distributed. As $W_{\sigma_t^2 t}$ has the same distribution as $\sigma_t W_t$, it follows that

$$X_t = \mu t + \sigma_t(\sigma_t\beta t + W_t)$$

in distribution. A consequence of the above is therefore that this stochastic volatility process has a generalized hyperbolic distribution at $t = 1$.

The Lévy measure for the generalized inverse Gaussian distribution is equal to

$$\nu_{\mathrm{GIG}}(dx) = \frac{e^{-\gamma^2 x/2}}{x}\left(\int_0^\infty \frac{e^{-xy}}{\pi^2 y\left(J^2_{|\lambda|}\left(\delta\sqrt{2y}\right) + Y^2_{|\lambda|}\left(\delta\sqrt{2y}\right)\right)}dy + \lambda\mathbb{1}_{\lambda\geq 0}\right)dx$$

according to Barndorff-Nielsen and Shephard (2001). It implies that

$$\nu_{\mathrm{GH}}(dx) = \frac{e^{\beta x}}{|x|}\left(\int_0^\infty \frac{e^{-\sqrt{2y+\alpha^2}|x|}}{\pi^2 y\left(J^2_{|\lambda|}\left(\delta\sqrt{2y}\right) + Y^2_{|\lambda|}\left(\delta\sqrt{2y}\right)\right)}dy + \lambda e^{-\alpha|x|}\mathbb{1}_{\lambda\geq 0}\right)dx$$

where J_λ and Y_λ are the first and second kind Bessel functions.

Having five free parameters, the generalized hyperbolic distribution can generate a large class of distributions and is therefore useful in financial modeling. It is in general not necessary to use all these degrees of freedom and we now consider some instances with fewer parameters for which the modeling is simpler.

Observe that the Bessel function K_λ that appears in the expression for the generalized hyperbolic distribution becomes particularly simple when $\delta \to 0$. Motivated by this fact, we consider the situation when $\delta = 0$, $\mu = 0$ and $\lambda > 0$. This is the *variance gamma distribution* and the characteristic function is given by

$$\Phi(k) = \left(\frac{1}{1 - i\theta v k + \sigma^2 v k^2/2}\right)^{1/v}$$

with $v = 1/\lambda$, $\theta = 2\beta\lambda/(\alpha^2 - \beta^2)$ and $\sigma = \sqrt{2\lambda}/(\alpha^2 - \beta^2)$.

From arguments given earlier in this section, we see that $X_t = \beta\tau(t) + W_{\tau(t)}$ gives a variance gamma distributed X_1 for $\tau(1)$ gamma distributed with parameters $(\lambda, 2/(\alpha^2 - \beta^2))$. We generalize this result by considering $X_t = \xi\tau(t) + \omega W_{\tau(t)}$ for $\tau(t)$ a gamma process, i.e. $\tau(t)$ is a gamma distributed Lévy process:

$$p_{\tau(t)} = \frac{b^{t/v}}{\Gamma(t/v)} x^{t/v-1} e^{-bx}$$

With straightforward computations we see that the characteristic function equals

$$E\left[e^{ik\tau(t)}\right] = \left(\frac{1}{1 - ik/b}\right)^{t/v}$$

and that the right-hand side can be expressed as

$$\exp\left(t\int_0^\infty \left(e^{ikx} - 1\right) v^{-1} x^{-1} e^{-bx} dx\right)$$

We conclude that the Lévy measure is given by

$$\nu(dx) = v^{-1}x^{-1}e^{-bx}\mathbb{1}_{x>0}dx$$

The characteristic function for the gamma subordinated process is equal to

$$\Phi_{X_t}(k) = \left(\frac{1}{1 - i\xi k/b + w^2k^2/2b}\right)^{t/v}$$

Apart from the choice $b = 1/v\sigma^2$, $\xi = \theta/\sigma^2$, $\omega = 1$ used above, the variance gamma distribution can also be obtained from the parametrization $b = 1/v$, $\xi = \theta$, $\omega = \sigma$. The latter is the most commonly used representation for the variance gamma process, i.e. the Lévy process that is variance gamma distributed at $t = 1$.

Observe that the characteristic function decomposes as

$$\left(\frac{1}{1 - i\theta vk + \sigma^2vk^2/2}\right)^{t/v} = \left(\frac{1}{1 - i\eta_+ k}\right)^{t/v}\left(\frac{1}{1 - i\eta_- k}\right)^{t/v}$$

where

$$\eta_\pm = \sqrt{\frac{\theta^2v^2}{4} + \frac{\sigma^2v}{2}} \pm \frac{\theta v}{2}$$

from which it follows that the variance gamma process can be written as the difference of two gamma processes with parameters $(1/v, \eta_+)$ and $(1/v, \eta_-)$, respectively. Using this representation, it was shown in Madan et al. (1998) that the Lévy measure for the variance gamma process is equal to

$$\nu(dx) = \frac{e^{-|x|/\eta}}{v|x|}dx, \quad \eta = \begin{cases} \eta_+ & x > 0 \\ \eta_- & x < 0 \end{cases}$$

Based on this expression, the variance gamma process was generalized in Carr et al. (2002) to a process defined by

$$\nu(dx) = \begin{cases} C\dfrac{e^{-M|x|}}{|x|^{1+Y}} & x > 0 \\ C\dfrac{e^{-G|x|}}{|x|^{1+Y}} & x < 0 \end{cases}$$

where $C, G, M > 0$ and $Y \in (-\infty, 2)$. For $Y < 0$ the characteristic function is given by

$$\Phi(k) = \exp\left(C\Gamma(-Y)\left((M - ik)^Y - M^Y + (G + ik)^Y - G^Y\right)\right)$$

Another important subclass of the generalized hyperbolic distributions is the class of *normal inverse Gaussian distributions*. They are obtained by using inverse Gaussian distributions instead of generalized inverse Gaussian distributions in the generation of the generalized hyperbolic distributions, which means that $\lambda = -1/2$. The characteristic function is

$$\Phi(k) = e^{i\mu k} \frac{e^{\delta\sqrt{\alpha^2-\beta^2}}}{e^{\delta\sqrt{\alpha^2-(\beta+ik)^2}}}$$

from which it follows that the sum of two random variables, which have normal inverse Gaussian distributions with parameters $(\alpha, \beta, \delta_1, \mu_1)$ and $(\alpha, \beta, \delta_2, \mu_2)$, has a normal inverse Gaussian distribution with parameters $(\alpha, \beta, \delta_1 + \delta_2, \mu_1 + \mu_2)$. Thus, unlike the class of generalized hyperbolic distributions, this subclass is closed under convolutions.

Finally, we would like to mention the subclass of *hyperbolic distributions*. They are obtained in the special case when $\lambda = 1$:

$$p(x) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1\left(\delta\sqrt{\alpha^2 - \beta^2}\right)} e^{-\alpha\sqrt{\delta^2+(x-\mu)^2}+\beta(x-\mu)}$$

8.5 Pricing

The pricing of derivatives can as usual be done by assuming that the quotient of the underlying and the numeraire is a martingale. Unfortunately, derivatives cannot be perfectly replicated in markets that support jumps. Markets for which replication is not possible are called *incomplete*. In this situation there is no unique derivatives price and the pricing can be done in various ways, including minimizing (under some appropriate norm) the absolute value of the terminal payoff of being long the derivative and short the replicating strategy.

8.6 Dynamics

Consider the forward European call option price $V = E[(F - K)_+]$, modeled with the lognormal SDE $dF_t = \sigma F_t dW_t$. Dividing by K in both the SDE and the pricing equation, we conclude that V/K depends on F and K only through the moneyness $x = F/K$. This result also holds true when replacing W with another driving process, for instance, a Lévy process. To find the implied volatility of a Lévy model, we need to solve for σ_{imp} in the equation

$$V^{\text{Lévy}}(F, K) = V^{\text{BS}}(F, K, \sigma_{\text{imp}}(F, K))$$

Dividing by K, we obtain

$$\frac{V^{\text{Lévy}}}{K}(x) = \frac{V^{\text{BS}}}{K}(x, \sigma_{\text{imp}}(F, K))$$

from which we see that σ_{imp} must be a function depending only on the moneyness x. We conclude that lognormal Lévy models have sticky-delta dynamics. Obviously, if using alternative SDEs such as of local volatility type, $\sigma = \sigma(F)$, different dynamics are obtained in analogy to the corresponding results for Brownian motions, see Sect. 6.4.

Bibliography

Barndorff-Nielsen OE, Halgreen C (1977) Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. Z Wahrscheinlichkeitstheorie verwandte Geb 38:309–312

Barndorff-Nielsen OE, Shephard N (2001) Non-Gaussian OU based models and some of their uses in financial economics. J Roy Stat Soc Ser B 63:167–241

Bates D (1996) Jumps and stochastic volatility: The exchange rate processes implicit in Deutschemark options. Rev Financial Stud 9:69–107

Carr P, Geman H, Madan D, Yor M (2002) The fine structure of asset returns: an empirical investigation. J Bus 75:305–332

Carr P, Geman H, Madan D, Yor M (2004) From local volatility to local Lévy models. Quant Finance 5:581–588

Madan DB, Carr P, Chang EC (1998) The variance gamma process and option pricing. Eur Finance Rev 2:79–105

Watson GN (1995) A treatise on the theory of bessel functions. Cambridge University Press, Cambridge

Part III
Exotic Derivatives

Chapter 9
Path-Dependent Derivatives

Path-dependent derivatives have payoffs that not only depend on the value S_T of the underlying at maturity but also on the values $\{S_t\}_{0\leq t\leq T}$ attained up to maturity. They can broadly be classified into two categories: weakly and strongly path dependent. The former only depends on the value of the underlying at one or a few instances in time. These points in time might not be known today but can be determined by the future path taken by the underlying. An example is given by the derivative that pays the difference between the maximum value obtained in the time up to maturity and the value at maturity. The maximum only occurs at one instance in time, but viewed from today, we do not know when that will be. In contrast, a strongly path-dependent derivative depends on the whole path. An example is given by *Asian options* that have payoffs linked to the average of $\{S_t\}_{0\leq t\leq T}$ when monitored daily.

Weakly path-dependent derivatives can sometimes be priced with tools similar to those used for vanilla options. For instance, barrier options can be priced by adding boundary conditions to the same PDE used for vanillas. For strongly path-dependent derivatives, on the other hand, the pricing is fundamentally different. Indeed, it is often necessary to introduce variables that depend on the path. For instance, Asian options are usually modeled with the average value of the path as an extra variable. This leads to a PDE of one dimension higher.

Path-dependent derivatives are often priced numerically. The pricing is then done with a model calibrated to vanilla prices. If skew and smile effects are included, the calibration to vanilla instruments is typically done with perturbative methods or through the evaluation of low-dimensional integrals. The model is typically solved by simulating a SDE or by a numerical solution of a PDE. The reason for using numerical solutions for path-dependent derivatives is twofold: first of all, it is often hard to find suitable models for path-dependent options that can be solved by low-dimensional integrals, by simple perturbative techniques or other semi-analytical methods. Secondly, the types of path-dependent products that are popular change from client to client and from year to year. This is in contrast to vanilla products that are few and remain the same over the years. It is therefore useful to have generic

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_9,

methods (by simulating SDEs or solving PDEs) for path-dependent derivatives rather than tailor-made methods.

We saw in Chap. 3 that the derivatives pricing problem can be formulated either in terms of SDEs or PDEs. From a theoretical point of view, these equivalent formulations complement each other: it is sometimes easier to analyze a pricing problem in terms of SDEs and sometimes in terms of PDEs. The same argument applies to numerical pricing. The main numerical difference between the two approaches is that SDEs are simulated forward in time while PDEs compute the price backwards in time from maturity to the pricing date. The PDE computations can be done either by using a finite difference approximation or through a tree structure using an explicit expression for the Green's function. Several investment banks nowadays use a version of PDE solver consisting of recombining trees that are generalizations of binomial and trinomial trees with equally many nodes for each time slice. When going backwards in time and computing expectations, splines are used to connect the node points and the integrals can be solved by fast methods, usually with Gaussian quadrature. The high accuracy allows for large time steps with a substantial performance increase compared to traditional PDE solvers.

With experience, one learns for which derivatives SDEs are best suited and for which PDEs should be used. The basic guidelines are: PDEs usually perform better for low-dimensional problems, i.e. when there are only a small number of variables to track. Because of the technique of working backwards in time, PDE methods handle American and Bermudan features with ease. Furthermore, they return stable risk values. The SDE approach, on the other hand, is better performing in higher dimensions (typically equal to 3 or higher). Also, this approach is often simpler to implement and can be easily generalized to different types of payoffs.

There is a vast array of publications on the implementation of SDEs and PDEs, and we advise the interested reader to consult these. Path-dependent derivatives are sometimes also priced semi-analytically, but as there is currently no consensus on preferred methods, we have decided not to include these methods in the book. We instead focus on formulating the problems mathematically and solve them analytically if possible. The aim is not to provide state-of-the-art formulae but rather to help the reader to develop an intuition about path-dependent derivatives.

We use the constant-parameter lognormal SDE as an illustrating model. Please be aware of the fact that limiting ourselves to a lognormal model means that we do not have any control of the dynamics (which, as we pointed out in Sect. 4.3, is important for the pricing of path-dependent derivatives) or the possibility to calibrate to the skew and the smile. Furthermore, using time-independent parameters means that it is only possible to calibrate to a single maturity. Because of the limited space in this book, we only consider a small selection of products consisting of barrier options, variance swaps, American options and callable products. We believe that this set of products is large enough for the reader to gain familiarity with the techniques and to be able to price general path-dependent products.

Just as we did in Sect. 2.4 for European call options, it is possible to derive no-arbitrage conditions for path-dependent derivatives. For instance, a knock-out option must be worth more than the underlying European option and the price must increase the further away the barrier is from the current value of the underlying. Apart from the brief discussion of barrier options in Chap. 2, we have chosen not to write down the no-arbitrage conditions and parity relations that the various path-dependent derivatives have to satisfy but we instead rely on the reader to work out the details when pricing such contracts.

9.1 Barrier Options

We price a barrier option that knocks out if the underlying goes below an exponentially increasing or decreasing barrier, i.e. if $S_t < Be^{-\lambda(T-t)}$. The usual situation of a constant barrier can be obtained from the special case $\lambda = 0$. For generality, we assume the barrier to be equipped with a *rebate*, which means that an amount $w(t)$ is paid to the option holder if the barrier is hit. Depending on the contract specifications, the rebate $w(t)$ can be paid either at the time t when the barrier is hit or at maturity. We assume the former, meaning that the latter case can be obtained by redefining $w(t) \longrightarrow w(t)e^{-r(T-t)}$. We price the option with a lognormal process and initially follow the corresponding approach for ordinary call options in Sects. 3.2 and 3.4.

The PDE in Sect. 3.2 is formulated in terms of the forward while the barrier condition is in terms of the underlying itself. To avoid this problem we assume a constant interest rate r up to the maturity T. The barrier then has the form $F_t = Be^{(r-\lambda)(T-t)}$.

We assume that a payment of $qSdt$ is received if holding the underlying during $[t, t + dt]$. This payment can represent continuous dividends or a foreign interest rate yield. Generalizing the derivation in Sect. 3.2 gives an extra term $-qFU_F$ in the PDE. The problem then reads

$$\begin{cases} U_t - qFU_F + \frac{1}{2}\sigma^2 F^2 U_{FF} & = 0 \\ U(t = T, F) & = (F - K)_+ \\ U\left(t, F_t = Be^{(r-\lambda)(T-t)}\right) & = w(t) \end{cases}$$

for $F_t > Be^{(r-\lambda)(T-t)}$.

We change variables according to

$$U(t, F) = Ke^{\alpha z + \beta\tau}\Phi(\tau, z)$$

where

$$\begin{cases} \tau = \sigma^2(T - t) \\ z = \ln\left(Fe^{-(r-\lambda)(T-t)}/K\right) = \ln(F/K) - (r - \lambda)(T - t) \end{cases}$$

It implies that

$$\begin{aligned}
U_t &= \left(\frac{\partial \tau}{\partial t}\partial_\tau + \frac{\partial z}{\partial t}\partial_z\right) U = \left(-\sigma^2\partial_\tau + (r-\lambda)\partial_z\right) U \\
&= Ke^{\alpha z+\beta\tau}\left(-\sigma^2(\beta+\partial_\tau) + (r-\lambda)(\alpha+\partial_z)\right)\Phi \\
U_F &= \left(\frac{\partial \tau}{\partial F}\partial_\tau + \frac{\partial z}{\partial F}\partial_z\right) U = \frac{1}{F}\partial_z U = \frac{1}{F}Ke^{\alpha z+\beta\tau}(\alpha+\partial_z)\Phi \\
\Rightarrow U_{FF} &= \frac{1}{F^2}\left(-\partial_z + \partial_z^2\right) U \\
&= \frac{1}{F^2}Ke^{\alpha z+\beta\tau}\left(-\alpha - \partial_z + \alpha^2 + 2\alpha\partial_z + \partial_z^2\right)\Phi
\end{aligned}$$

The Φ_z term in the PDE vanishes if $\alpha = \frac{1}{2} - \frac{r-q-\lambda}{\sigma^2}$ while the Φ term vanishes if $\beta = -\alpha^2/2$. With these choices we obtain the heat equation. Together with the terminal condition and the boundary condition:

$$\begin{cases}
(F-K)_+ &= U(t=T,F) = Ke^{\alpha z}\Phi(\tau=0,z) \\
w(t) &= U\left(t, F_t = Be^{(r-\lambda)(T-t)}\right) = Ke^{\alpha\ln(B/K)-\alpha^2\tau/2}\Phi(\tau,\ln(B/K))
\end{cases}$$

we obtain

$$\begin{cases}
\Phi_\tau - \frac{1}{2}\Phi_{zz} &= 0 \\
\Phi(\tau=0,z) &= e^{-\alpha z}\left(e^z-1\right)_+ \\
\Phi(\tau, z=\tilde{B}) &= \tilde{w}(\tau)
\end{cases}$$

where $\tilde{B} = \ln(B/K)$ and $\tilde{w}(\tau) = \frac{1}{K}e^{-\alpha\tilde{B}+\alpha^2\tau/2}w(T-\tau/\sigma^2)$.

We use the Green's function $p(\tau, z; \tau', z')$, defined by

$$\begin{cases}
p_\tau - \frac{1}{2}p_{zz} &= 0 \\
p(\tau=\tau', z; \tau', z') &= \delta(z-z')
\end{cases}$$

This should be compared with Sect. 3.4 in which we used the related function $\tilde{p}(\tau, z, z') = p(\tau+\tau', z; \tau', z')$ that satisfies the initial condition

$$\tilde{p}(\tau=0,z,z') = \delta(z-z')$$

The initial condition can be moved to the PDE, resulting in

$$\tilde{p}_\tau - \frac{1}{2}\tilde{p}_{zz} = \delta(\tau)\delta(z-z')$$

This follows since if we apply the Laplace transform

$$\mathcal{L}\{\tilde{p}(\tau)\} = \int_0^\infty e^{-su}\tilde{p}(u)du$$

to both PDEs, the identical result

$$-\delta(z-z') + s\mathcal{L}\{\tilde{p}(\tau)\} - \frac{1}{2}\partial_z^2\mathcal{L}\{\tilde{p}(\tau)\} = 0$$

is obtained. According to Lerch's theorem, functions with equal Laplace transforms are unique up to null function, i.e. functions satisfying

$$\int_0^t f(t')dt' = 0, \quad \forall t$$

We therefore obtain an equivalent definition of the Green's function p:

$$p_\tau - \frac{1}{2}p_{zz} = \delta(\tau-\tau')\delta(z-z')$$

Following the approach in Sect. 3.7, we see that the equation in the backward coordinates reads

$$-p_{\tau'} - \frac{1}{2}p_{z'z'} = \delta(\tau-\tau')\delta(z-z')$$

Alternatively, it can be verified from the expression below for p that $p_{\tau'} = -p_\tau$ and $p_{z'z'} = p_{zz}$ are satisfied.

The following general computation can now be done:

$$\begin{aligned}
\Phi(\tau,z) &= \int_0^\infty \int_{\tilde{B}}^\infty \delta(\tau-\tau')\delta(z-z')\Phi(\tau',z')d\tau'dz' \\
&= \int_0^\infty \int_{\tilde{B}}^\infty \left(-p_{\tau'} - \frac{1}{2}p_{z'z'}\right)\Phi(\tau',z')d\tau'dz' \\
&= \int_0^\infty \int_{\tilde{B}}^\infty \Big(-\partial_{\tau'}(p\Phi) + p\Phi_{\tau'} - \frac{1}{2}\partial_{z'}(p_{z'}\Phi) \\
&\qquad + \frac{1}{2}\partial_{z'}(p\Phi_{z'}) - \frac{1}{2}p\Phi_{z'z'}\Big)d\tau'dz' \\
&= \int_{\tilde{B}}^\infty p(\tau,z;0,z')\Phi(0,z')dz' \\
&\qquad + \frac{1}{2}\int_0^\infty \left(p_{z'}(\tau,z;\tau',\tilde{B})\Phi(\tau',\tilde{B}) - p(\tau,z;\tau',\tilde{B})\Phi_{z'}(\tau',\tilde{B})\right)d\tau'
\end{aligned}$$

The values of $\Phi(0,z')$ and $\Phi(\tau',\tilde{B})$ are known from the initial condition and the boundary condition, but not $\Phi_{z'}(\tau',\tilde{B})$. To obtain an expression for $\Phi(\tau,z)$ that does not contain this unknown value, it is necessary impose the condition $p(\tau,z;\tau',\tilde{B})=0$, $\forall\tau' = 0$. It is at this stage where our solution starts to differ from the pricing of vanilla options in Sect. 3.4, where we implicitly assumed the boundary condition $p \to 0$ for $|z| \to \infty$ and obtained

$$p_0(\tau,z;\tau',z') = \frac{1}{\sqrt{2\pi(\tau-\tau')}}e^{-(z-z')^2/2(\tau-\tau')}$$

which is called the *fundamental solution* of the PDE. It is possible to use the fundamental solution to construct a Green's function that satisfies the boundary condition for a barrier:

$$p(\tau,z;\tau',z') = p_0(\tau,z;\tau',z') - p_0(\tau,z;\tau',2\tilde{B}-z')$$

This function clearly satisfies the boundary condition at $\tilde{B}$ and also the PDE as

$$\begin{aligned}&\left(\partial_\tau - \frac{1}{2}\partial_{zz}\right)(p_0(\tau,z;\tau',z') - p_0(\tau,z;\tau',2\tilde{B}-z'))\\ &\quad = \delta(\tau-\tau')\delta(z-z') - \delta(\tau-\tau')\delta(z+z'-2\tilde{B}) = \delta(\tau-\tau')\delta(z-z')\end{aligned}$$

where the last equality holds because $z, z' > \tilde{B}$

Using the Green's function, we obtain the solution

$$\begin{aligned}\Phi(\tau,z) &= \int_{\tilde{B}}^{\infty} \frac{1}{\sqrt{2\pi\tau}}\left(e^{-(z-z')^2/2\tau} - e^{-(z+z'-2\tilde{B})^2/2\tau}\right)e^{-\alpha z'}\left(e^{z'}-1\right)_+ dz'\\ &\quad + \int_0^{\infty} \frac{1}{\sqrt{2\pi(\tau-\tau')}}\frac{z-\tilde{B}}{\tau-\tau'}e^{-(z-\tilde{B})^2/2(\tau-\tau')}\tilde{w}(\tau')d\tau'\end{aligned}$$

For simplicity, we only evaluate this expression for zero rebate, i.e. $\tilde{w}(\tau')=0$, and we assume that the barrier is below the strike $B<K$. The instance of a constant barrier above the strike can then be obtained by using static replication methods, see Sect. 2.6. The first of the two remaining terms in the above expression equals

$$\begin{aligned}I(\tau,z) &= \int_{\tilde{B}}^{\infty} \frac{1}{\sqrt{2\pi\tau}}e^{-(z-z')^2/2\tau}e^{-\alpha z'}\left(e^{z'}-1\right)_+ dz'\\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi\tau}}e^{-(z-z')^2/2\tau}e^{-\alpha z'}\left(e^{z'}-1\right)dz'\end{aligned}$$

since $\tilde{B} < 0$. The second term is given by

$$\begin{aligned}
\int_0^\infty \frac{1}{\sqrt{2\pi\tau}} e^{-(z-z')^2/2\tau} e^{-\alpha z'} dz' &= \frac{1}{\sqrt{2\pi}} e^{-\alpha z} \int_{-z/\sqrt{\tau}}^\infty e^{-z'^2/2} e^{-\alpha z'\sqrt{\tau}} dz' \\
&= e^{-\alpha z + \alpha^2\tau/2} N\left(\frac{z}{\sqrt{\tau}} - \alpha\sqrt{\tau}\right)
\end{aligned}$$

The first term can then be obtained by replacing α with $\alpha - 1$. We now arrive at

$$\begin{aligned}
\Phi(\tau, z) &= I(\tau, z) - I(\tau, 2\tilde{B} - z) \\
&= e^{-(\alpha-1)z + (\alpha-1)^2\tau/2} N\left(\frac{z}{\sqrt{\tau}} + (1-\alpha)\sqrt{\tau}\right) \\
&\quad - e^{-\alpha z + \alpha^2\tau/2} N\left(\frac{z}{\sqrt{\tau}} - \alpha\sqrt{\tau}\right) \\
&\quad - e^{-(\alpha-1)(2\tilde{B}-z) + (\alpha-1)^2\tau/2} N\left(\frac{2\tilde{B} - z}{\sqrt{\tau}} + (1-\alpha)\sqrt{\tau}\right) \\
&\quad + e^{-\alpha(2\tilde{B}-z) + \alpha^2\tau/2} N\left(\frac{2\tilde{B} - z}{\sqrt{\tau}} - \alpha\sqrt{\tau}\right) \\
\Rightarrow U(t, F) &= K e^{\alpha z - \alpha^2\tau/2} \Phi(\tau, z) \\
&= K e^{z - \alpha\tau + \tau/2} N\left(\frac{z}{\sqrt{\tau}} + (1-\alpha)\sqrt{\tau}\right) \\
&\quad - K N\left(\frac{z}{\sqrt{\tau}} - \alpha\sqrt{\tau}\right) \\
&\quad - K e^{-2(\alpha-1)\tilde{B} + (2\alpha-1)z - \alpha\tau + \tau/2} N\left(\frac{2\tilde{B} - z}{\sqrt{\tau}} + (1-\alpha)\sqrt{\tau}\right) \\
&\quad + K e^{-2\alpha\tilde{B} + 2\alpha z} N\left(\frac{2\tilde{B} - z}{\sqrt{\tau}} - \alpha\sqrt{\tau}\right) \\
&= F e^{-q(T-t)} N\left(\frac{\ln(F/K) - q(T-t)}{\sigma\sqrt{T-t}} + \frac{1}{2}\sigma\sqrt{T-t}\right) \\
&\quad - K N\left(\frac{\ln(F/K) - q(T-t)}{\sigma\sqrt{T-t}} - \frac{1}{2}\sigma\sqrt{T-t}\right) \\
&\quad - F e^{-q(T-t)} \left(\frac{F}{B} e^{-(r-\lambda)(T-t)}\right)^{2(\alpha-1)}
\end{aligned}$$

$$
\begin{aligned}
&N\left(\frac{\ln(B^2/FK)+2(r-\lambda)(T-t)-q(T-t)}{\sigma\sqrt{T-t}}+\frac{1}{2}\sigma\sqrt{T-t}\right)\\
&+K\left(\frac{F}{B}e^{-(r-\lambda)(T-t)}\right)^{2\alpha}\\
&N\left(\frac{\ln(B^2/FK)+2(r-\lambda)(T-t)-q(T-t)}{\sigma\sqrt{T-t}}-\frac{1}{2}\sigma\sqrt{T-t}\right)\\
\Rightarrow V(t,S) = {}& Se^{-q(T-t)}N\left(\frac{\ln(S/K)+(r-q)(T-t)}{\sigma\sqrt{T-t}}+\frac{1}{2}\sigma\sqrt{T-t}\right)\\
&-KN\left(\frac{\ln(S/K)+(r-q)(T-t)}{\sigma\sqrt{T-t}}-\frac{1}{2}\sigma\sqrt{T-t}\right)\\
&-\frac{B^2}{S}e^{-(q+2\lambda)(T-t)}\left(\frac{Se^{\lambda(T-t)}}{B}\right)^{2\alpha}\\
&N\left(\frac{\ln(B^2/SK)+(r-q-2\lambda)(T-t)}{\sigma\sqrt{T-t}}+\frac{1}{2}\sigma\sqrt{T-t}\right)\\
&+K\left(\frac{Se^{\lambda(T-t)}}{B}\right)^{2\alpha}N\left(\frac{\ln(B^2/SK)+(r-q-2\lambda)(T-t)}{\sigma\sqrt{T-t}}\right.\\
&\qquad\left.-\frac{1}{2}\sigma\sqrt{T-t}\right)
\end{aligned}
$$

Denote the sum of the first two terms by $C_q(t,S)$. This is the price of a call option without a barrier but with dividend payments. We can then formulate the price of the barrier option as

$$
V(t,S) = C_{q+\lambda}\left(t, Se^{\lambda(T-t)}\right) - \left(\frac{Se^{\lambda(T-t)}}{B}\right)^{1-(r-q-\lambda)/\sigma^2} C_{q+\lambda}\left(t, \frac{B^2}{Se^{\lambda(T-t)}}\right)
$$

When $\lambda = 0$ we see from this formula that the barrier condition has been satisfied by adding a reflective term to the European call option price, by replacing S with B^2/S and multiplying by a certain factor. The condition $V(t,B) = 0$ is then obviously true. We also note that the formula for an exponential barrier $\lambda \neq 0$ is obtained from the constant barrier case by replacing S with $Se^{\lambda(T-t)}$ and q with $q+\lambda$.

Assume that a barrier option is close to knock out at a certain observation point. As the barrier in practice is often only monitored at certain discrete time points, for example, by daily observations, there is a chance that the option would have knocked out in the time interval between the previous and current observation point if the observation had been continuous. It means that if we use a model with a continuous barrier, the model barrier B' needs to be below the product barrier B, assuming that the barrier is located below the underlying. In Broadie et al. (1997), an approximative relation between a continuously and discretely observed barrier was derived:

$$B' = Be^{\sigma\sqrt{T}\zeta(1/2)/\sqrt{2\pi m}}$$

where it was assumed that the underlying follows a lognormal process. ζ is the Riemann zeta function with $\zeta(1/2) \approx -1.46$ and m is the number of regularly distributed observation points.

It is often not sufficient to use lognormal processes for pricing path-dependent derivatives because of the poor agreement with vanilla prices. To calibrate to a rich set of vanilla instruments, more sophisticated models have to be used, e.g. local or stochastic volatility models. The pricing is typically done numerically by using n number of discrete time steps. Because of performance limitations, n is typically smaller than the number of observations stated in the contract. Using the above result, the level B'' of the barrier in the implementation is approximately related to the barrier B in the contract according to

$$B'' = Be^{\sigma\sqrt{T}\zeta(1/2)\left(\frac{1}{\sqrt{m}}-\frac{1}{\sqrt{n}}\right)/\sqrt{2\pi}}$$

The value of σ is determined from the lognormal process that best matches the implemented model. This result only holds if the implemented model is not too different from a lognormal model. If the implemented model is locally (in time) close to a lognormal process, the result can be generalized to a time-dependent B''. In Broadie et al. (1999) it is shown how the method of shifting a parameter, such as the barrier, can be used for other path-dependent derivatives such as a lookback option, for which the distinction between continuous and discrete monitoring is important.

There are several proposed semi-analytical methods that can be used when not all of the above assumptions are satisfied, e.g. in the presence of non-lognormal processes, time-dependent parameters or when having a double barrier. Another solution method is, as we saw in Sect. 2.7, to statically replicate a knock-in call by a put and a strip of smaller digital puts with shorter maturities. We now show that it is actually possible to do the static replication using only European options with equal maturities.

We price a down-and-in call option with barrier $B < K$ and maturity T. We do the pricing using an auxiliary European option with a payoff $h(S)$ such that $h(S) = 0$ if $S > B$, see Fig. 9.1. We then determine $h(S)$ so that this option is worth as much as the underlying European option along the barrier. With similar arguments as in Sect. 2.7 it follows that the knock-in has the same price as the European option with payoff $h(S)$. This price can be determined as the latter option can be priced through static replication of a bond, the underlying and European puts, see Sect. 2.3.

It remains to determine $h(S)$ so that the conditional expectations are equal:

$$E_t[(S - K)_+] = E_t[h(S)]$$

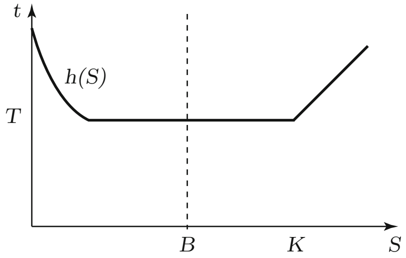

Fig. 9.1 Relication of a knock-in call by a European option with payoff $h(S)$

for all times t and for $S_t = B$. Assuming constant interest rates and a lognormal process for the underlying gives with $\gamma = r - \frac{1}{2}\sigma^2$ the result

$$\begin{aligned}
e^{r\tau} BN(d_+) - KN(d_-) &= \int_0^B \frac{1}{\sqrt{2\pi\sigma^2\tau}S} e^{-(\ln(S/B)-\gamma\tau)^2/2\sigma^2\tau} h(S)dS \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau}} e^{-(x+\gamma\tau)^2/2\sigma^2\tau} h(Be^{-x})dx \\
&= e^{-\gamma^2\tau/2\sigma^2} \int_0^\infty \frac{1}{2\sqrt{\pi x\tau}} e^{-x/\tau-\gamma\sqrt{2x}/\sigma} h\big(Be^{-\sigma\sqrt{2x}}\big)dx \\
\Leftrightarrow \int_0^\infty & e^{-x\tau} \frac{1}{\sqrt{x}} e^{-\gamma\sqrt{2x}/\sigma} h(Be^{-\sigma\sqrt{2x}})dx \\
&= e^{\gamma^2/2\sigma^2\tau} \frac{2\sqrt{\pi}}{\sqrt{\tau}} \bigg(e^{r\tau^{-1}} BN\bigg(\frac{\ln(B/K) + r\tau^{-1}}{\sigma\tau^{-1/2}} + \frac{1}{2}\sigma\tau^{-1/2} \bigg) \\
& -KN\left(\frac{\ln(B/K) + r\tau^{-1/2}}{\sigma\tau^{-1}} - \frac{1}{2}\sigma\tau^{-1/2} \right) \bigg)
\end{aligned}$$

$h(S)$ can be computed from this relation as the right-hand side is the Laplace transform of $\frac{1}{\sqrt{x}} e^{-\gamma\sqrt{2x}/\sigma} h\big(Be^{-\sigma\sqrt{2x}}\big)$.

Observe that this static replication technique, as well as the one in Sect. 2.7, is model dependent. For instance, the equality in payoffs was only computed with the present value of the volatility and will not be valid at later times should the implied volatility change.

9.2 Volatility Products

If being long a European call option, we make a gain with increasing underlying value or with increasing volatility. We have therefore taken a position in the underlying value as well as in the volatility. If we instead want a pure exposure to the underlying value, this can be obtained by purchasing the underlying itself or by entering a forward or futures contract. This brings us to the question of how to obtain a pure exposure to the volatility, which is the topic of this section.

One approach to obtain a pure volatility dependence is to hold an option that is delta hedged. Unfortunately, dynamic replication is not exact in practice because of the idealized assumptions of: imperfect modeling of the underlying process, discrete-time trading, presence of transaction costs, etc. It means that even after the delta hedge, there remains some exposure to the underlying. Another disadvantage of this approach is that the moneyness of the option changes during the lifetime: sometimes the option is close to ATM and sometimes it is far away, depending on the underlying value. As the vega is heavily dependent on the moneyness, the volatility dependence will be unpredictable with the passage of time.

A more direct way to gain exposure to the volatility is through volatility products. They mainly consist of *variance swaps* and *volatility swaps*. The former has a payment at maturity T proportional to

$$\frac{1}{n-1}\sum_{i=1}^{n}(\ln(S_{t_i}/S_{t_{i-1}}))^2 - K$$

while the latter has a time T payment proportional to

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\ln(S_{t_i}/S_{t_{i-1}}))^2} - K$$

The observations are usually made on a daily basis.

We now describe how variance swaps can be priced in the limit $n \to \infty$ of continuous observations. The underlying returns $(S_{t_i} - S_{t_{i-1}})/S_{t_{i-1}}$ are then small meaning that we can use Taylor approximation twice on $\ln(S_{t_i}/S_{t_{i-1}})$ to obtain

$$\begin{aligned}
\sum_{i=1}^{n}(\ln(S_{t_i}/S_{t_{i-1}}))^2 &= \sum_{i=1}^{n}\left(\ln\left(1+\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}}\right)\right)^2 \\
&\approx \sum_{i=1}^{n}\left(\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}}\right)^2 \\
&= 2\sum_{i=1}^{n}\left(\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}} - \left(\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}} - \frac{1}{2}\left(\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}}\right)^2\right)\right) \\
&\approx 2\sum_{i=1}^{n}\left(\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}} - \ln(S_{t_i}/S_{t_{i-1}})\right) \\
&= 2\sum_{i=1}^{n}\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}} - 2\ln(S_{t_n}/S_{t_0})
\end{aligned}$$

We assume zero interest rates and consider the zero-cost t_{i-1} investment of $1/S_{t_{i-1}}$ underlyings and –1 bonds. This gives a payment of $(S_{t_i} - S_{t_{i-1}})/S_{t_{i-1}}$ at t_i which means that the first term on the right-hand side can be obtained with zero initial investment. Thus, the two time T payments of $\sum_{i=1}^{n}(\ln(S_{t_i}/S_{t_{i-1}}))^2$ and $-2\ln(S_T/S_0)$ are worth equally much. We also know how to price the latter payment with static replication. We have thereby shown how variance swaps can be priced with static replication when interest rates are zero and in the limit of continuous observations. We then obtain the interesting result that this approximation of the variance swap price only depends on the underlying process through European option prices. Assuming a lognormal process for the underlying, we obtain $(dS/S)^2 = \sigma^2 dt$ and

$$\sum_{i=1}^{n}\left(\frac{S_{t_i} - S_{t_{i-1}}}{S_{t_{i-1}}}\right)^2 \approx \sum_{i=1}^{n}\sigma_i^2 dt \to \int_0^T \sigma^2 dt$$

which explains the name variance swap.

Consider now the pricing of a variance swap at the date $t = t_k$. The payment can then be decomposed into a known part and an unknown part

$$\sum_{i=1}^{n}(\ln(S_{t_i}/S_{t_{i-1}})) = \sum_{i=1}^{k}(\ln(S_{t_i}/S_{t_{i-1}})) + \sum_{i=k+1}^{n}(\ln(S_{t_i}/S_{t_{i-1}}))$$

It means that historical information of the underlying is necessary for the pricing. This is a typical property of strongly path dependent options.

9.3 American Options

American style options were defined in Sect. 2.5 as options that can be exercised at any time up to the maturity. The advantage of not exercising early is that the exercise decision can be based on information about the underlying all the way up to the maturity. The exercise decision also depends on the discounting of the strike from the maturity to the pricing date: the longer we wait with the exercise, the smaller the value of the discounted strike. These factors are both in favor of not exercising American call options early, which means that these options must have the same price as their European counterparts. For put options, on the other hand, the discounting effect encourages early exercise, which means that there are situations when this is the optimal strategy. With similar arguments it can be proven that American call options on futures contracts should never be exercised early while the corresponding statement for put options is not necessarily true.

In parallel to our treatment of barrier options, we assume that the underlying pays continuous dividends to obtain more general and non-trivial results. An early exercise then yields dividend payments and might therefore be feasible for call options as well.

To start off lightly, we first price an American digital option that pays $\theta(S-K)$ upon exercise. Whenever the value of S exceeds K, the holder is entitled a payment of one dollar. As it is preferable to receive one dollar sooner than later in time, the optimal strategy is to exercise the option as soon as $S \geq K$. This product is therefore equivalent to a digital up and out barrier option that pays a rebate of one dollar and can be priced as in Sect. 9.1.

Assume that the condition $S=K$ means that the price of a European digital option is close to $\frac{1}{2}$ as there is about as high probability of S_T to end up below as above K. An American digital call can then be replicated by two European digital call options as both strategies are worth 1 if the barrier is hit and are worthless otherwise.

The pricing is less trivial for American call (or put) options on dividend-paying underlyings. The reason is that the exercise boundary is not known in advance. As there are no known analytical solutions, numerical or semi-analytical methods have to be used for these products.

We set up the pricing equations for American call options on dividend-paying underlyings. The instance of put options can be handled in a similar way. First, consider the situation when the underlying has a very high value. We are then certain to end up ITM at maturity and it can therefore be preferable to exercise the option early to cash in the dividend payments. If the underlying has a low value, on the other hand, it is not advisable to exercise early. We conclude that for each time t there exists a boundary point B_t, possibly equal to infinity, such that the option should be exercised if $S_t \geq B_t$. The early exercise boundary is given by the function B_t, where $t \in [0, T]$.

We find it more convenient to formulate the problem in terms of the spot variables S and V rather than using the forward variables F and U. Assuming a lognormal underlying, the pricing problem can be written as

$$\begin{cases} -rV + V_t + (r-q)SV_S + \frac{1}{2}\sigma^2 S^2 V_{SS} & = 0 \\ V(t=T, S) & = (S-K)_+ \\ V\left(t, S_t = B_t\right) & = B_t - K \end{cases}$$

where $S_t \leq B_t$ and we have assumed a time-independent interest rate and dividend yield. The exercise boundary B_t should be determined so that V becomes maximized. This type of problem is called a *free-boundary problem.*

The option value is obviously continuous across the boundary. We now argue that the first-order (mathematical) derivative is also continuous, a consequence of the fact that the exercise boundary is chosen to optimize the option value. Indeed, for any given time, let $V(S, B)$ be the option value and $h(B)$ the value if exercised early, e.g. $h(B) = B - K$ for a call option. The continuity condition can then be written as $h(B) = V(S, B)|_{S=B}$. Taking the B derivative on both sides gives

$$\frac{dh}{dB} = \left.\frac{\partial V}{\partial S}\right|_{S=B} + \left.\frac{\partial V}{\partial B}\right|_{S=B} = \left.\frac{\partial V}{\partial S}\right|_{S=B}$$

where the last equality follows as B has been chosen to optimize V.

Before discussing techniques for determining the early exercise boundary, we describe its shape qualitatively. Assume first that the maturity is very short. Clearly, $B_T > K$ as we do not want to exercise an OTM option. To determine when an ITM option should be exercised, the proceeds of an early exercise have to be compared with an exercise at maturity. An early exercise gives $S - K$ which is equivalent with a payment of $Se^{q(T-t)} - Ke^{r(T-t)} \approx S - K + Sq(T-t) - Kr(T-t)$ at T. Thus, the option should only be exercised early when

$$S - K + Sq(T-t) - Kr(T-t) > S - K \Leftrightarrow S > \frac{r}{q}K$$

as more cash is generated from the dividends than from the interest rate of being short K amount of cash. We conclude that $B_T > K \max\{1, r/q\}$.

We argued in Sect. 4.3 that the European option price typically decreases with t if all other variables are unchanged. The American option price must also decrease with t as we lose the optionality of exercising in $[t, t')$ when time goes from t to t'. Because the early exercise boundary is where an American call option price is equal to $S - K$, the decreasing option price implies that the exercise boundary B_t decreases with t.

Let us now determine the exercise boundary B_t when the maturity T is far in the future. The long maturity implies a weak t-dependence that can be omitted in the lowest-order approximation. We obtain the PDE

$$\begin{cases} -rV + (r-q)SV_S + \frac{1}{2}\sigma^2 S^2 V_{SS} & = 0 \\ V(S = B) & = B - K \end{cases}$$

which is of Euler type and has the solution

$$V = A_+ S^{\mu_+} + A_- S^{\mu_-}, \quad \mu_\pm = \frac{-(r - q - \sigma^2/2) \pm \sqrt{(r - q - \sigma^2/2)^2 + 2\sigma^2 r}}{\sigma^2}$$

The requirement of a well-defined solution at $S = 0$ gives $A_- = 0$ while the condition at the exercise boundary implies that

$$V = (B - K)\left(\frac{S}{B}\right)^{\mu_+}$$

Maximizing this solution with respect to the early exercise point B gives

$$B = K\frac{\mu_+}{\mu_+ - 1}$$

We thereby conclude that the exercise boundary decreases from the value $K\frac{\mu_+}{\mu_+ - 1}$ when far from maturity to a value $B_T > K \max\{1, r/q\}$ at maturity.

We now briefly discuss some techniques that can be used to price American options. We start by observing that the American call option price can be expressed semi-analytically as a function of the underlying and the exercise boundary: $V = V(t, S_t, \{B_{t'}\}_{t<t'<T})$, with the continuity condition of the form $V(t, B_t, \{B_{t'}\}_{t<t'<T}) = B_t - K$. The exercise boundary can be determined iteratively by taking small time steps backwards from the maturity T until today's date. In each step, the option price is computed whereafter the boundary can be determined. The computations are tedious because of the complicated expression for V. An alternative approach is to use the formula for the continuity of the first derivative instead of the continuity of the price, as it gives slightly simpler equations.

The exercise boundary can also be computed by using the fact that it optimizes $V(t, S_t, \{B_{t'}\}_{t<t'<T})$. For example, the boundary can be parametrized with a function that depends on a small number of variables. The boundary can then be determined by maximizing the price with respect to these variables. The choice of function should be guided by our knowledge of its value at maturity and that it increases (for call options) with the distance to maturity to a known limiting value. If choosing an exponential boundary, the analytic expressions of Sect. 9.1 can be used to obtain high-performing calculations.

An immediate exercise of an ITM option gives the lower bound of $(S - K)_+$ on the American call option price. If, on the other hand, we choose to never exercise the option, i.e. $B_t \rightarrow \infty$, we see that the European price is a lower bound. For general choices of B_t, more interesting lower bounds can be found such as those obtained from a constant or exponential exercise boundary. Unfortunately, it is hard to know how tight these bounds are (i.e. how far they are from the optimal exercise). This question can be answered if tight upper bounds are found and much research has been devoted to this. We saw a particularly simple upper bound in our discussion of static replication of Americans in Sect. 2.5 and in the same way it is possible to find an upper bound for American call options on dividend paying underlyings.

When it comes to pure numerical solutions, American options can, just like exotic options, be priced with a PDE solver or an SDE simulator. The preferred solution is often through PDEs as they work backward in time and enable us to determine the optimal exercise boundary iteratively. This is done by comparing the intrinsic option value with the exercise value in each node and choosing the maximum of the two. When it comes to simulation, there have been major developments in Monte Carlo methods for American option pricing during the last decade and satisfactory results are now produced.

There also exist several types of semi-analytic techniques for pricing American options. One of the simplest is that of Bermudan approximation. The American option is then approximated with a Bermudan option V_n that can be exercised at n points in time. As the analytical computations for Bermudan options become increasingly complex with increasing n, they are often only carried out for small values, $n = 1, 2$ and 3. The American option price V_∞ can then be estimated with extrapolation techniques.

9.4 Callable Products

We will here mainly be concerned with swaps, i.e. products that exchange one type of cash flow for another. The coupons are typically determined from an equity index, a commodity price, an FX rate, an interest rate or are equal to a constant. The cash flows are obtained by multiplying the coupons with a notional or a quantity and possibly an FX rate (for interest rates there is also a multiplication with a day count fraction). An example of a swap is given by the product that once per year exchanges the oil price for a fixed price (multiplied by the number of barrels) during five years' time.

A swap is said to be *callable* if there exists a triggering event that results in the cancelation of the future cash flows. A callable swap is said to be a *Bermudan swaption* if the triggering event is such that one of the counterparties can cancel the deal at any cash flow date. The triggering event can also be a market event. The swap is then said to be *autocallable*. An example of an autocallable is a fixed-for-floating swap that gets canceled when the underlying (that determines the floating flows) reaches a certain level.

A closely related product is the *enter-into swap* for which a trigger event leads to an activation of the cash flows. By going long this product and short the corresponding vanilla swap, a cancelable swap is obtained. Because of this parity relation, we focus exclusively on callable swaps.

There are several possible types of trigger events that can be used for autocallables. For instance, a swap can be canceled when an external index, i.e. different from the underlying that determines the floating flows, reaches a certain level. Another popular type of autocallables is represented by the *target redemption notes (TARNs)*, for which the cancelation occurs when the cumulated floating coupons exceed a certain level.

There also exist autocallables that depend on several trigger events. One example is the *auto cap* which just like an ordinary cap consists of a set of caplets, see Sect. 13.3. The auto cap gets canceled when n of the caplets have ended up ITM at their maturities, where n is an integer determined at the trade date. A related product is the *chooser cap*, where the receiver of the caplets decides if the cash flows should be paid out, with a maximum of n received payments.

It is difficult to find suitable analytical or semi-analytical formulae for callable swaps and these products are therefore priced numerically via PDE or SDE methods. In fact, it can even be a hard task to find a suitable numerical method. For instance, the choice of the optimal exercise date for Bermudan swaptions is a non-trivial problem when using Monte Carlo simulations and is often solved by using methods similar to those developed in Longstaff and Schwartz (2001) and Andersen (2000). An example of the opposite situation, when a Monte Carlo simulation is easy while a tree implementation is harder to use, is found in the pricing of TARNs. This problem is usually solved by adding another dimension to the tree, representing discrete levels of the cumulative coupons.

The reader might at this stage wonder why we even bother about a difficult method when there exists a simpler method. There are several reasons for doing so. First of all, we might be located at a trading desk with limited resources and can therefore only afford to implement one of the numerical methods, typically the Monte Carlo simulation as it is more flexible and easier to work with. In this situation we must solve all problems with Monte Carlo simulations, even the ones more suitable for trees. Another reason can be that the model is only suitable for one of the numerical methods. For instance, the high dimensionality of the LMM model makes it hard to find a suitable tree implementation, see Sect. 13.17.

It is also common to encounter products that contain some features not suitable for tree pricers and some other features not suitable for Monte Carlo pricers. An example is given by Bermudan swaptions for which the coupons depend on the previous coupon payments. The simplest, and the most common, of these products has coupons that depend only on the previous coupon. The relation between the coupons is typically positive, meaning that a high previous coupon makes it likely that the current coupon will be high as well. The consequence can be a runaway effect with increasing coupons. Products with coupon dependencies are called *snowballs* because of the analogy of a snowball rolling down a hill and growing larger and larger. A Monte Carlo pricer can easily incorporate the coupon dependencies while the callability is harder to handle. For a tree pricer, on the other hand, it is the coupon dependency that makes the problem difficult.

Another example of a product for which both a Monte Carlo and a tree implementation are non-trivial is the chooser cap. The Monte Carlo pricer typically has to include the number of exercised caplets among the state variables in a generalization of the Longstaff and Schwartz or Andersen methods. For the tree, the number of exercised caplets can be included as an extra dimension of the tree.

Barrier options and American options can be viewed as cancelable products for which the cancelation affects a single cash flow. For general callable swaps, on the other hand, the triggering event affects several cash flows, which results in a higher sensitivity to the input parameters and therefore a risk that is hard to compute.

Because of the digital feature of the triggering event in a callable swap, the computation of the risk is arguably the most important, and the most difficult, task for a pricer. The trick is often to make a modification to a more conservative product for which the risk is better behaved, as was done in Sect. 4.4 for digital options and barrier options. For instance, the method of improving the risk for digital options can be viewed as letting the notional go linearly to zero in an interval next to the strike. Similarly, we could let the notional of a TARN go linearly to zero in an interval next to the target level. Similar ideas can be applied to other types of callable swaps.

The risk computations for Bermudan swaptions are straightforward for tree pricers but can be more complicated when using Monte Carlo simulations. Fortunately, there is a trick that simplifies the computation of the first-order greeks. To explain how it works, note that the pricing is done by first computing the early exercise boundary and then doing the pricing itself. As the early exercise boundary is determined from an optimizing condition, that of obtaining the best possible payoff, it remains unchanged under first-order changes of the model parameters.

It means that the early exercise boundary does not need to be computed in the bump-and-revalue process. Avoiding the recomputation of the early exercise boundary increases the performance and gives more stable first-order greeks.

If a Bermudan swaption is such that there is only a small chance of it to be called, a Monte Carlo simulator does not effectively determine the early exercise boundary as only a few paths contribute to this information. For the purpose of determining the early exercise boundary it is a good idea to shift the current value of the underlying to a value for which an early exercise is more probable. This leads to a higher accuracy of the early exercise boundary as it is independent of the current value of the underlying.

Bibliography

Andersen L (2000) A Simple approach to the pricing of bermudan swaptions in the multifactor LIBOR Market Model. J Comput Finance 3:5–32

Broadie M, Glasserman P, Kou S (1997) A continuity correction for discrete barrier options. Math Finance 7:325–349

Broadie M, Glasserman P, Kou S (1999) Connecting discrete and continuous path-dependent options. Finance Stochast 3:55–82

Longstaff FA, Schwartz ES (2001) Valuing american options by simulation: a simple least-squares approach. Rev Financ Stud 14:113–147

Chapter 10
High-Dimensional Derivatives

We cover the pricing of path-independent derivatives such that the payoff at a given time T depends on the values of several underlyings $F_1, F_2, \ldots, F_n$. The general case of path-dependent higher-dimensional derivatives can be handled by combining the methods of this chapter with those in Chap. 9. The fundamental theorem of asset pricing states that the price can be computed from

$$E\left[g(F_1, F_2, \ldots, F_n)\right]$$

where g is the payoff function and the expectation is in the T-forward measure.

Just as for path-dependent derivatives, there are several possible payoff types. In a similar way, we here also choose to leave it as an exercise for the reader to work out the no-arbitrage conditions and parity relations. We assume that the individual distributions of $F_1, F_2, \ldots, F_n$ are known from prices of vanilla options. As the expectation not only depends on the marginal distributions, but also on the joint distribution, it is necessary to account for the correlation between the underlyings.

The first solution that comes to mind when attacking higher-dimensional problem is the use of copulas. We start by covering this technique. Although a general method, the implementation is often of low performance. We therefore proceed by considering special cases for which high-performance implementations are possible. The methods are: variable freezing, moment matching, quadratic functional modeling and the change of measure technique. We also consider the important special cases of digital options and spread options. We end the chapter with discussions of correlations and calibration.

10.1 Copulas

The natural approach for obtaining a joint distribution of $F_1, F_2, \ldots, F_n$ from the marginal distributions is to use a copula, see Appendix. We usually have an idea about the value of the correlation between the underlyings, but not for the

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_10,

higher-order moments. It therefore makes sense to use the simplest possible copula that only depends on the correlation matrix, which happens to be the Gaussian copula. Unfortunately, just as the normal variable assigns too low a probability to extreme events in financial applications, the Gaussian copula implies too weak a correlation between extreme events. For this purpose, alternative copulas are often used in financial modeling, in particular since the financial crisis starting in 2007 which has been claimed to be partly the result of underestimating extreme event correlations by using Gaussian copula. To focus on the essentials, however, we only consider the simple case of the Gaussian copula.

To gain a better understanding of the Gaussian copula, let us limit ourselves to the instance of two underlyings: F_1 and F_2. Motivated by the form of several of the SDEs in Chap. 5, we assume that the distributions can be written as $F_i = h_i(X_i)$, where $\{X_i\}$ are standard normal variables and $\{h_i\}$ are monotonic increasing functions. We assume that X_1 and X_2 have correlation ρ and are the components of a 2-dimensional Gaussian variable. The joint distribution is then given by

$$P(F_1 < f_1, F_2 < f_2) = P(X_1 < h_1^{-1}(f_1), X_2 < h_2^{-1}(f_2))$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{h_1^{-1}(f_1)} \int_{-\infty}^{h_2^{-1}(f_2)} \exp\left(-\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)/2\left(1-\rho^2\right)\right) dz_1 dz_2$$

Observe that if we instead had correlated F_1 and F_2 by using their marginal distributions

$$P(F_i < f_i) = P(X_i < h_i^{-1}(f_i)) = N\left(h_i^{-1}(f_i)\right)$$

in the Gaussian copula, see Appendix, we would have obtained the same expression for $P(F_1 < f_1, F_2 < f_2)$. We conclude that using the Gaussian copula is the correct way to patch together marginal distributions if they can be written as functions of the components of a 2-dimensional Gaussian copula. Furthermore, the correlation in the copula has to be equal to the correlation between the normal variables. Note that the 2-dimensional case was only used as an illustration and the generalization to higher dimensions is straightforward.

Consider now the situation when the underlyings follow an SDE driven by a Brownian motion in such a way that

$$F_i = F_i\left(\int_0^T \sigma_i dW_i\right)$$

at the maturity T. Examples include the normal and lognormal process. Assume that the Brownian motion can be written as

$$W_i(t) = \sum_j \int_0^t a_{ij}(u) dZ_j(u)$$

where $\{Z_i\}$ are independent standard Brownian motions and the non-stochastic variables a_{ij} satisfy $\sum_j a_{ij}^2(t) = 1$. As

$$(dW_i(t))^2 = \sum_j a_{ij}^2 dt = dt$$

W_i is a standard Brownian motion. The covariation is given by

$$dW_i dW_j = \sum_k a_{ik} dZ_k \sum_m a_{jm} dZ_m = \sum_k a_{ik} a_{jk} dt =: \rho_{ij} dt$$

The correlation ρ_{ij} between the Brownian motions is restricted by

$$|\rho_{ij}| = |\sum_k a_{ik} a_{jk}| \leq \left(\sum_k a_{ik}^2\right)^{1/2} \left(\sum_k a_{jk}^2\right)^{1/2} = 1$$

where Cauchy-Schwarz inequality has been used. As

$$\sum_i \lambda_i \int_0^T \sigma_i dW_i = \sum_j \int_0^T \left(\sum_i \lambda_i \sigma_i a_{ij}\right) dZ_j$$

is normally distributed for arbitrary $\{\lambda_i\}$, $\{X_i\} = \{\int_0^T \sigma_i dW_i\}$ must be Gaussian, see Appendix. It is therefore possible to use the above result and patch together the marginal distributions $\{F_i\}$ using a Gaussian copula.

We now compute the correlation $\operatorname{corr}(X_i, X_j)$ that should be used in the copula. For completeness, we allow the X_is to be observed at different points in times, i.e.

$$X_i = \int_0^{T_i} \sigma_i dW_i$$

It is shown in the Appendix that

$$\operatorname{Var}(X_i) = \int_0^{T_i} \sigma_i^2 dt$$

and with similar ideas it can be proven that

$$\operatorname{Covar}(X_i, X_j) = \int_0^{\min(T_i, T_j)} \sigma_i \sigma_j \rho_{ij} dt$$

from which we obtain

$$\operatorname{Corr}(X_i, X_j) = \frac{\int_0^{\min(T_i, T_j)} \sigma_i \sigma_j \rho_{ij} dt}{\sqrt{\int_0^{T_i} \sigma_i^2 dt} \sqrt{\int_0^{T_j} \sigma_j^2 dt}}$$

To avoid confusion, we refer to this as the statistical or *terminal correlation* while ρ_{ij} is called the *instantaneous* or *local correlation*. Similarly, $(\frac{1}{T}\int_0^T \sigma_i^2(u)du)^{1/2}$ is called the *terminal volatility* and σ_i the local volatility. If the X_is are all reset at the same time, i.e. $T_i = T, \forall i$, and the local volatilities and correlations are time independent, the statistical correlation is equal to the local correlation. In general, however, the correlations are not equal and it is important to remember to use the statistical correlation in the Gaussian copula.

When using copulas for modeling high-dimensional derivatives, we obtain the probability density function p after which the price can be computed through the integral

$$\int g(x_1, x_2, \ldots, x_n)p(x_1, x_2, \ldots, x_n)dx_1, dx_2, \ldots, dx_n$$

Although theoretically attractive, this approach is in general only possible for low dimensions, typically $n = 2$ and $n = 3$. For higher dimensions, the implementation is low performing because of the difficulty in evaluating a high-dimensional integral numerically. Sometimes the method is too slow even in 2 or 3 dimensions, depending on the product type and the demands of the users.

10.2 Variable Freezing

We illustrate this method by considering payoffs of the form

$$g(F_1, F_2, \ldots, F_n) = (h(F_1, F_2, \ldots, F_n) - K)_+$$

Examples include *basket options*: $h(F_1, F_2, \ldots, F_n) = \sum_i w_i F_i$, and *index options*: $h(F_1, F_2, \ldots, F_n) = \prod_i F_i^{w_i}$. For a motivation of the approach, consider a basket option when the underlyings follow a normal SDE,

$$dF_i = \sigma_i dW_i$$

The basket price also satisfies a normal SDE,

$$d\sum_i w_i F_i = \sum_i w_i \sigma_i dW_i = \sigma dW, \quad \sigma = \sqrt{\sum_{ij} w_i w_j \sigma_i \sigma_j \rho_{ij}}$$

and can therefore be priced with standard techniques, see Sect. 5.3. Similarly, if the underlyings follow a lognormal SDE,

$$dF_i = \sigma_i F_i dW_i$$

then the index satisfies a lognormal SDE with drift:

$$
\begin{aligned}
d\prod\nolimits_i F_i^{w_i} &= \left(\prod_i F_i^{w_i}\right)\left(\sum_i \frac{dF_i^{w_i}}{F_i^{w_i}} + \frac{1}{2}\sum_{ij}\frac{dF_i^{w_i}}{F_i^{w_i}}\frac{dF_j^{w_j}}{F_j^{w_j}}\right)\\
&= \left(\prod_i F_i^{w_i}\right)\left(\sum_i w_i\sigma_i dW_i + \frac{1}{2}\sum_i w_i(w_i-1)\sigma_i^2 dt\right.\\
&\qquad \left. +\frac{1}{2}\sum_{ij} w_i w_j \sigma_i\sigma_j\rho_{ij}\, dt\right)\\
&= \left(\prod_i F_i^{w_i}\right)\left(\sigma dW + \frac{1}{2}\sum_i w_i(w_i-1)\sigma_i^2 dt + \frac{1}{2}\sigma^2 dt\right),\\
\sigma &= \sqrt{\sum\nolimits_{ij} w_i w_j \sigma_i\sigma_j\rho_{ij}}
\end{aligned}
$$

and the pricing can again be done with standard techniques.

Consider now a general function $h(F_1, F_2, \ldots, F_n)$ and assume the underlyings to satisfy

$$dF_i = \sigma_i(F_i)dW_i$$

The SDE for h is then

$$dh(F_1, F_2, \ldots, F_n) = \sum_i (\partial_i h)\sigma_i dW_i + \frac{1}{2}\sum_{ij}(\partial_{ij} h)\sigma_i\sigma_j\rho_{ij}\, dt$$

Observe that in general there is no analytical solution to the SDE as $(\partial_i h)\sigma_i$ and $(\partial_{ij} h)\sigma_i\sigma_j$ are functions of $F_1, F_2, \ldots, F_n$. However, if we believe that h approximately behaves as a normal variable, we might try to replace these functions with the results obtained by evaluating them at today's values of the underlyings. Thus, by freezing $(\partial_i h)\sigma_i$ and $(\partial_{ij} h)\sigma_i\sigma_j$ at today's value, we obtain a normal SDE for h that can be solved analytically. Similarly, if we believe h to be close to lognormal, we write $\sigma_i\partial_i h = \sigma_i h\partial_i h/h$ and freeze $\sigma_i(\partial_i h)/h$ at today's value, and similar for $\sigma_i\sigma_j\partial_{ij} h$.

The technique of freezing variables is most successful for short and possibly also intermediate maturities. Unfortunately, the tails of the resulting distributions are incorrectly modeled and that the outcome is dependent on the choice of freezing SDE. It is particularly popular to freeze the equation into the normal, the lognormal or the shifted lognormal SDE, because of the attractive analytical properties of these processes.

10.3 Moment Matching

Just as in the previous section, we price high-dimensional derivatives by considering the function $h(F_1, F_2, \ldots, F_n)$ as an underlying itself. We use the fact that it is possible to compute the lowest-order moments of h from the marginal distributions and some knowledge about the joint distribution, e.g. the correlations. We can then construct an analytically solvable distribution with identical moments and use it as an approximation of h in the pricing.

For example, assume that the distributions F_i can be written as functions of standard normal variables X_i with correlations ρ_{ij}. By expressing $h(F_1, F_2, \ldots, F_n)$ in terms of $\{X_i\}$, the two lowest moments $E[h^1]$ and $E[h^2]$ can be computed and h can be approximated with a distribution with two degrees of freedom, for instance, the normal or lognormal distribution. If it is possible to compute the three lowest moments, h can be approximated by a distribution with three degrees of freedom, for example, the shifted lognormal distribution.

The matching of moments gives in general a better approximation than the method of freezing the coefficients. This approach unfortunately also gives an incorrect tail distribution. Furthermore, h is only known at a fixed point T in time, meaning that nothing is known about the time dependence of the process. It means that path-dependent derivatives cannot be priced.

10.4 Quadratic Functional Modeling

We here limit ourselves to underlyings F_i that depend on standard normal variables X_i. The function h can be viewed as depending directly on $\{X_i\}$ instead of indirectly via the underlyings. For a more compact notation, we use matrices and let X be an $n \times 1$ matrix with components $\{X_i\}$. The method that we introduce works when h is a quadratic function

$$h(X) = a + BX + \frac{1}{2} X^T C X$$

where B is a $1 \times n$ matrix and C is a symmetric $n \times n$ matrix. As a quadratic function appears naturally in a second-order Taylor expansion, the method is useful when the dependence on X is weak, for example, for products with short maturity. Furthermore, as the quadratic function contains three free parameters, it can be used as an approximate distribution obtained by moment matching arbitrary distributions up to the third order.

To obtain independent variables, we write $X = MX'$, where M is an $n \times n$ matrix and $\{X_i'\}_{i=1}^n$ are independent $\mathcal{N}(0, 1)$ distributed variables. It implies that

$$h(X) = a + B'X' + \frac{1}{2} X'^T C' X'$$

where $B' = BM$ and $C' = M^T CM$. As C' is symmetric, there exists an $n \times n$ matrix O that is orthogonal, $O^T O = 1$ and $OO^T = 1$, and diagonalizes C': $O^T C' O = \text{diag}(\{c_j\})$. Introducing the $\mathcal{N}(0, 1)$ independent variables $\{Y_j\}_{j=1}^n$ by $X' = OY$ and the matrix $B'' = B'O$ gives

$$h(X) = a + B''Y + \frac{1}{2}Y^T\mathrm{diag}(\{c_j\})Y = a + \sum_{j=1}^{n}\left(b_jY_j + \frac{1}{2}c_jY_j^2\right)$$

Assume that the variables have been labeled so that $c_j \neq 0$ for $j \leq \nu$ and $c_j = 0$ otherwise:

$$h(X) = a + \sum_{j=\nu+1}^{n} b_jY_j + \sum_{j=1}^{\nu}\frac{1}{2}c_j\left(Y_j + \frac{b_j}{c_j}\right)^2 + \sum_{j=1}^{\nu}\frac{b_j^2}{2c_j}$$

The second term is a sum of independent normal variables with zero mean and is therefore itself a normal variable with zero mean. We can therefore write

$$h(X) = \alpha + \delta Y_0 + \sum_{j=1}^{\nu}\frac{1}{2}\gamma_j\left(Y_j + \beta_j\right)^2$$

To price derivatives depending on $h(X)$, we now derive the probability density function. As $h(X)$ is a sum of independent variables, the characteristic function is particularly simple to compute:

$$\hat{p}(k) = E\left[e^{ikh(X)}\right] = e^{ik\alpha}E\left[e^{ik\delta Y_0}\right]\prod_{j=1}^{\nu}E\left[e^{\frac{1}{2}ik\gamma_j(Y_j+\beta_j)^2}\right]$$

For Y a standard normal variable it holds that $E\left[e^{ikY}\right] = e^{-k^2/2}$ and it follows from Sect. 5.9 that

$$\begin{aligned} E\left[e^{ik(Y+\beta)^2}\right] &= \frac{1}{\sqrt{1-2ik}}e^{ik\beta^2/(1-2ik)} \\ &= (1+4k^2)^{-1/4}e^{i\theta/2}\exp\left(ik\beta^2(1+2ik)/(1+4k^2)\right) \end{aligned}$$

where θ is defined from $1-2ik = \sqrt{1+4k^2}e^{-i\theta}$, which means that $\theta = \arctan 2k$. Thus, the characteristic function has the form

$$\begin{aligned} \hat{p}(k) = e^{ik\alpha}e^{-\delta^2k^2/2}\prod_{j=1}^{\nu}&(1+\gamma_j^2k^2)^{-1/4}e^{i\theta_j/2} \\ &\exp\left(\frac{1}{2}i\gamma_jk\beta_j^2(1+i\gamma_jk)/(1+\gamma_j^2k^2)\right) \end{aligned}$$

We now follow the computations in the beginning of Sect. 7.4 and express the Heaviside function as a complex integral to obtain

$$
\begin{aligned}
P(h(X) > x) &= \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}\left(\hat{p}(k)\frac{e^{-ikx}}{ik}\right) dk \\
&= \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Im}\left(\hat{p}(k)\frac{e^{-ikx}}{k}\right) dk \\
&= \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{1}{k}e^{-\delta^2k^2/2}\left(\prod_{j=1}^{\nu}(1+\gamma_j^2k^2)^{-1/4}\exp\left(-\frac{1}{2}\gamma_j^2k^2\beta_j^2/(1+\gamma_j^2k^2)\right)\right) \\
&\quad \mathrm{Im}\left(\exp\left(ik\alpha - ikx + \frac{i}{2}\sum_{j=1}^{\nu}\theta_j + \frac{1}{2}i\sum_{j=1}^{\nu}\gamma_j k\beta_j^2/(1+\gamma_j^2k^2)\right)\right) dk \\
&= \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{1}{k}\left(\prod_{j=1}^{\nu}(1+\gamma_j^2k^2)^{-1/4}\right) \\
&\quad \exp\left(-\frac{1}{2}k^2\left(\delta^2 + \sum_{j=1}^{\nu}\gamma_j^2\beta_j^2/(1+\gamma_j^2k^2)\right)\right) \\
&\quad \sin\left(k(\alpha - x) + \frac{1}{2}\sum_{j=1}^{\nu}\arctan(\gamma_j k) + \frac{1}{2}k\sum_{j=1}^{\nu}\gamma_j\beta_j^2/(1+\gamma_j^2k^2)\right) dk \\
\Rightarrow p_{h(X)}(x) &= \frac{1}{\pi}\int_0^\infty \left(\prod_{j=1}^{\nu}(1+\gamma_j^2k^2)^{-1/4}\right) \\
&\quad \exp\left(-\frac{1}{2}k^2\left(\delta^2 + \sum_{j=1}^{\nu}\gamma_j^2\beta_j^2/(1+\gamma_j^2k^2)\right)\right) \\
&\quad \cos\left(k(\alpha - x) + \frac{1}{2}\sum_{j=1}^{\nu}\arctan(\gamma_j k) + \frac{1}{2}k\sum_{j=1}^{\nu}\gamma_j\beta_j^2/(1+\gamma_j^2k^2)\right) dk
\end{aligned}
$$

which concludes our derivation of the probability density function.

The existence of a closed-form expression has made the quadratic functional approach popular for approximating distributions. Furthermore, in the 1-dimensional case, it is even possible to derive a Black–Scholes type of expression for European options. Indeed, consider

$$
h(X) = a + bX + \frac{1}{2}cX^2 = a - \frac{1}{2}\frac{b^2}{c} + \frac{1}{2}c\left(X + \frac{b}{c}\right)^2 = \alpha + \frac{1}{2}\gamma\,(X+\beta)^2
$$

where we have assumed that $c \neq 0$ to avoid reduction to the simple first-order Taylor expansion. The PDF can be computed by using

$$P(h(X) < x) = P\left((X+\beta)^2 < y^2\right) = N(y-\beta) - N(-y-\beta),$$
$$y = \sqrt{2(x-\alpha)/\gamma}$$

if $x > \alpha$ and 0 otherwise. It follows that

$$p_{h(X)}(x) = \frac{d}{dx} P(h(X) < x) = (n(y-\beta) + n(y+\beta))\frac{dy}{dx}$$

The call option price can be computed by

$$E[(h(X)-K)_+] = \int_{\max(K,\alpha)}^{\infty} (x-K)(n(y-\beta)+n(y+\beta))\frac{dy}{dx}dx$$

$$= \int_{K'}^{\infty} (\gamma y^2/2 + \alpha - K)(n(y-\beta)+n(y+\beta))dy$$

where $K' = \sqrt{2(K-\alpha)_+/\gamma}$. We focus on the term containing $y-\beta$ as the term with $y+\beta$ can be obtained by substituting β with $-\beta$:

$$\int_{K'}^{\infty} (\gamma y^2/2 + \alpha - K)n(y-\beta)dy$$
$$= \int_{K'-\beta}^{\infty} (\gamma y^2/2 + \gamma y\beta + \alpha - K + \gamma\beta^2/2)n(y)dy$$
$$= \int_{-\infty}^{-K'+\beta} \left(-\gamma/2\frac{d}{dy}(yn(y)) + \beta\gamma\frac{d}{dy}n(y) + (\gamma/2 + \alpha - K + \gamma\beta^2/2)n(y)\right)dy$$
$$= \gamma/2(\beta + K')n(\beta - K') + (\gamma/2 + \alpha - K + \gamma\beta^2/2)N(\beta - K')$$

The end result is

$$E[(h(X)-K)_+]$$
$$= \gamma/2(\beta + K')n(\beta - K') + (\gamma/2 + \alpha - K + \gamma\beta^2/2)N(\beta - K')$$
$$+ \gamma/2(-\beta + K')n(-\beta - K') + (\gamma/2 + \alpha - K + \gamma\beta^2/2)N(-\beta - K')$$

10.5 Change of Measure

The method that we discuss here reduces the dimensionality by one unit. It is therefore not particularly useful when the dimension is high. For low dimensions, on the other hand, the method can lead to a substantial performance improvement. For instance, the evaluation of an integral in 1 or 2 dimensions is much faster than if there were an additional dimension.

We illustrate the method by considering a product that pays the positive part of the difference between two equities S_1 and S_2:

$$V = P_{0T} E\left[(S_1 - S_2)_+\right]$$

where the expectation is with respect to the forward measure. A straightforward computation involves the evaluation of a 2-dimensional integral. An alternative approach is to change numeraire to one of the equities, for example, S_2:

$$V = S_2 E^* \left[\left(\frac{S_1}{S_2} - 1\right)_+\right]$$

where the expectation E^* is with respect to the measure with S_2 as numeraire. The pricing can be done by interpreting S_1/S_2 as the single underlying of the problem. This underlying is clearly driftless as it is the quotient of a tradable and the numeraire. Furthermore, it is trivial to compute the volatility as the diffusion parts of S_1 and S_2 remain unchanged under a change of measure.

10.6 Digital Options

We now focus on payoffs that have a digital dependence on at least one of the underlyings. It is then often possible to reduce the dimensionality of the pricing. Consider, for example, an option with payoff $\theta(F_1 - F_2 + K)$. Assuming that both F_1 and F_2 can be written as functions h_1 and h_2 of standard normal variables X_1 and $X_2 = \rho X_1 + \sqrt{1-\rho^2} X_1^{\perp}$, where h_2 is monotonically increasing, we obtain

$$\begin{aligned}
E[\theta(F_1 - F_2 + K)] &= P(F_1 > F_2 - K) = P(h_1(X_1) > h_2(X_2) - K) \\
&= P\left(\rho X_1 + \sqrt{1-\rho^2} X_1^{\perp} < h_2^{-1}\left(h_1(X_1) + K\right)\right) \\
&= P\left(X_1^{\perp} < \frac{h_2^{-1}\left(h_1(X_1) + K\right) - \rho X_1}{\sqrt{1-\rho^2}}\right) \\
&= E\left[N\left(\frac{h_2^{-1}\left(h_1(X) + K\right) - \rho X}{\sqrt{1-\rho^2}}\right)\right]
\end{aligned}$$

As there exist efficient implementations of the cumulative normal function $N(\cdot)$, the integrand can be considered as analytic and we have thereby reduced the problem to a 1-dimensional integral. It is straightforward to extend this approach to other types of payoffs, for instance, $h(F_1)\theta(F_1 - F_2 - K)$, or to higher dimensions.

The above technique usually only reduces the dimension by 1 and is therefore uninteresting for high dimensions. As a complement, we now go through a method that works particularly well in this situation. To illustrate the approach, we price a digital call option on the maximum of weighted underlyings, i.e. the payoff takes the form

$$\begin{aligned}\theta\left(\max_{1\le i\le n} w_i F_i - K\right) &= 1 - \theta(K - w_1F_1)\theta(K - w_2F_2)...\theta(K - w_nF_n)\\ &= 1 - \theta(K_1 - F_1)\theta(K_2 - F_2)...\theta(K_n - F_n)\end{aligned}$$

where $K_i = K/w_i$. We consider underlyings that can be written as monotonically increasing functions of standard normal variables $\{X_i\}$. The main assumption of this technique is that the variables can be written as

$$X_i = \lambda_i Y + \sqrt{1-\lambda_i^2}Y_i$$

where $\{Y, \{Y_i\}\}$ are independent standard normal variables. We obtain the equivalent payoff

$$1 - \theta\left(\frac{h_1^{-1}(K_1) - \lambda_1 Y}{\sqrt{1-\lambda_1^2}} - Y_1\right)\theta\left(\frac{h_2^{-1}(K_2) - \lambda_2 Y}{\sqrt{1-\lambda_2^2}} - Y_2\right)...$$
$$\theta\left(\frac{h_n^{-1}(K_n) - \lambda_n Y}{\sqrt{1-\lambda_n^2}} - Y_n\right)$$

Taking the expectation gives a 1-dimensional integral

$$1 - \int p(y)N\left(\frac{h_1^{-1}(K_1) - \lambda_1 y}{\sqrt{1-\lambda_1^2}}\right)N\left(\frac{h_2^{-1}(K_2) - \lambda_2 y}{\sqrt{1-\lambda_2^2}}\right)...$$
$$N\left(\frac{h_n^{-1}(K_n) - \lambda_n y}{\sqrt{1-\lambda_n^2}}\right)dy$$

The problem of evaluating an n-dimensional integral has thereby been reduced to an integral over a single dimension.

The above result can be used to price ordinary calls on the maximum of a weighted set of underlyings. For example

$$\left(\max_{1\le i\le n} w_i F_i - K\right)_+ = \int_K^\infty \theta\left(\max_{1\le i\le n} w_i F_i - K'\right) dK'$$

can be computed through a 2-dimensional integral. Observe that a 2-dimensional maximum option with zero strike is preferally computed by

$$\max(F_1, F_2) = F_2 + (F_1 - F_2)_+$$

and changing the numeraire as in Sect. 10.5.

One of the limitations with the above technique is that the correlation between the X_is is restricted to the form

$$\operatorname{corr}\left(X_i, X_j\right) = \lambda_i \lambda_j + \left(1 - \lambda_i^2\right)\delta_{ij}$$

If this correlation structure is considered to be too restrictive, the definition of $\{X_i\}$ can be generalized to

$$X_i = \lambda_i Y + \lambda_i' Y' + \sqrt{1 - \lambda_i^2 - {\lambda_i'}^2}\, Y_i$$

This leads to integrals of one dimension higher. By the same token, more variables $Y^{(n)}$ can be introduced until the user obtains the appropriate balance between dimensionality and flexibility in the correlation structure.

The approach is limited in the number of supported payoff types. Fortunately, the special cases to which it can be applied are often useful for derivatives pricing. For instance, the method can be used to compute the cumulative function $P(F_1 < x_1, F_2 < x_2, \ldots, F_n < x_n)$.

10.7 Spread Options

We here discuss spread options that pay $(F_1 - F_2 - K)_+$ at time T. Because of the strike K, it is not possible to use the change of numeraire technique of Sect. 10.5. We instead do the pricing by computing

$$\int (F_1 - F_2 - K)_+ p(F_1, F_2) dF_1 dF_2$$

A 2-dimensional integral can be too slow to compute in many practical applications. We therefore use $p(f_1, f_2) = \frac{d^2}{df_1 df_2} P(F_1 < f_1, F_2 < f_2)$ to obtain

$$\begin{aligned}
&\int (f_1 - f_2 - K)_+ \frac{d^2}{df_1 df_2} P(F_1 < f_1, F_2 < f_2) df_1 df_2 \\
&\quad = \int \delta(f_1 - f_2 - K) P(F_1 < f_1, F_2 < f_2) df_1 df_2 \\
&\quad = \int P(F_1 < f + K, F_2 < f) df
\end{aligned}$$

As it is often possible to find an approximate expression for the cumulative distribution P, e.g. when the distribution is Gaussian, only a 1-dimensional integral remains.

By combining the inequalities

$$(F_1 - K_1)_+ - (F_2 - K_2)_+ \le (F_1 - F_2 - (K_1 - K_2))_+ \le (F_1 - K_1)_+ + (K_2 - F_2)_+$$

with static replication techniques, upper and lower bounds on spread options are given in terms of ordinary call options. Although the bounds do not appear to be too tight, violations of the lower bound occurred in 2009 for interest rate products with F_1 and F_2 EUR CMSs with 10Y and 2Y tenors starting in 5Y and with a maturity of 20Y, see McCloud (2011).

10.8 Correlations

The correlation is often the parameter to which the price of a higher-dimensional payoff is most sensible. Unfortunately, the number of liquid correlation-dependent products in the market is limited. It means that the correlation often has to be estimated rather then calibrated. A consequence is that the correlation dependence then cannot be hedged. It means that the sellers of higher-dimensional derivatives are relatively highly exposed to market risk and therefore charge a high margin.

We would like to issue a warning on the currency effect on volatilities and correlations. Most financial practitioners are aware that the volatility depends on the currency for which the underlyings are valued. For instance, the dollar value of gold has a different volatility from the euro value of gold. It is not, however, as well known that the correlation is a statistical measure that also depends on the currency in which the measurement is made. This fact can easily be verified by inspecting the defining formulae for the local and terminal correlation. Alternatively, use any information services such as Bloomberg or Reuters to compare, for example, the correlation between EURSEK and USDSEK with the correlation between EURCHF and USDCHF, which reflects that the correlation between euro and dollar depends on whether the measurement is done in Swedish kronor or Swiss francs. The dependency on the currency is not marginal but can lead to relatively large differences in correlation. It means that the European branch of a bank cannot use USD-based correlations computed by a US-based research department for their EUR-based models.

Another misconception is the belief that a high correlation between assets means that they are interchangeable. This is far from true as can be seen from the fact that the difference $F_1 - F_2$ between two assets with equal normal volatilities σ and 99% correlation has a normal volatility of

$$\sqrt{\sigma^2 + \sigma^2 - 2\rho\sigma\sigma} = \sigma\sqrt{2}\sqrt{1-\rho} \approx \frac{1}{7}\sigma$$

which is not as significant reduction of the volatility as one naively would expect.

10.9 Calibration

Because of the lack of liquid market data, the correlation (or any other relevant copula parameter) can often not be calibrated. One of the few exceptions for which liquid calibration instruments can be found is in the pricing of equity basket options. We first discuss the question whether there are enough calibration instruments to uniquely determine the correlation, or more generally the joint PDF $p(\{F_i\})$. We then give a concrete example of how a calibration can be performed.

If, for a basket $F = \sum_i w_i F_i$ and a given maturity T, option prices are given for all weights $\{w_i\}$ and strikes K, it is intuitively clear that the joint PDF $p(\{F_i\})$ is known. The statement can be proven mathematically by observing that when differentiating the forward option price

$$U = \int \left(\sum_i w_i F_i - K \right)_+ p(\{F_i\}) d\{F_i\}$$

twice with respect to the strike, the *Radon transform* of the PDF is obtained:

$$\frac{d^2}{dK^2} U = \int \delta \left(\sum_i w_i F_i - K \right) p(\{F_i\}) d\{F_i\} =: \mathcal{R}[p](\{w_i\}, K)$$

The inverse Radon transform then gives p as a function of $\frac{d^2}{dK^2}U$, see Carr and Laurence (2011) and references therein.

In general, liquid equity basket option prices only exist for indices, which means that the weights $w_1, w_2, \ldots, w_n$ are fixed for a given set of underlyings $\{F_i\}$. The weights can vary with time, but for simplicity we assume them to be time independent. The consequence is that there are not enough calibration instruments in the market for a unique determination of the joint distribution. A common way to obtain a calibration problem with a unique solution is to assume a certain parametric function for the correlation. In Sect. 12.4 we give concrete examples of such parameterizations when modeling commodity futures.

Based on the simplicity of calibrating local volatility models, we attempt to extend that approach to include correlations. Models obtained in this way are called *local correlation models*.

We assume that there exist liquid options for all maturities and strikes on the basket $F = \sum_i w_i F_i$ and on the underlyings $\{F_i\}$. We assume that it is possible to simultaneously calibrate local volatility models for both the basket and the underlyings. It means that there exist functions $\sigma(t, F)$ and $\{\sigma(t, F_i)\}$ that correctly price the calibration instruments and are related by

$$\sigma(t, F) F dW = dF = \sum_i w_i dF_i = \sum_i w_i \sigma(t, F_i) F_i dW_i$$

Squaring both sides gives

$$\sigma(t, F)^2 F^2 = \sum_{ij} w_i w_j \sigma(t, F_i) \sigma(t, F_j) F_i F_j \rho_{ij}$$

We conclude that simultaneous local volatility models for the basket and the underlyings can be constructed if ρ_{ij} satisfies the above equation. In Langnau (2010) it was shown how (non-unique) solutions can be found by constructing continuous 1-dimensional parameterizations of the correlation such that the right-hand side is increasing and has values both above and below the value of the left-hand side. The solution can then be found via a 1-dimensional root finder.

Bibliography

Carr P, Laurence P (2011) Multi-asset stochastic local variance contracts. Math Finance 21:21-52
Langnau A (2010) A dynamic model for correlation. Risk April:74–78
McCloud P (2011) The CMS triangle arbitrage. Risk January:126–131

Part IV
Asset Class Specific Modeling

Chapter 11
Equities

This chapter deals with derivatives that depend on the value of one or several equity stocks or equity indices. We argue that equity stocks are similar to the idealized underlying S that has been considered in previous chapters and it is therefore possible to use the models discussed there. There are, however, certain aspects of the equity stock behavior that do not agree with the idealized underlying. One example is that a company can default, which results in a worthless stock. The theory treating such credit defaults is a quantitative area of its own. Some of the methods used in that field are similar to the ones in this book, e.g. hazard-rate models for defaults corresponds to short-rate models for interest rates, but most often they are fundamentally different. Derivatives that depend on credit defaults often rely on copula models and on models for extreme events. This makes credit default modeling closely linked to auctorial mathematics and we have therefore chosen not to include this area in the set of the major asset classes discussed in this book.

Equity derivatives can be priced accurately without using models for defaults of the underlying stock. The reason is that the credit information is accounted for by the market through the value of the underlying and its implied volatility. It is, however, necessary to be aware of the way the credit exposure affects these values, e.g through occasional downward jumps in the underlying process and a skew in the implied volatility.

Another example of how equity stocks behave differently from an idealized underlying is that they pay discrete dividend cash flows throughout their life. The implication is that the assumptions in Chap. 1 are violated and that the models discussed so far in the book need to be modified before they can be applied to equities.

Equities are different from the other major asset classes as they are almost exclusively used for investment and speculation. Commodities, interest rates and foreign exchange, on the other hand, are often traded for hedging purposes as well as for investment and speculation.

A particularly interesting equity derivative is the *warrant*. This instrument type is similar to a call option with a few exceptions. For instance, it usually has a

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_11,

longer maturity than ordinary equity options. More important, from a modeling perspective, is the fact that new stocks are issued when a warrant is exercised. This dilutes the value of the stock and leads to a lower stock price if several warrants are exercised simultaneously. It is straightforward to modify the option pricing formula to account for this diluting effect.

We first discuss characteristic behavior of equity prices and volatilities. This guides us in the choice of an appropriate model. The effects of dividend payments, and how to modify derivatives models to account for them, are also covered.

11.1 Stylized Facts

The price of an equity stock is driven by news relating to the issuing company and the competitors, and by global news affecting the stock market as a whole. The news flow typically results in more extreme events in the stock returns than predicted by Gaussian models such as the normal or lognormal process. This is represented by fat tails in the probability distributions. Furthermore, the center of the distribution has a higher peak than a Gaussian model. This behavior is not unique to equities but occurs within all asset classes.

Unexpected news gives jumps in the stock price. Equities are different from the other asset classes because large jumps are more often negative than positive. This behavior reflects the investor fear of negative news and the associated credit exposure. It manifests itself in a fatter left tail and a skew for the implied volatility.

For most asset classes, the distribution looks less lognormal the smaller the time scale of the distribution. This is particularly pronounced for equities and commodities where jumps are frequent. For equities, the volatility skew for short maturities is sometimes so strong that if described by a shifted lognormal model, an extrapolation beyond the normal process is necessary.

Consider the situation when a company reports unexpectedly low profits. The consequence is a sell-off and a lower stock price. This scenario makes investors uncertain and implies higher trading volumes and a higher volatility. This is one of the main reasons for the negative correlation between the price of an equity stock and the volatility. The outcome is a skewed implied volatility surface.

11.2 Dividends

An equity stock pays discrete *dividend* cash flows throughout its life. The payment dates and the size of the dividends are not known unless they are in the near future. We later discuss how known dividend payments can be accounted for in derivatives pricing. For now, we assume that all future dividend payments are unknown both in size and in the payment date. The simplest way to model this situation is to assume an equal probability of receiving a dividend proportional to the stock value at any given date. The benefit of the possibility of receiving dividend payments in a time interval $[t, t + dt)$ can then be written as $qSdt$ where the variable q is called the *dividend yield*.

We start the discussion of derivatives pricing on dividend-paying equities by considering forward contracts. The pricing can be done by repeating the static replication argument of Sect. 2.1 in the presence of a dividend yield. For this purpose, we compare the strategy of holding $e^{-q(T-t)}$ underlyings with the holding of a forward contract with strike K and a cash amount KP_{tT}. As both strategies are worth S_T at the maturity T, the fair value of the strike, i.e. the forward, is $F_t = e^{-q(T-t)} P_{tT}^{-1} S_t$, or $F_t = e^{(r-q)(T-t)} S_t$ with r the continuous compounded interest rate for the period $[t, T]$.

If dividends are accounted for, the purchase of an equity stock at $t = 0$ gives a holding of $S_t e^{qt}$ at t. The fundamental theorem of asset pricing states that the forward $F_t = e^{-q(T-t)} P_{tT}^{-1} S_t$ is a martingale in the T-forward measure as it is the quotient of an investment strategy and P_{tT}. As the lognormal price of a call option is

$$V_t = P_{tT} \left(FN\left(d_+\right) - KN\left(d_-\right) \right)$$

the above expression for the forward gives us the following generalization of the Black–Scholes formula:

$$V_t = e^{-q(T-t)} S_t N\left(d_+\right) - P_{tT} KN\left(d_-\right),$$

$$d_\pm = \frac{\ln(e^{-q(T-t)} S_t/(P_{tT} K))}{\sigma\sqrt{T-t}} \pm \frac{1}{2}\sigma\sqrt{T-t}$$

It is straightforward to extend the formula to a time-dependent dividend yield by replacing qt with $\int_0^t q_u du$.

With B_t being the money market account, $S_t e^{qt}/B_t$ is a martingale in the risk-neutral measure. For simplicity, we assume it to be a driftless lognormal process

$$d\left(S_t e^{qt}/B_t\right) = \sigma \left(S_t e^{qt}/B_t\right) dW_t$$

Combining this equation with the product rule of differentiation

$$d\left(S_t e^{qt}/B_t\right) = (q - r)\left(S_t e^{qt}/B_t\right) dt + \left(e^{qt}/B_t\right) dS_t$$

gives

$$dS_t = (r - q) S_t dt + \sigma S_t dW_t$$

which generalizes the option pricing SDE in earlier chapters. This SDE can be used to numerically price exotic equity options, for example, by simulations.

By representing the unknown future dividend payments by a dividend yield, we have seen that it is straightforward to generalize the derivative models we have used for idealized underlyings. Observe that many of the results in the previous chapters are affected by such a generalization. For instance, we leave it as an exercise to the reader to derive the corresponding asymptotic limits and no-arbitrage conditions of Sect. 2.4. Another consequence concerns American call options for which the lower bound $(S - Ke^{qT}P_{0T})_+$ invalidates the argument in Sect. 2.5 that it is never optimal to exercise an American call option early. There are now situations when an early exercise is preferable in order to obtain the dividend payments, which are otherwise missed out on by the option holder. The consequence is that American call options for equities can be worth more than their European counterparts, see Sect. 9.3 for more details.

We now turn to the analysis of known future dividend payments and start by assuming a single dividend payment D that occurs at t_D. To avoid arbitrage, the equity stock must drop by the same amount: $S_{t_{D+}} = S_{t_{D-}} - D$. A European option with maturity $T > t_D$ can therefore equivalently be viewed as an option on a fictional non-dividend paying underlying

$$\tilde{S} = \begin{cases} S - DP_{tt_D} & t < t_D \\ S & t \geq t_D \end{cases}$$

This new underlying has the same time T value as S and as

$$\tilde{S}_{t_{D-}} = S_{t_{D-}} - D = S_{t_{D+}} = \tilde{S}_{t_{D+}}$$

it is also continuous. We conclude that if the option is priced at $t < t_D$ with the Black–Scholes formula, S should be replaced with $S - DP_{tt_D}$.

It is not sufficient to change the spot value from S to $\tilde{S}$, but the volatility needs to be modified as well. To understand why this is the case, assume for the sake of the argument that interest rates are zero and that the option matures just after the dividend payment. It means that $\tilde{S} = S - D$ for essentially the whole life of the option. We then obtain

$$d\tilde{S}_t = dS_t = \sigma S_t dW_t = \sigma \frac{S_t}{S_t - D} \tilde{S}_t dW_t$$

Freezing to today's spot prices gives an estimate for the volatility: $\tilde{\sigma} = \sigma \frac{S_0}{S_0 - D}$.

When it comes to American call options, there are no dividend payments in (t_{D+}, T) and they should therefore not be exercised early in this interval. We conclude that they have the same price at t_{D+} as their European counterparts. There are, however, situations when it is optimal to exercise just before the dividend payment date t_D. Indeed, an American call option not exercised at t_D has the value

$$V_{\text{American}}(t_{D+}, S_{t_{D+}}) = V_{\text{European}}(t_{D+}, S_{t_{D+}}) = V_{\text{European}}(t_{D+}, S_{t_{D-}} - D)$$

t_{D+}. If, on the other hand, an American call option has been exercised early, a payment $S_{t_{D-}} - K$ would have been obtained. This gives us the breakpoint B for the early exercise boundary, defined by

$$B - K = V_{\text{European}}(t_D, B - D)$$

If exercising at t_{D-}, we lose the time t_D value of $K - KP_{t_D T}$ in interest rate cost compared to an exercise at maturity. As the dividend payment must compensate for this loss if early exercise should be optimal, the above equation only has a solution if

$$D \geq K(1 - P_{t_D T})$$

For the same reason as a non-dividend paying American call option should not be exercised early, the American call option in our example can only be optimally exercised early at t_{D-}, and not before. Thus, if the underlying pays dividends on a discrete set of dates $t_{D_1}, t_{D_2}, \ldots, t_{D_N}$, the option can only be optimally exercised just before any of the payment dates. The American option therefore coincides with the corresponding Bermudan option and can be priced with the same methods.

We now turn our attention to American put options, where we again start with a single dividend payment D at time t_D. We conclude from Sect. 9.3 that the exercise boundary between t_D and T is an increasing function reaching the strike level K at maturity. When it comes to the exercise decision for $t' < t_D$, the gain in interest rate $KP^{-1}_{t't_D} - K$ needs to be compared with the gain from the drop in the underlying value by D at t_D. Assuming constant interest rates: $P^{-1}_{t't_D} = e^{r(t_D - t')}$, gives the break-even time

$$t' = t_D - \frac{1}{r} \ln\left(1 + \frac{D}{K}\right)$$

We see that it is never optimal to exercise an American put in the interval $[t', t_D)$, regardless of the value of the underlying. Continuing backwards in time, the exercise boundary increases from the zero value at t' to a local maximum at some $t'' < t'$ after which it will continue its more normal behavior and decrease with the time to maturity. It is straightforward to generalize this argument to obtain a qualitative understanding of the early exercise when there are several dividend payments. It is important to be aware of the implications of the simplifying assumption that the dividend for certain pays D at t_D. For instance, if the underlying is close to the exercise boundary just before t', the value is very low and it is unlikely that the whole dividend amount is paid.

11.3 More Advanced Models

Although good to a lowest-order approximation, the lognormal model does not account for the stylized fact in Sect. 11.1. For instance, a proper model should account for the correlation between the underlying and the volatility by including a stochastic volatility. This introduces a skew in the implied volatility which is found to be weaker than the market skew, in particular for short maturities. Because of their ability to produce a pronounced skew for short maturities, Lévy processes are popular tools for equity derivatives models. The use of such models is also in agreement with the observational fact that the prices of equity stocks do jump, in particular downwards.

11.4 Volatilities and Correlations

We now explain in more detail how exotic equity options can be priced. To make the discussion as simple as possible, we use the lognormal process of Sect. 11.2. As the drift was given by a no-arbitrage condition, it remains to determine the value of the volatility. This is usually done by calibrating to European options.

We illustrate the procedure by calibrating a lognormal model to European ATM options with maturities T_1, $T_2, \ldots, T_N$ and assume a piece-wise constant local volatility: $\sigma_t = \sigma_i$ for $t \in (T_{i-1}, T_i]$, where T_0 is today's date. The following chain of equations is then obtained:

$$\begin{aligned}
\sigma_1^2 T_1 &= \sigma_{1,\text{imp}}^2 T_1 \\
\sigma_1^2 T_1 + \sigma_2^2 (T_2 - T_1) &= \sigma_{2,\text{imp}}^2 T_2 \\
\ldots &
\end{aligned}$$

from which the local volatility can be computed. This method can, of course, be extended to other interpolation methods for the local volatility, though the result is more complicated formulae. When generalizing to non-lognormal processes, it is often necessary to use approximate expressions for the implied volatility. Another technique for obtaining the local volatility is to assume a parametric form and rely on a least squares fit. For a more general discussion of calibration, see Sect. 4.2.

In this and the following chapters we limit the discussion to calibration to ATM options. The calibration to skew and smile is more complicated and can, for example, be done by using the perturbative methods in Sects. 6.1 and 7.2.

Recall that European option pricing is based on the forward being a martingale. Being the quotient of the spot price and a zero-coupon bond, the forward volatility has components coming from both the spot volatility and interest rate volatility. To obtain the spot volatility from the calibration, it is necessary to strip out the interest rate volatility. Fortunately, if the maturity is not too long, the interest rate contribution is small. The interest rate effect for longer maturities is discussed in Sect. 13.21.

Regarding the correlation, the situation is similar to that for idealized underlyings. We therefore refer the reader to Sects. 10.8 and 10.9. It is, however, important to be aware of the fact that correlations tend to increase in times of a crisis, when equity prices are falling.

Chapter 12
Commodities

This chapter is dedicated to the pricing and risk management of financial derivatives for which the underlying is a commodity. The asset class of commodities is vast and can roughly be subdivided into energy, agriculture, base metals and precious metals, see Fig. 12.1 for some examples of commodities in each class. Each commodity has its own characteristics that need to be accounted for in the modeling. Nevertheless, our goal is to set up a general framework that can guide the reader in the modeling of derivatives that depend on any type of commodity.

From a derivatives modeling perspective, the most important differences between commodities and purely financial asset classes are the facts that commodities are constantly produced and consumed, and that for every commodity there is a storage cost, i.e. a cost for physically holding the commodity. The immediate consequence is that the forward price argument of Sect. 2.1 breaks down. Indeed, instead of storing a commodity up to the maturity of the forward contract in a static replication argument, it is often cheaper to purchase a newly produced commodity at maturity. The conclusion is that there is no such strong connection between the spot price and the forward prices as is found in purely financial markets such as the equity and the FX markets. The modeling of commodities therefore involves the evolution of the forward curve, which is similar to yield curve modeling for interest rates.

The cost of storage depends on the commodity type, with precious metals found in one extreme. They are cheap to store and the amount consumed and produced every day is low compared to the existing stock. It means that the forward-replication argument is approximately valid. Precious metals can therefore with a high degree of accuracy be described by using only the spot value and in many pricing systems they are not modeled as other commodities but rather as FX rates. We do therefore not discuss this commodity class in detail but rather focus on the other extreme where we find electricity, which is very expensive to store. In this instance the link between the spot and the forward prices is weak. Because of its extreme properties, electricity is arguably the most complex commodity from a modeling perspective. It is no exaggeration to claim that mastering the art of electricity derivatives modeling makes it possible to model any other commodity derivative. For this reason we often use electricity as an illustrating example.

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_12,

Energy	Petroleum Products Crude oil, heating oil	Other Fossil Fuels Coal, natural gas	Electricity, uranium
Agricultural	Grains & Oilseeds Wheat, soybeans	Livestock & Meat Live cattle, pork bellies	Softs Coffee, sugar
Base metals	Aluminium, copper		
Precious metals	Gold, silver		

Fig. 12.1 Classification of the major commodities

As commodities are only weakly correlated to interest rates, the forward and futures prices are approximately equal. We therefore use the words forward and futures interchangeably in this chapter.

We start the chapter by classifying the various commodities traders and discuss their incentives. We then review the main characteristics of commodity prices. Based on these observations, we describe how commodity derivatives can be modeled. Finally, we discuss in detail how volatilities and correlations enter the models in practical applications.

12.1 Commodities Trading and Investment

With the exception of precious metals, it is in general not profitable to invest in commodities physically because of the high storage costs. Instead, commodities are mainly traded through futures contracts (swaps are used for oil distillates, e.g. heating oil, gasoline and jet fuel). The counterparties agree, via an exchange, on a future delivery of the underlying commodity in return for a cash payment.

Let us consider the example of a crude oil futures contract with a delivery in March next year. This contract is traded, and settled daily, up to some date in February, the *last trade date*, that has been defined by the exchange. The oil then needs to be delivered some time between two exchange-specific dates in March, the *first delivery date* and the *last delivery date*. The futures contract also states the place of the delivery, the quality of the oil, how the cash deposit should be made, etc.

The buyers of March oil futures contracts are not prepared to pay much more than what they believe the oil price will be in March. By symmetry, the sell side are not prepared to sell for much less than what they believe the oil price will be in March. The intersection of these supply and demand forces determines the futures prices. It follows that the March futures price of oil is close to what oil analysts expect the price to be at that time.

It is in general not possible to make a profit from the information that commodities prices are expected to increase. Assume, for instance, that the media report that oil prices are expected to increase for the next couple of years and that you decide to take advantage of this by purchasing oil. As the high storage cost prevents a physical holding, you turn to the futures market. But as the futures price reflects the expected price of oil in the future, the increase reported by the media has already been included in the futures price, meaning that no profit can be made.

The only way to make a profit from increasing prices in the commodities market is if the commodity price increases more than what is predicted by the futures prices. However, as the prices of futures are mainly determined by the consensus on the expected future spot price, it appears that the return on commodity investments should be zero on average. Despite this, investments in commodity futures have historically been as attractive as investments in equity stocks. As we now show, the good returns can partly be explained by a supply and demand argument by Keynes (1930) that originates in the fact that not all participants in the commodities market have purely financial interests.

The producer of a commodity usually knows a certain time in advance (typically a couple of months) how much of the commodity is going to be produced. The price of the commodity can then be secured by entering long-dated futures contracts right away instead of selling at a later date when the commodity has be produced, and thereby taking the risk of price fluctuations. The purchaser of a commodity, on the other hand, typically has customers interested in buying at the spot price. To minimize the price risk between the buy side and sell side, the purchaser buys short-dated futures.

The supply and demand forces described above affect most commodities. The result is that the discounted price is, on the average, higher at the short end than at the long end of the futures curve. A profit can therefore be made by purchasing long-dated futures and selling short-dated futures. In practice, this can be done by holding futures contracts until they have short maturity, whereafter they are sold and new long-dated futures are purchased. This strategy, which can be perpetual, is referred to as *rolling* futures contracts. As the futures contracts are never held beyond the last trade date, the strategy is ideal for financial investors which have no interest in delivering, or taking delivery of, the underlying commodity.

The expected returns in the commodities market do not come from expected increases in commodities prices, but rather from the premium that certain market participants are prepared to pay to secure the price. It means that a profit can be made in times of expected decreasing spot prices as well as in times of expected increasing spot prices. Thus, a purely financial market participant can take advantage of the supply and demand forces and obtain a yield in return for taking on the exposure to price fluctuations. An investment in commodities is therefore similar to any other financial investment: a certain positive return is expected for a given risk.

The arguments of Keynes' work in the reverse direction for certain commodities. For instance, as refineries buy crude oil at the short end of the futures curve they would also like to sell their end products, the oil distillates, at the short end to minimize their financial risk. The refineries are then only exposed to the price of the *crack spreads*, which are the price differences between the distillates and crude oil, and avoid any temporal risk. Some purchasers of oil distillates have the foresight to hedge themselves longer out on the futures curve. A financial investor aiming to pick up a risk premium in such a market must therefore sell long-dated futures and buy them back the short end, i.e to roll a short commodities position.

We have so far described the trading activity of industrial market participants and how this opens up for the possibility of obtaining a positive return. Let us now focus on the purely financial investor. Apart from the attractive returns, there are a couple of other reasons for investing in the commodities markets.

The global financial market is currently in an inflationary period that has lasted since the beginning of the twentieth century, see Fischer (1996) and references within. It means that the price of tradable goods increases with time, at least when viewed over long time periods of a decade or more. As commodities (which we in this book define as exchange-traded goods) are tradable goods, their values appreciate on the average. It means that holding a commodity with a low storage cost (e.g. gold) is in the long run at least as good an investment as holding cash. Furthermore, it has been observed that rolling strategies of commodity futures have higher correlation to inflation than equity and interest rate investments. Commodity investments are therefore popular with investors such as pension funds that seek inflation-linked investments. It should be noted, however, that the correlation between commodities investments and inflation is much lower than the common belief, especially for shorter time scales (of a decade or less) where it can be indistinguishable from zero.

Investments in commodities have historically given high returns at different times in the business cycle than equities and bonds. An investment portfolio containing commodities therefore allows access to a high-performing market at times when neither equities nor bonds are performing. Empirical studies show that commodities investments have low correlation to other asset classes such as equities and interest rates. Furthermore, investments in the various commodity classes are themselves only weakly correlated. It means that adding a component of commodities to the traditional equity and bond investment portfolio leads to a diversification that can result in a higher return for a given risk.

The above reasons have led to a large number of financial market participants entering the commodities markets. In fact, the influx has been so large that they outnumber the industrial market participants. It is therefore no longer certain whether the traditional Keynes' supply and demand argument is valid. Even though financial investors take both sides of the trades, i.e. they can go both long and short the commodity, they tend to be predominately long. It means that the rolling strategies used by the financial investors can very well be such that their supply and demand forces are stronger than those coming from the industry. The result is that the returns originating from supply and demand could be negative when rolling long futures positions. This applies in particular to rolling strategies in the short end of the futures curve as most financial investors have chosen to do the rolling there because their clients expect an investment that follows the spot price as closely as possible and because the liquidity dries out quickly when moving away from the very short end. Despite this effect, long investors have anyway been rewarded by high returns recently, and the reason depends on the commodity type. For instance, gold prices have been driven by the flight to security during the financial crisis starting in 2007 while the prices for industrial metals have been driven by the booming Asian economies.

12.2 Commodity Price Characteristics

With commodites being traded goods, their prices are driven by supply and demand. For instance, if there is an abundance of a certain commodity, the price tends to be low. Because of low profitability, the producers tend to lower the production in such times. Consumption then erases the abundance with a return to more normal prices. The reverse situation occurs in times of scarcity. The consequence is a mean-reverting behavior for commodity prices. A purely financial market such as the equity market can only display a weak mean reversion as it is otherwise possible to take advantage of this behavior by buying at low prices and selling at high prices. For commodities, on the other hand, the success of such strategies is limited by storage costs, which means that a relatively strong mean reversion can exist. As a low storage cost has a dampening effect on the fluctuations around the mean-reversion level, the mean-reverting behavior is more pronounced when the storage costs are high.

After this short introduction to the behavior of the spot price, we proceed to discuss the dynamics of the forward (futures) curve. We start by investigating relationships between the forward and the spot. Recall that a forward contract on an idealized underlying, i.e. without credit exposure, dividends and storage costs, can be replicated by purchasing the underlying at spot and holding it until maturity. The outcome is that the discounted forward value equals the spot. As we now see, this equality is replaced with an inequality for commodities.

Consider the strategy $V_{\text{phys.}}$ of storing a commodity up to a certain future time T. Starting today with one unit of the commodity S, we end up with commodities worth $e^{-cT} S$ at T, where we have assumed a constant proportional storage cost $c \geq 0$ per unit time. We compare with the strategy V_{roll}^{0} of holding e^{-cT} forward contracts f and $S_{t=0}$ worth of bonds maturing at T. Assuming constant interest rates, this strategy is worth

$$e^{-cT} f + e^{rT} S_0 = e^{-cT}(f + F) + \xi = e^{-cT} S + \xi, \quad \xi = e^{rT} S_0 - e^{-cT} F$$

at T, where F as usual denotes the futures price. This strategy has the same initial value as $V_{\text{phys.}}$ while the terminal value differs by a term ξ that is already known at today's date. If ξ is negative, $V_{\text{phys.}}$ is preferable to V_{roll}^{0}, which means that it is possible to arbitrage the forward market by physically holding the commodity. We conclude that $\xi \geq 0$ which gives an upper bound on commodities forward prices:

$$F \leq e^{(r+c)T} S_0$$

Recall from Chap. 1 that it was necessary to exclude the strategy of holding cash from our theoretical framework because of the existence of a better performing strategy, the money market account. We now use a similar argument to exclude the physical holding of a commodity.

First of all, we would like to point out that the above argument of V_{roll}^{0} being better than $V_{\text{phys.}}$ at T is not sufficient for this purpose. The reason is that someone that holds a commodity physically probably does so to be able to take advantage of any surges in the price permitted by the mean-reverting behavior. It means that we instead should consider the strategy $V'_{\text{phys.}}$ in which the investor is allowed to sell the commodity at the time of choice and not only at T. As the holder of V_{roll}^{0} is not able to take advantage of any surges in the spot price, it is not true that V_{roll}^{0} is better than $V'_{\text{phys.}}$. For this purpose, we now modify V_{roll}^{0} to a strategy that dominates $V'_{\text{phys.}}$.

Consider the strategy V_{roll}^{1} consisting of $e^{-cT/2}$ forward contracts maturing at $T/2$ and S_0 worth of bonds. It follows from the above that this investment is worth $e^{-cT/2}S + \xi$ at $T/2$, for ξ positive. If then investing $e^{-cT/2}S_{T/2} + \xi$ in bonds and entering e^{-cT} forward contracts, we obtain $e^{-cT}S + \xi'$ at T, for ξ' positive. We have thereby constructed a strategy that is better than $V_{\text{phys.}}$ if observed at T as well as at $T/2$. By introducing more and more intermediate dates, we obtain a strategy V_{roll}^{∞} that performs better than $V_{\text{phys.}}$ if viewed at any date in $[0, T]$. As we can exit the strategy V_{roll}^{∞} at the time of choice, V_{roll}^{∞} performs better than $V'_{\text{phys.}}$. To avoid arbitrage it is therefore necessary to exclude the strategy of physically holding a commodity.

Observe that in the strategies V_{roll}^{k} we enter futures contracts and when they are close to maturing, we sell them and purchase new futures contracts. This is just the rolling strategy described earlier. The limit $k \to \infty$ means that the rolling is done with contracts of infinitesimal maturity. This enables us to follow the behavior of the spot without being subjected to storage costs. Thus, infinitesimal rolling (which excludes the strategy of physically holding of a commodity) is similar to the money market account (which excludes the strategy of holding cash). Observe that as infinitesimal-maturing futures do not exist in reality, it is anyway common to hold commodities physically to benefit from the timing option of being able to sell the commodity at the time of choice. The infinitesimal rolling strategy is also limited in practice by the bid-offer spread.

The difference in value between $V'_{\text{phys.}}$ and $V_{\text{phys.}}$ represents the benefit of being able to sell the commodity at the time of choice before T. If expressing this value as $e^{yT}S_0$, then y is called the *convenience yield*. We would like to draw the reader's attention to the fact that there exist several different definitions of convenience yield in the literature. For instance, it is common to define it through the relation $F = S_0 \exp((r + y)T)$ between the forward and the spot. When this definition is used, the convenience yield no longer represents the benefit of a timing option.

We have argued that the link between futures prices and the spot price is vague when the storage costs are high. Market participants instead tend to agree on a futures price that is close to the expected value of the spot at the maturity. If there is more interest from sellers than from buyers at a certain maturity, a supply-demand argument leads to a futures price below the expected spot price, and vice versa. The higher the variance of the spot distribution at maturity, the more the sellers (or buyers) are prepared to pay to lock in their price risk, which results in a bigger difference between the futures price and the expected spot price. Furthermore, we have also seen that the storage cost, interest rates and the convenience yield have an impact on the shape of the futures curve. The conclusion is that the futures curve has a complex behavior that depends on several factors.

A curve for which the futures prices decrease with maturity is said to be in *backwardation*. The reverse situation, seen in the equities market (but also for commodities), is referred to as *contango* (Fig, 12.2).

As the futures prices are often mainly driven by the expected value of the spot price, we take a closer look at the factors that affect the spot price. As it is difficult to make general statements regarding the consumers on the demand side, we focus on the supply side. We observe that the prices of commodities are heavily influenced by the inventories: the larger the inventories, the larger the supply of the commodity, which results in low prices. Thus, we conclude that the spot price is inversely correlated to the level of inventories. Furthermore, the expected future levels of the inventories have a significant effect on the shape of the futures curve. This can lead to many different shapes of the curve, for instance, it can increase for the first few months after which it starts to decrease. There exist several other factors that influence the futures curve, such as expected weather conditions and political decisions.

The price behavior of renewable and non-renewable commodities can be fundamentally different. From an economic viewpoint, a commodity is considered to be non-renewable if it is expected to last some decades or perhaps up to a century. It means that we consider oil and natural gas as non-renewable but not coal as it is expected to last for at least another thousand years (based on current production and consumption levels). The price of non-renewable commodities is influenced by news regarding the estimated reserves. The reserves can also be classified into different price layers. To illustrate this fact, assume that the price of oil increases and stays at high levels for several years. Oil companies can then make a profit by extracting oil at places that were not economically viable before, e.g. deep water drilling. When the new drilling platforms come in use, the supply of oil increases, which leads to a decrease in the price in order to be linked to the marginal costs of the extraction.

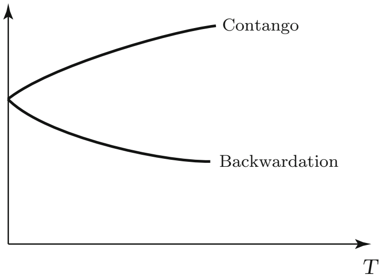

Fig. 12.2 Commodity futures prices can increase (contango) or decrease (backwardation) with maturity

Many commodities display a seasonal pattern in their prices. This behavior can originate in the supply side as well as in the demand side. An example of seasonality on the demand side is found in countries in the Far North where more electricity is needed during the winter than during the summer. On the other hand, if these countries rely on hydropower, there can be an abundance of electricity during the spring due to snow melting. This is an example of seasonality originating in the supply side. For a commodity like electricity, which is economically unviable to transport over vast distances, the seasonal patterns depend on the geographic location. For instance, many warm locations have higher electricity prices during the summer (because of air conditioning) than in the winter.

The seasonality is not restricted to the yearly seasons. Consider again the example of electricity where the demand is high during the day but low in the night when most people are asleep. The result is higher prices during the day. Apart from the daily profile, there is also a weekly seasonal pattern as the energy consumption is different on weekdays versus weekends. Furthermore, the daily profile looks different on holidays, weekends and working days.

As the futures prices reflect the expected future spot price, an increase in the spot price implies an increase in the futures prices, though with a smaller amplitude because of the mean-reverting behavior. It means that a backwardated curve has an even more pronounced backwardation after an increase in spot. Similarly, a decreasing spot price gives less backwardation. The same type of argument applies to a curve in contango. The conclusion is that the slope of the futures curve is positively correlated to the inventories and both these quantities are negatively correlated to the spot. Another implication is that the volatility of futures contracts decreases with the maturity.

Recall that the (lognormal) volatility for an equity stock generally goes up when the stock value decreases. This is because a decrease in the stock value is often related to a market uncertainty regarding the underlying company, which in turn implies a higher volatility. The situation can be the reverse for commodities. If a commodity price goes up, this could, for example, be because of unexpectedly low inventories. As the inventories act as a buffer on the price towards unexpected news flows, the volatility increases in this situation. The spot price and the volatility are therefore sometimes positively correlated for commodities. This is one of the explanations why the implied volatility skew for commodities can be found in the opposite direction compared to equity options. It also follows that a commodity with a seasonal price dependence has a seasonal volatility as well.

We conclude that the spot price, the futures prices, the inverse slope and the volatility are often positively correlated for commodities. These variables are in turn negatively correlated to the inventory levels. Several empirical studies, for example, by Deaton and Laroque (1992), Fama and French (1987), Fama and French (1988) and Ng and Pirrong (1994), have been done that support these conclusions. We would like to point out, however, that these relations are violated relatively often and can therefore not be taken as a rule.

The cost of storage makes commodities highly volatile. The general rule is that the higher the storage cost, the higher the volatility. The reason is that commodities with low storage costs have high inventories, which acts as a buffer on any sudden changes in the supply or the demand. The most extreme example of high storage costs is electricity with limiting storage techniques, such as pumping up water into reservoirs during night time. The consequence is a spot volatility that can exceed a hundred percent. This should be compared with equity stocks that seldom have volatilities above 50%. The large volatility (and volatility of volatility) observed for many commodities can make it impractical to use perturbation schemes like those in Chaps. 6 and 7. The volatility for the futures contracts, however, is often substantially smaller than the spot volatility.

For commodities with high storage costs and low inventories, a sudden disruption in the supply or the demand causes the price to jump. Again, the most extreme case is electricity for which the spot price can jump by several hundred percent. It can result in jumps of the futures prices, but of more modest amplitudes. This explains the popularity of using jump processes for the modeling of electricity and other commodities with high storage costs. The jumps are often followed by a correction/jump of roughly the same size but in the opposite direction. The reason is that the disruption that caused the initial jump was temporary and when corrected the spot price returns to more normal levels. This behavior of the jumps is typical of commodities and does not characterize financial assets because of arbitrage opportunities. Thus, commodities require different types of jump models than, for example, equities.

12.3 Commodities Derivatives Modeling

We base our commodities modeling on the evolution of the futures prices as these are the liquidly traded contracts for commodities. The modeling becomes particularly simple as the futures have zero drift in the risk-neutral measure according to the result of Sect. 3.10.

We start by pricing a European option on a futures price. This product only depends on one point on the futures curve and a modeling of the full curve is not necessary. As the futures price is a martingale in the risk-neutral measure, we can pick a driftless SDE of choice for the modeling. Let us for simplicity use a lognormal SDE

$$dF_t = \sigma_t F_t dW_t$$

The price of a European call option is then given by

$$V_t = E\left[(F_T - K)_+ / e^{\int_t^T r_u du}\right] = e^{-\int_t^T r_u du}\left(F_t N(d_+) - K N(d_-)\right)$$

where we have used the computations in Sect. 5.2 together with the assumption that interest rates are deterministic.

When it comes to exotic options pricing, models of the full futures curve can be required. Models of this type are similar to yield curve models. For example, it is popular to base commodity models on the evaluation of the spot, see, for example, the models by Gibson and Schwartz (1990) and Gabillon (1992), which correspond to short-rate models for interest rates. We instead follow an approach similar to Andersen (2008) and choose to work with models based directly on the futures curve, in analogy to the market model approach for interest rates. We show below that this approach gives a relatively simple model to work with that at the same time can reflect the stylized facts of the previous section.

We choose to work in the risk-neutral measure where futures prices are martingales. For concreteness, we assume lognormal dynamics:

$$dF_{tT} = \sigma_{tT} F_{tT} dW_t$$

where F_{tT} denotes the time t value of the futures price with maturity T. We assume a zero correlation to interest rates, making the calibration to European option prices straightforward. Indeed, the change to the T-forward measure does not alter the SDE, which implies a trivial relation between the local volatility and the implied volatility. The pricing of exotics can then be done by solving the above SDE (actually chain of SDEs as we have one for each T) either analytically or numerically.

The futures prices are totally correlated in the SDE we have chosen. A straightforward decorrelating generalization can be obtained by allowing each futures price F_{tT} to have its own Brownian driver $W_t(T)$. This gives an infinite set of Brownian motions as T is a continuous variable. For a practical implementation, however, a finite set will suffice. This can be achieved by setting

$$W(T) = \sum_i a_i(T) Z_i$$

where $\{Z_i\}$ are independent Brownian motions and $\sum_i a_i(T)^2 = 1$ in order for $W(T)$ to be a standard Brownian motion. The price to pay for this flexibility in the correlation structure is a lower performance of the implementation as values of several Brownian motions need to be computed and stored.

Even when there is only a single driver, the implementation can have a higher-dimensional character. To understand how this can be the case, consider the situation when futures prices F_{tT_1}, F_{tT_2} and F_{tT_3} of three maturities are simulated over the time steps $\{t_i\}$. Assume that the simulation has been made up to t_i and that we are about to evolve the SDE to t_{i+1}. We use

$$F_{t_{i+1}T_j} = F_{t_i T_j} \exp\left(-\frac{1}{2} \int_{t_i}^{t_{i+1}} \sigma_{sT_j}^2 \, ds + \int_{t_i}^{t_{i+1}} \sigma_{sT_j} \, dW_s \right)$$

where $F_{t_i T_j}$ is assumed to be known. The problem is that for each path of the simulation, the vector $(F_{t_i T_1}, F_{t_i T_2}, F_{t_i T_3})$ is needed for the computation of $(F_{t_{i+1}T_1}, F_{t_{i+1}T_2}, F_{t_{i+1}T_3})$. The storage and access of such a vector can be a performance bottleneck in an implementation of a Monte Carlo simulator if many futures prices $\{F_{tT_j}\}_j$ need to be evolved. Even worse, a tree (or PDE) solver has the same dimension as the number of futures prices. As the performance of tree solvers is heavily dependent on the dimension, they are practically impossible to use on these types of models.

The dimensionality problem can be solved by allowing the futures prices $\{F_{tT}\}_T$ to depend on one (or a couple) common factors. For instance, with a separable volatility $\sigma_{tT} = \Lambda_T \sigma_t$ we obtain

$$F_{t_{i+1}T} = F_{t_i T} \exp\left(-\frac{1}{2}\Lambda_T^2 \int_{t_i}^{t_i+1} \sigma_s^2 ds + \Lambda_T \int_{t_i}^{t_i+1} \sigma_s dW_s\right)$$

We see that if $\{F_{t_i T}\}_T$ are known, $\{F_{t_{i+1}T}\}_T$ can be computed by only using the stochastic information $\int_{t_i}^{t_i+1} \sigma_s dW_s$. This process is independent of T and all futures prices can therefore be obtained by simulating $\int_0^{t_i} \sigma_s dW_s$ over the time steps $\{t_i\}$. This is equivalent to simulating a standard Brownian motion over the scaled time steps $t_i' = \int_0^{t_i} \sigma_s^2 ds$. As only a single random process needs to be stored, an efficient tree implementation can be done. The approach can be generalized by setting $\sigma_{tT} = \sum_{k=1}^{N} \Lambda_T^k \sigma_t^k$. This method gives a greater flexibility in the choice of volatility while the performance decreases with increasing N.

Recall that the volatility typically decreases with the time to maturity of the futures contract. This can be accounted for by including a factor of the form $e^{k(t-T)}$ in the volatility. A possible choice of separable volatility function is then $\sigma_{tT} = \eta_t e^{k(t-T)} = e^{-kT} \eta_t e^{kt}$.

The penalty to pay for using a separable volatility $\sigma_{tT} = \Lambda_T \sigma_t$ is that the futures curve becomes highly correlated. For instance, with similar computations as those in Sect. 10.1, we obtain an expression for the terminal correlation:

$$\mathrm{Corr}(\ln F_{tT}, \ln F_{t'T'}) = \sqrt{\int_0^{\min(t,t')} \sigma_s^2 ds \Big/ \int_0^{\max(t,t')} \sigma_s^2 ds}$$

We see that σ_t can be used to control the correlation. However, as pointed out in Sect. 4.2, constant parameter values should guide us in what can be done and what should be avoided in a model. In our example, a constant σ_t gives no control over the correlation. Furthermore, σ_t already has the responsibility of controlling the volatility of the futures prices. Therefore, to use this parameter for controlling the correlation can be going one step too far, and even if we did this, the futures prices $\{F_{tT}\}_T$ would still be perfectly correlated. This type of model should therefore not be used to price products that have a strong dependence on the correlation between different futures prices.

The optimal solution that allows for non-trivial correlation appears to be a combination of separable volatility and multiple drivers. For instance, a driver of the form $\sigma_{tT} dW_t = \Lambda_T^{(1)} \sigma_t^{(1)} dZ_t^{(1)} + \Lambda_T^{(2)} \sigma_t^{(2)} dZ_t^{(2)}$ allows for a low dimensional implementation at the same time as it has a non-trivial correlation structure.

As we show in Chap. 13, the extension to skew and smile models is complicated for interest rate models because the SDEs get modified under transformations between the pricing measure and the calibration measures. For commodities, on the other hand, we do not have this problem when assuming zero correlation to interest rates as the SDEs look the same in the risk-neutral measure and the forward measures. It is therefore straightforward to extend the model to include skew and smile, either through local volatility, or by stochastic volatility techniques. In the same manner, jumps can be incorporated, which is particularly important for commodities with high storage costs.

As pointed out earlier in the chapter, a commodity futures curve can have a peculiar appearance governed by supply and demand, expected future inventory levels, seasonality, etc. These factors can be hard to take into account in models based on the evaluation of the spot price. Furthermore, such models are typically based on variables defined from fixed time periods beginning today. For example, for the spot price itself this time period is zero as the spot is defined at today's date. When doing a simulation of such models, the variables are rolling along with the simulation date keeping a constant distance from it. This makes it hard to model effects that are defined for fixed points in time and not for fixed time periods from today, e.g. seasonality. From this perspective it is more natural to use models based on futures prices F_{tT} that are defined for fixed time points T. It is then trivial to calibrate to today's value of the futures curve as it is given by the initial state of the SDE.

As far as modeling is concerned, it can be useful to divide commodities into three distinct classes. The first one consists of commodities with low storage costs such as gold. The dynamics of the future curve can then to a high degree of accuracy be described by using only the spot S_t. We have chosen not to cover this type of commodities as the modeling can be done with foreign-exchange techniques. The second class comprises commodities that can be described by their futures prices F_{tT}, as discussed above. Example of commodities in this class are oil and copper. The third class consists of commodities that are not delivered at a certain point in time but rather during a fixed time period, such as electricity. The modeling can then be based on the futures price $F_{tTT'}$ representing the price per time unit for delivering the commodity during the time period $[T, T')$. Observe that the number of time parameters increases with the numbering of the commodity classes. From a modeling perspective, the increasing number of time parameters makes the first class the simplest while the third is the most difficult.

Regarding the modeling of commodities with a delivery period, observe that $(T' - T)F_{tTT'}$ must obey the *cocycle relation* (we borrow this expression from the mathematical branch of homology)

$$(T' - T)F_{tTT'} + (T'' - T')F_{tT'T''} = (T'' - T)F_{tTT''}$$

This relation restricts the set of possible evaluations for $F_{tTT'}$. For example, if $F_{tTT'}$ and $F_{tT'T''}$ satisfy lognormal SDEs, $F_{tTT''}$ cannot be lognormal. The normal SDE, on the other hand, is closed under the cocycle relation. Unfortunately, commodities with delivery periods generally have high storage costs, which leads to high volatilities. The existence of negative values of the underlying can therefore cause problems when using a normal SDE. Despite the fact that the cocycle relation needs to be fulfilled from a theoretical perspective, it is often possible to violate it in practice without too severe consequences. It is, however necessary to be aware of the pitfalls of doing so.

An alternative approach to construct a consistent model that satisfies the cocycle relation is to base the theory on prices for infinitesimal delivery periods, i.e. $F_{tTT'}$ in the limit when $T' \to T$. The finite delivery futures prices $F_{tTT'}$ can then be obtained from the infinitesimals through integration. The advantage of this approach is twofold: the modeling is reduced to two time variables instead of three and the cocycle relation is automatically satisfied. This technique is also used for interest rates and in Chap. 13 we show how it can be applied in practice.

We now clarify why it is possible to use a martingale (e.g. the lognormal SDE) to describe a market that is mean reverting. The reason is that we are referring to two fundamentally different market variables, the spot price and the futures prices. It is the former that is mean reverting while the latter is modeled by martingales. The consistency in this argument originates in the fact that the spot is a rolling (with respect to today's date) financial variable while the futures prices are stationary in time. For a mathematical explanation, assume for simplicity that the futures prices are normal with a separable volatility

$$
\begin{aligned}
dF_{tT} &= \Lambda_T \sigma_t dW_t \\
\Leftrightarrow F_{tT} &= F_{0T} + \Lambda_T \int_0^t \sigma_s dW_s
\end{aligned}
$$

The spot price $S_t = F_{tt}$ satisfies

$$
\begin{aligned}
dS_t &= S_{t+dt} - S_t = F_{t+dt,t+dt} - F_{tt} \\
&= F_{0,t+dt} + \Lambda_{t+dt} \int_0^{t+dt} \sigma_s dW_s - F_{0t} - \Lambda_t \int_0^t \sigma_s dW_s \\
&= \left(\partial_t F_{0t} + \frac{\partial_t \Lambda_t}{\Lambda_t} (F_{tt} - F_{0t}) \right) dt + \Lambda_t \sigma_t dW_t \\
&= -\frac{\partial_t \Lambda_t}{\Lambda_t} \left(\tilde{S}_t - S_t \right) dt + \Lambda_t \sigma_t dW_t, \quad \tilde{S}_t = F_{0t} - \frac{\Lambda_t}{\partial_t \Lambda_t} \partial_t F_{0t}
\end{aligned}
$$

which is a mean-reverting process. We discuss the relation between mean-reverting rolling variables and martingale stationary variables in more detail for interest rates in Chap. 13.

An industrial user is often exposed to the daily price of a commodity. For this reason, Asian options are very popular for commodities. Another reason for their popularity is that they help to avoid price manipulation in illiquid markets, which can occur for European options. Asian options can be priced by solving models like the one developed above either numerically or by using analytic approximations.

The physical user of commodities can also face *volume risk* on top of the price risk. For example, more electricity is consumed during harsh winters. As the demand is high during such times, so is the price. The user should therefore ideally hedge the volume risk as well as the price risk. In the electricity market, the most popular products for handling this risk are *swing options*. They typically work in the following way: for a predetermined number of days during a period, say a year, the holder of the option can consume electricity between a lower and upper bound for a fixed price. The option holder decides on which days to use this optionality. The total amount of electricity purchased for this price during the year must also be between a certain lower and upper bound.

12.4 Volatilities and Correlations

In the same way as for equities, the drift is given by a no-arbitrage condition while the volatility typically needs to be determined through calibration. The difference from equities is that each commodity futures contract has its own volatility and the model therefore contains a volatility surface: σ_{tT}, $t \leq T$ instead of just a volatility curve. The calibration can be done separately for each liquid maturity in the same way as it was done for equities. The value of the local volatility for an arbitrary T can then be found by interpolation.

There are often too few liquid calibration instruments for the above calibration procedure to succeed. For instance, liquid option quotes in general only exist when the option expiry equals (or is close to) the maturity of the underlying futures contract. An alternative calibration approach is to reduce the dimensionality of the problem by only considering certain shapes of the volatility surface. Such a restriction is also necessary when a tailor-made calibration is done to a subset of the possible liquid calibration instruments, see Sect. 4.3. Furthermore, we saw above that restricted forms of the volatility can enhance the performance. Examples of volatility surfaces include $\sigma_{tT} = \Lambda_T \sigma_t$, $\sigma_{tT} = \eta_t e^{k(t-T)}$ and $\sigma_{tT} = Ae^{k(t-T)} + B$, where the latter choice is based on the observation that commodities with high storage costs have a dependence on $T - t$ that is stronger than the dependence on t or T alone. The choice of functional form depends on the product to be priced as well as the set of liquid calibration products.

When it comes to commodities (and interest rates), it is necessary to distinguish between *intercorrelation*, i.e. correlation between different commodities, and *intracorrelation*, which is the correlation between futures contracts on the same commodity but with different maturity months. There exists some limited market information on intercorrelation. For instance, New York Mercantile Exchange (NYMEX) lists options on certain crack spreads. The intercorrelation clearly also depends on the maturity months for which the correlation is measured. This temporal dependence can to some extent be derived from the intracorrelation.

We here choose to focus on intracorrelation, for which there is only limited information available in the market. An example is given by the various calendar spreads that are traded at NYMEX. The intracorrelation modeling is rather complex as the correlation $\rho(t, T, T')$ depends on one more time variable than the volatility. In analogy with the above discussion for the volatility, we assume that the correlation only depends on the relative times to maturity $\rho = \rho(T - t, T' - t)$. We can then without loss of generality assume that $t = 0$. Because of the limited relevant market data, we choose to model intracorrelation by finding a suitable functional form.

As opposed to volatility modeling, there exist several constraints on the correlation, which makes it a complex problem to find a useful parametric form. The conditions are:

1. $\rho(T, T') = \rho(T', T)$
2. $\rho(T, T) = 1$
3. ρ should be positive semidefinite
4. $\rho(T, T') \geq \rho(T, T' + \Delta T)$
5. $\rho(T - \Delta T, T') \leq \rho(T, T')$
6. $\rho(T, T') \leq \rho(T + \Delta T, T' + \Delta T)$
7. $\theta \leq \rho(T, T')$

for $T \leq T'$. The first two conditions are obvious. For the third condition, assume, for example, that commodity futures satisfy $dF_T = \sigma_T dW_T$, where σ_T is allowed to depend on the futures price F_T. As a weighted sum of futures prices has a positive variance:

$$0 \leq \left(d \sum_i w_i F_i \right)^2 = \sum_{ij} w_i w_j \sigma_i \sigma_j \rho(T_i, T_j)$$

for an arbitrary vector $\{w_i\}$, the third condition must hold. It is natural to assume that the correlation between two futures should decrease with the distance between them, which explains conditions 4 and 5. Furthermore, we also believe that the correlation between two equidistant rates should increase with the distance from today's date. Indeed, the further away the rates are from today, the more indistinguishable they become. This is the content of condition 6. The upper bound of 1 for a correlation follows from the above conditions. For the lower bound, the theoretical limit is -1, corresponding to perfect anti-correlation. However, in practice we do not expect the correlation to become as low and we introduce a lower bound θ. For example, we seldom find futures prices that are negatively correlated, so $\theta = 0$ appears to be a natural constraint. Apart from the above conditions, it is also important that the correlation matrix should be in agreement with certain observational "facts", which we see examples of later.

We would like to point out that although conditions 4, 5 and 6 are often satisfied in the market, they are not written in stone. For example, the existence of a seasonal pattern can violate them. Despite this, we still believe that the conditions serve as a good base for finding a suitable functional form.

One of the simplest functional forms of the correlation is based on the assumption that there are only two driving factors Z and Z' of the futures curve. The drivers W_T of the futures prices F_T can be written as

$$W_T = a_T Z + b_T Z'$$

The constraint $a_T^2 + b_T^2 = 1$, necessary for $\{W_T\}$ to be standard Brownian motions, suggests the introduction of the variable ϕ_T defined by

$$\begin{cases} a_T = \cos \phi_T \\ b_T = \sin \phi_T \end{cases}$$

which implies that

$$\rho(T, T') = a_T a_{T'} + b_T b_{T'} = \cos(\phi_{T'} - \phi_T)$$

We have thereby simplified the parametrization of a 2-dimensional function $\rho(T, T')$ to the simpler problem of finding a parametrization of a 1-dimensional function ϕ_T.

As a general correlation matrix, the ρ constructed above satisfies conditions 1, 2, 3 and 7. We now choose ϕ_T so that the remaining conditions are satisfied. By a redefinition of the drivers Z and Z', we can assume that $\phi_0 = 0$ and $\phi_T \geq 0$. We also make the natural assumption that $0 \leq \phi_T \leq \pi$ for all T. Condition 4 then implies

$$\cos(\phi_T) \geq \cos(\phi_{T+\Delta T}) \Leftrightarrow \phi_T \leq \phi_{T+\Delta T}$$

i.e. the function $\phi = \phi(T)$ is increasing. Condition 6 implies

$$\cos(\phi_T - \phi_{T-\Delta T}) \leq \cos(\phi_{T+\Delta T} - \phi_T) \Leftrightarrow \phi_T - \phi_{T-\Delta T} \geq \phi_{T+\Delta T} - \phi_T$$

so $\phi = \phi(T)$ is concave. Note that condition 5 follows automatically when ϕ is an increasing function.

It remains to find an increasing and concave function limited from above. One example of such a function is

$$\phi_T = \alpha \left(1 - \exp\left(-\beta T^{\gamma}\right)\right)$$

which implies

$$\rho(T, T') = \cos\left(\alpha \left(\exp\left(-\beta T^{\gamma}\right) - \exp\left(-\beta (T')^{\gamma}\right)\right)\right)$$

For $T = 0$, it follows that $\phi_0 = 0$ and $a_0 = 1$, $b_0 = 0$, so Z is the driver W_0 for futures close to today's date. The second driver Z' becomes more pronounced

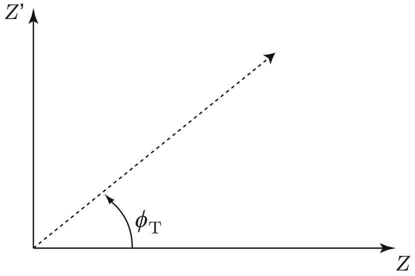

Fig. 12.3 The Brownian driver as an angle in the Z-Z' plane for a 2-factor model

with increasing T. The variable ϕ_T can be viewed as an angle between Z and Z', see Fig. 12.3.

We now consider an alternative parametrization that allows for more driving factors and decorrelation. Based on the ease through which the conditions on the correlation could be incorporated into a 2-factor model, we here also attempt to reduce the correlation matrix $\rho(T,T')$ to a function of a single variable ϕ_T.

Consider three Brownian drivers W_T, $W_{T'}$ and $W_{T''}$, $T < T' < T''$. We write

$$\begin{cases} W_{T'} &= \rho(T,T')W_T + \sqrt{1-\rho(T,T')^2}W_T^{\perp} \\ W_{T''} &= \rho(T',T'')W_{T'} + \sqrt{1-\rho(T',T'')^2}W_{T'}^{\perp} \end{cases}$$

where $W_T^{\perp}$ is independent of W_T and $W_{T'}^{\perp}$ is independent of $W_{T'}$. As the driver $W_{T'}^{\perp}$ is a part of $W_{T''}$ but does not affect $W_{T'}$, it seems reasonable to assume that it should not affect W_T since $T < T'$. Following Shoenmakers and Coffee (2000), we therefore assume that W_T and $W_{T'}^{\perp}$ are independent. We obtain the cocycle relation:

$$\rho(T,T'') = \rho(T,T')\rho(T',T'')$$

Limiting ourselves to positive correlations and setting $a_T = \rho(0,T)^{-1}$ gives that

$$\rho(T,T') = \rho(0,T')\rho(0,T)^{-1} = a_T/a_{T'}, \quad T \le T'$$

and $\rho(T,T') = \rho(T',T)$ for $T > T'$.

Again, conditions 1,2, 3 and 7 are automatically satisfied while the remaining conditions can be used to determine an appropriate functional form for a_T. Conditions 4 and 5 imply that a_T is an increasing function. With $\phi_T = \ln a_T$ condition 6 becomes

$$a_T/a_{T'} \le a_{T+\Delta T}/a_{T'+\Delta T} \Leftrightarrow \phi_T - \phi_{T'} \le \phi_{T+\Delta T} - \phi_{T'+\Delta T}$$
$$\Leftrightarrow \phi_T - \phi_{T+\Delta T} \le \phi_{T'} - \phi_{T'+\Delta T} \Leftrightarrow \phi_T - \phi_{T+\Delta T} \le \phi_{T+\Delta T} - \phi_{T+2\Delta T}$$

so ϕ must be concave. Without lack of generality, we can assume that $a_0 = 1$ which means that $\phi_0 = 0$. We therefore end up with the same demands on ϕ as for the

2-factor model and can choose it in the same way as was done there:

$$\rho(T, T') = a_T / a_{T'}, \quad T \leq T', \quad a_T = \exp\left(\alpha\left(1 - \exp\left(-\beta T^{\gamma}\right)\right)\right)$$

We would like to point out that although the cocycle relation might look intuitively appealing at first sight, its validity should be taken with a grain of salt. For instance, the 2-factor model does not satisfy this condition. Furthermore, the cocycle relation together with condition 6 implies that $\rho(T_0, T)$ is a convex function of T. This is in contrast with the 2-factor model for which the function is concave for T close to T_0 and only becomes convex for later maturities. We discuss this issue in more detail in Sect. 13.20 on interest rates where we also use empirical results to guide us in the choice of an appropriate model.

In the implementation of a commodities model, only a discrete set of futures prices is used. The correlation then takes the form of a matrix. In the functional form chosen above, the matrix has full rank. It means that there are as many Brownian drivers as there are futures prices. This is sometimes a bit too many drivers as it can cause performance problems in the implementation. For this reason, it is common to use an approximate matrix with a lower rank. The approximation is done with respect to some appropriate matrix norm. This technique is commonly referred to as *factor reduction*. It usually performs well, i.e. without modifying the original full-rank matrix too much, if the reduced rank is not too low. However, one needs to take care so that the correlation conditions are not violated too much (or at all). When the rank is low (2 or 3), it is obviously better to immediately use a low-rank method as the 2-factor parametrization.

The choice of a suitable set of Brownian drivers, or equivalently the choice of correlation matrix, is important. Realistic drivers means that the correct instruments are used for the hedging. The choice can also have an impact on the price. This can be understood by considering the pricing of an option on the difference between two futures contracts with different maturities. If using a single driving factor, the two futures prices move up and down together implying a low volatility for the spread. By introducing a second factor the futures prices become decorrelated which implies a higher spread volatility and an increase in the option value. In this situation, it is sufficient to use two factors. Assume now that we use a single model for pricing several spread options which have underlying futures of different maturities. The reason for using a single model is to be able to hedge consistently within the model, see Chap. 4. It is then no longer sufficient to use two factors. Instead, it is necessary to find an appropriate representation of the evaluation of the futures curve in terms of the drivers in order to hedge and price correctly.

The correlation often needs to be bumped in order to investigate the market risk and sometimes, when appropriate correlation-dependent products are available, to find the hedge. Bumping intracorrelations is more difficult than bumping volatilities as the correlation constraints should preferably be preserved. For intercorrelations the problem of the bumping comes from the constraint of a positive semidefinite matrix. The bumping obstacles can be avoided by assuming a suitable parametrization of the correlation matrix as we have done above for the intracorrelation.

Bibliography

Andersen L (2008) Markov models for commodity futures: theory and practice. Social Science Research Network. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1138782. Accessed 16 May 2011

Deaton A, Laroque G (1992) On the behaviour of commodity prices. Rev Econ Stud 59:1–23

Fama EF, French KR (1987) Commodity futures prices: some evidence on forecast power, premiums and the theory of storage. J Bus 60:55–73

Fama EF, French KR (1988) Business cycles and the behavior of metal prices. J Finance 43(5):1075–1093

Fischer DH (1996) The great wave. price revolutions and the rhythm of history. Oxford University Press, Oxford

Gabillon J (1992) The term structure of oil futures prices. Working paper. Oxford Institute for Energy Studies

Gibson R, Schwartz ES (1990) Stochastic convenience yield and the pricing of oil contingent claims. J Finance 45:959–976

Keynes JM (1930) The applied theory of money. Macmillan & Co., London

Ng VK, Pirrong SC (1994) Fundamentals and volatility: storage, spreads, and the dynamics of metals prices. J Bus 67(2):203–230

Shoenmakers J, Coffee B (2000) Stable implied calibration of a multi-factor LIBOR model via a Semi-parametric correlation structure. Weierstrass Institute, Preprint No. 611

Chapter 13
Interest Rates

The topic of this chapter is pricing and risk management of derivatives that depend on interest rates. These products can be divided into two classes. The first class consists of derivatives that only depend on interest rates and no other asset classes. An example is given by an option paying the positive part of the difference between an interest rate and a fixed rate. Regarding the second product type, observe that derivatives on other asset classes such as equity, commodities and FX contain discount factors from the payment dates. This introduces an interest rate component into the pricing. As we see in this chapter, the impact of the interest rate volatility is usually much smaller than the contribution from the volatility of the main underlying in the contract. It is therefore often possible to assume deterministic interest rates without too much loss of accuracy. The exceptions are typically for long-dated products where a stochastic model for interest rates is necessary for a proper pricing and risk management. In this second class of derivatives we also include interest rate hybrids for which the dependence on interest rate is explicit, e.g. convertible bonds.

Just as for other asset classes, interest rate derivatives can be divided into vanillas and exotics. We define the former as consisting of products that can be modeled with a stochastic process of only a single underlying rate. For the pricing of exotics, the evaluation of the whole yield curve is often necessary. This modeling is similar to commodity exotics pricing where a whole curve also needs to be evolved. The most important difference is that measure transformations are non-trivial for yield curve models. As such transformations are often necessary for the calibration, interest rates modeling can be quite complicated. We have therefore chosen to provide a detailed discussion on interest rate modeling which results in a longer chapter than for the other asset classes.

We start with a review basic interest rate terminology and explain how vanillas can be priced. After a short discussion of convexity adjustment, we turn our attention to yield curve models. Our main focus is on a normally distributed model for which we discuss the relation to short-rate models, HJM models, market models and Markov-functional models. We then consider the stochastic impact of interest rates on other asset classes. We disregard liquidity and credit risk in the first part of

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_13,

the chapter. It means that we limit ourselves to the way interest rate modeling was done before 2007. The more general situation is considered from Sect. 13.22 and onwards.

13.1 Interest Rates and Conventions

Recall that the zero-coupon bond P_{tT} represents the time t value of the contract paying \$1 at T. Therefore, a fair loan of \$1 at t should be repaid with P_{tT}^{-1} at T. The earning per time for the lender is

$$L_{tT} = \frac{P_{tT}^{-1} - 1}{T - t}$$

The instantaneous rate of earning is obtained from

$$r_t = \lim_{dt \to 0} L_{t,t+dt} = \lim_{dt \to 0} \frac{P_{t,t+dt}^{-1} - P_{tt}^{-1}}{dt} = \frac{d}{ds} P_{ts}^{-1}|_{s=t} = -\frac{d}{ds} P_{ts}|_{s=t}$$

r_t is called the *short rate*.

Assuming that the short rate is independent of time, $r_t = r$, an investment of \$1 grows to \$$(1 + rdt)$ during a small time interval dt. Separating the time period $[t, T]$ into N parts of size $(T - t)/N$, we see that \$1 grows to

$$\left(1 + r\frac{T - t}{N}\right)^N \to e^{r(T-t)} \text{ when } N \to \infty$$

dollars from t to T. It means that

$$P_{tT}^{-1} = e^{r(T-t)}$$

Based on the above discussion we define the *simple compounded interest rate* L_{tT} by

$$P_{tT}^{-1} = 1 + (T - t)L_{tT}$$

and the *continuously compounded interest rate* r_{tT} by

$$P_{tT}^{-1} = e^{r_{tT}(T-t)}$$

Another important concept is the *n-compounded interest rate* r'_{tT} defined by

$$P_{tT}^{-1} = (1 + r'_{tT}/n)^{n(T-t)}$$

The continuously compounded rate is obtained in the limit $n \to \infty$. Of the above rates, the most commonly used are the annual compounded rate ($n = 1$), the semi-annual compounded rate ($n = 2$), the quarterly compounded rate ($n = 4$) and the simple compounded rate. The continuously compounded rate is rarely used in practice. It is, however, often used in interest rate modeling because of the cocycle relation: $P_{tT_1} P_{T_1 T_2} = P_{tT_2}$, which holds when the continuously compounded rate is constant.

An important example of a simple compounded interest rate is the *London interbank offered rate (LIBOR)* produced by *British Bankers' Association (BBA).* It is an average of the offered lending rates between major banks in the London interbank market. LIBOR is given each day for 15 different maturities up to one year and for 10 currencies (AUD, CAD, CHF, DKK, EUR, BGP, JPY, NZD, SEK and USD). There also exist rates defined in a similar way but on the local interbank markets: EURIBOR (EUR), HIBOR (HKD), SABOR (ZAR), SIBOR (SGD), TIBOR (JPY), etc. To simplify the terminology, we from now on refer to all simple compounded rates as LIBOR rates.

The value of a zero-coupon bond varies widely with the values of t and T. The limiting values are 0 when $T \to \infty$ and 1 when $t \to T$. Interest rates, on the other hand, have a much weaker dependence on t and T. Indeed, interest rates have historically most often been in the order of magnitude of 5%. As rates have more stable values than zero-coupon bonds, they are more suitable to use for modeling. This is analogous to the preference of working with volatilities instead of directly with option prices.

An ambiguity in the definition of interest rates is the computation of $T - t$. For instance, if $t = 26$ Aug and $T = 26$ Sep, should $T - t$ be defined as 1/12 since it is exactly one month or do we define it as 31/365 or in any other way? In fact, there is no standard convention and to define an interest rate it is necessary to specify the *day-count convention* that is used.

Commonly used day-count conventions are actual/365 and actual/360, for which the *day-count fraction* $T - t$ is computed by dividing the number of days between t and T by 365 or 360. For example, $(T - t)_{\text{actual/365}}$ is equal to $31/365 \approx 0.0849$ in the above example. Another common day-count convention is 30/360 where every month is counted as 30 days and every year as consisting of 360 days, and actual/actual for which the actual number of days between t and T is divided by the actual number of days in the year. There are several versions of the mentioned day-count conventions depending on how to count the number of days in a month, how to treat the short month of February and how to treat leap years.

A *forward interest rate* is an interest rate for a time period $[T_0, T_1]$ in the future, $t < T_0 < T_1$. For example, the LIBOR rate for $[T_0, T_1]$, as observed from t, is defined by

$$P_{tT_1}^{-1} / P_{tT_0}^{-1} = 1 + (T_1 - T_0) L_{tT_0T_1}$$

The continuously compounded forward interest rate $f_{tT_0T_1}$ is defined by

$$P_{tT_1}^{-1} / P_{tT_0}^{-1} = e^{f_{tT_0T_1}(T_1 - T_0)}$$

It follows in particular that $L_{tT} = L_{ttT}$ and $r_{tT} = f_{ttT}$. By letting $T_0 \to T_1$ we obtain the instantaneous forward rate

$$f_{tT} = -\frac{d}{dT} \ln P_{tT}$$

which is related to the short rate by $r_t = f_{tt}$. As the above expression is invertible,

$$P_{tT} = \exp\left(-\int_t^T f_{tT'} dT'\right)$$

the zero-coupon bond prices can be computed from the instantaneous forward interest rates and vice versa.

13.2 Static Replication

This section deals with interest rate products that can be statically replicated with zero-coupon bonds. Example of such products are coupon-paying bonds, FRAs and swaps.

We start by discussing the relation between coupon-paying bonds and zero-coupon bonds. The time t value of a bond with notional N, fixed coupon c and payment dates $T_0, T_1, \ldots, T_n$ is given by

$$V_n = N\left(c \sum_{i=0}^{n} P_{tT_i} + P_{tT_n}\right)$$

which shows how to express the price of a coupon-paying bond in terms of the prices of zero-coupon bonds. To obtain as simple formulae as possible, we assume that the notional is equal to 1 unless otherwise stated. Subtracting the above price with the bond V_{n-1} with one fewer payment date gives

$$V_n - V_{n-1} = cP_{tT_n} + P_{tT_n} - P_{tT_{n-1}} \Leftrightarrow P_{tT_n} = \frac{V_n - V_{n-1} + P_{tT_{n-1}}}{1+c}$$

As $V_0 = (c+1)P_{tT_0}$, it follows that zero-coupon bond prices can be bootstrapped from bond prices V_i.

For a given future period $[T_0, T_1]$ a *forward rate agreement (FRA)* pays the LIBOR rate $L_{T_0T_1}$ in return for a fixed rate K. As typical for interest rate products, the payment is multiplied by the day-count fraction and notional (which we have assumed to be equal to 1):

$$\text{FRA}(T_1) = (T_1 - T_0)(L_{T_0T_1} - K)$$

As $L_{T_0T_1}$ is known at T_0, the value of the FRA can be discounted to T_0 and this is where the payment of the FRA usually takes place:

$$\mathrm{FRA}(T_0) = (T_1 - T_0)(L_{T_0T_1} - K)P_{T_0T_1} = (T_1 - T_0)\frac{L_{T_0T_1} - K}{1 + (T_1 - T_0)L_{T_0T_1}}$$

The part containing K consists of a payment $-(T_1 - T_0)K$ at T_1 and is therefore worth $-(T_1 - T_0)KP_{tT_1}$ at t. The part containing $L_{T_0T_1}$ can be written as

$$(T_1 - T_0)L_{T_0T_1}P_{T_0T_1} = P_{T_0T_0} - P_{T_0T_1}$$

at T_0, which means that it is worth

$$P_{tT_0} - P_{tT_1} = (T_1 - T_0)L_{tT_0T_1}P_{tT_1}$$

at t. We conclude that

$$\mathrm{FRA}(t) = (T_1 - T_0)(L_{tT_0T_1} - K)P_{tT_1}$$

At the initiation of the contract, K is chosen so that the contract values to par: $K = L_{tT_0T_1}$. A FRA represents the difference between a loan with a *floating rate* $L_{tT_0T_1}$, i.e. a rate that depends on t, and a loan with a *fixed rate* K, i.e. a rate that is independent of t.

For a given set of dates $T_0, T_1, \ldots, T_n$, a *vanilla interest rate swap* pays the floating rate $L_{T_iT_{i+1}}$ versus a fixed rate K for each interval $[T_i, T_{i+1}]$. Thus, it consists of a strip of FRAs with the same fixed rate on consecutive and adjacent time periods. The time t value of a swap is therefore given by

$$\begin{aligned}
V_t &= \sum_{i=0}^{n-1}(T_{i+1} - T_i)(L_{tT_iT_{i+1}} - K)P_{tT_{i+1}} \\
&= \sum_{i=0}^{n-1}((P_{tT_i} - P_{tT_{i+1}}) - (T_{i+1} - T_i)KP_{tT_{i+1}}) \\
&= P_{tT_0} - P_{tT_n} - K\sum_{i=0}^{n-1}(T_{i+1} - T_i)P_{tT_{i+1}}
\end{aligned}$$

As opposed to FRAs, the LIBOR payments $L_{T_iT_{i+1}}$ are made at the end T_{i+1} of the period.

Just as for forward contracts, K is chosen so that the swap value is zero at initialization:

$$K = \frac{P_{tT_0} - P_{tT_n}}{\sum_{i=0}^{n-1}(T_{i+1} - T_i)P_{tT_{i+1}}}$$

This value of K is called the *swap rate*. The set of dates $T_0, T_1, \ldots, T_n$ is called the *tenor structure*, the length of the swap $T_n - T_0$ is called the tenor and the denominator in the above expression is called the *annuity* and is denoted by A. Observe that a LIBOR rate is a special case of a swap rate for which there are only two tenor dates. The swap often starts at today's date: $t = T_0$. If $T_0 > t$, it is called a *forward starting swap*.

We denote the swap rate by R. A swap initialized at $t = 0$ has strike K equal to

$$R_0 = \frac{P_{0T_0} - P_{0T_n}}{A_0}$$

Using the swap rate, the value of a swap can be expressed as

$$V_t = \left(\frac{P_{tT_0} - P_{tT_n}}{\sum_{i=0}^{n-1}(T_{i+1} - T_i)P_{tT_{i+1}}} - K \right) \sum_{i=0}^{n-1}(T_{i+1} - T_i)P_{tT_{i+1}} = (R_t - R_0)A_t$$

This expression is both simple and natural. For example, it follows immediately that $V_{t=0} = 0$.

The use of swap rates for the modeling of swaps is as natural as the use of LIBOR rates for FRAs. The validity of this statement becomes explicit in the next section where we price options on FRAs and swaps. Furthermore, as LIBOR rates can be bootstrapped from swap rates (by bootstrapping zero-coupon bonds as an intermediate step), and vice versa, swap rates are as fundamental quantities for the yield curve modeling as LIBOR rates. The choice of rate type depends on the product to be modeled.

The tenor dates are generally different for the two payment legs in a swap. For example, the floating-rate payments can be semi-annual while the fixed-rate payments can be annual. Furthermore, the business-day adjustments and the day-count conventions can differ between the legs. This leads to minor (but important) modifications of the swap pricing formula.

We would like to emphasize that in practice, zero-coupon bonds (or equivalently, loans) cannot always be used directly in replication formulae. Instead, we rather view them as useful mathematical building blocks. For example, replicating a FRA with coupon-paying bonds is most easily done by replicating the FRA with zero-coupon bonds as an intermediate step and after that replicating the zero-coupon bonds with coupon-paying bonds.

13.3 Caps, Floors and Swaptions

We now turn our attention to interest rate products that cannot be priced with static replication. In this section, we discuss the simplest of such products, namely caps, floors and swaptions. The reason for their simplicity is that they can be modeled as depending on only a single underlying. The resulting pricing expressions are similar to the Black–Scholes formula.

A *caplet* with strike K and period $[T_0, T_1]$ pays

$$C(T_1) = (T_1 - T_0)(L_{T_0T_1} - K)_+$$

which is the positive part of a FRA. Similarly, a *floorlet* pays

$$F(T_1) = (T_1 - T_0)(K - L_{T_0T_1})_+$$

at T_1. As

$$C - F = FRA$$

a floorlet can be statically replicated using a caplet and a FRA. Therefore, it is sufficient to focus on caplet pricing. Being a non-linear function of the LIBOR rate, a caplet cannot be priced through static replication of FRAs. We instead use the fundamental theorem of asset pricing to price caplets with dynamic replication.

As

$$L_{tT_0T_1} = \frac{P_{tT_0} - P_{tT_1}}{(T_1 - T_0)P_{tT_1}}$$

the LIBOR rate can be written as a quotient of tradable assets. With P_{tT_1} as the numeraire, $L_{tT_0T_1}$ is a martingale. We make the standard assumption of a lognormal process

$$dL_{tT_0T_1} = \sigma L_{tT_0T_1} dW_t$$

where W is a Brownian motion in the forward measure. As $C(t)/P_{tT_1}$ is also a martingale, it follows that

$$C(t) = P_{tT_1} E[C(T_1)/P_{T_1T_1}] = (T_1 - T_0)P_{tT_1} E[(L_{T_0T_1} - K)_+]$$

Using the calculations of Sect. 5.2, we obtain

$$C(t) = (T_1 - T_0)P_{tT_1}(L_{tT_0T_1}N(d_+) - KN(d_-)),$$
$$d_\pm = \frac{\ln(L_{tT_0T_1}/K)}{\sigma\sqrt{T_0 - t}} \pm \frac{1}{2}\sigma\sqrt{T_0 - t}$$

The caplet pricing can also be done by using a direct hedging argument as we did in Sect. 3.2 for European call options. It is then made explicit that caplets are hedged with FRAs.

Caplets are often not traded themselves, but they rather appear as constituents of a cap. A *cap* defined on a tenor structure $T_0, T_1, \ldots, T_n$ consists of a strip of caplets with identical fixed rate on the consecutive and adjacent time periods. Let c_i be the price of a caplet with time period $[T_i, T_{i+1}]$. As the caplets constituting a cap can be priced independently, the price of a cap is given by $\sum_{i=0}^{n-1} c_i$. A *floor* is defined from floorlets in a similar way. Caps (and floors) are quoted in the market in terms of their implied volatility. By definition, the market price is obtained by inserting the implied volatility into the above caplet formula for all the caplets constituting a cap.

Just as for European options, a lognormal process only gives a rough approximation to the caplet and floorlet prices. More accurate models can be obtained by using, for example, local volatility or stochastic volatility. The most popular model for caplets and floorlets is currently the SABR model.

Caplets are important instruments in interest rate modeling as they are well suited for calibration of exotic interest rate products, see Sect. 13.20. It is then first necessary to back out the caplet prices from cap quotes. In theory, this looks simple: the price difference between two caps with tenor structure $T_0, \ldots, T_{i+1}$ and $T_0, \ldots, T_i$ is exactly that of a caplet c_i. In practice, however, it can be quite complicated as only caps with annual tenors are liquid, while caplets are semi-annual or quarterly depending on the currency. It means that there are two or four caplets between consecutive liquid caps. The caplet bootstrapping is therefore non-trivial since it does not give the price of individual caplets directly but only of the sum of two or four of them.

The market conventions for caps and floors are derived from the underlying FRAs, see Table 13.2 in Sect. 13.6, with the exception of the 1Y cap (and sometimes the 2Y cap) which is quarterly even when the standard market convention is semi-annual. Furthermore, the first caplet, which would have been fixed today, is always missing. For instance, a 3Y EUR cap, has a premium payment at the spot date, i.e. 2 business days after today according to the target calendar. The underlying rates starts at 6M, 12M, 18M, 24M and 30M and have lengths of 6M. The ATM strike is determined from the underlying swap rate which in our example starts at 6M and ends at 3Y.

A *swaption* is an option on a forward starting swap. At the swap start date, the swaption value is the positive part of the swap price:

$$V(T_0) = (R_{T_0} - R_{t=0})_+ A_{T_0}$$

Using

$$R_t = \frac{P_{tT_0} - P_{tT_N}}{A_t}$$

and the fact that A_t is a tradable as it is a linear combination of tradables, it follows that $V(t)/A_t$ and R_t are martingales in the measure for which A_t is a numeraire. The fundamental theorem of asset pricing implies that

$$V(t) = A_t E[V(T_0)/A_{T_0}] = A_t E[(R_{T_0} - R_{t=0})_+]$$

Assuming a lognormal process for the swap rate, we obtain

$$V(t) = A_t(R_t N(d_+) - R_0 N(d_-)), \quad d_\pm = \frac{\ln(R_t/R_0)}{\sigma\sqrt{T_0 - t}} \pm \frac{1}{2}\sigma\sqrt{T_0 - t}$$

Just as for caplets, floorlets and European options, a non-lognormal process can account for skew and smile. The most popular model is currently the SABR model.

There is a credit default risk embedded in swaptions as the premium is paid at spot date while the additional cash flows occur after (or at, in the instance when the swaption is cash settled) the swap start date. The effect of this credit risk became particularly apparent during the financial crisis starting in 2007. The outcome has been that the swaption conventions have changed for many currencies so that the premium is paid at the swap start date.

A swaption is exercised when the underlying swap has a cash flow that is positive in average. Such positive net position introduces a credit exposure, in particular when the swap tenor is long. The credit exposure can be avoided via a one time payment of the swap market value at the swaption exercise date. The swaption is then said to be *cash-settled*. The problem is that the counterparties do not always agree on the market value as their models predict different values of the discount factors that enter the swap pricing via the annuity.

For GBP and EUR swaptions, the problem is solved by agreeing that the discount factors in the annuity should be computed using the swap rate, for which there is no ambiguity. More precisely, these swaptions pay by definition the amount

$$P(t_{\mathrm{exp}}, T_0)A(S_{t_{\mathrm{exp}}})(S_{t_{\mathrm{exp}}} - K)_+$$

where t_{exp} is the expiry of the swaption and $A(S)$ can be viewed as an approximation of the annuity using a fixed rate S:

$$A(S) = \sum_{i=1}^{n-1} \frac{\tau}{(1+\tau S)^i} = \frac{1}{S}\left(1 - \frac{1}{(1+\tau S)^{n-1}}\right)$$

where we for simplicity have assumed a constant day count fraction τ. The market convention for cash-settled swaptions makes them harder to price than the *physically settled swaptions* that we considered earlier. For instance, using the annuity as the numeraire, we obtain the following expression for the price:

$$A_t E[P(t_{\mathrm{exp}}, T_0)A(S_{t_{\mathrm{exp}}})(S_{t_{\mathrm{exp}}} - K)_+/A_{t_{\mathrm{exp}}}]$$

which cannot easily be computed since $A_{t_{\mathrm{exp}}}$ depends on various interest rates and not only on S_{exp}. As swaptions are liquid vanilla products, performance is important and models depending on the whole yield curve are therefore in general not an option. A popular approach to obtain an efficient expression is to observe that

$P(t_{\exp}, T_0)A(S_{t_{\exp}})/A_{t_{\exp}}$ is a low volatility process and can be frozen to its value at t. The expression can then be taken outside the expectation, resulting in the price

$$P(t_, T_0)A(S_t)E[(S_{t_{\exp}} - K)_+]$$

where the expectation can be computed in the same way as for physically settled swaptions.

Needless to say, care should to be taken when using this kind of simplifying assumptions. For instance, as pointed out in Mercurio (2008), the above formula is in general not arbitrage free.

Interest rate and commodities markets are different from equities and FX since for a given underlying, there is often only a single exercise date for which liquid options exist. For instance, swaptions on swaps starting two years from now are only liquidly traded when the exercise is a market-specific number of business days (typically 2) before the start date of the swap. As options with different exercise dates are on different underlyings, the temporal no-arbitrage condition derived in Sect. 2.4 cannot be used for liquid options in interest rate and commodities markets.

For the liquid volatility products in interest rate markets, there are two other non-trivial no-arbitrage conditions that need to be taken into account. The first one follows from the obvious fact that a swaption on a swap starting in m years and ending in n years must be worth less than the difference between the n year cap and the m year cap with the same strike as the swaption. Indeed, for the most favorable market moves for the swaption all the cash flows will be positive and agree with those coming from the cap spread.

The second condition states that a swaption with tenor structure $T_0, T_1, \ldots, T_n$ is worth less than the sum of the two swaptions with the same strike and tenor structures $T_0, T_1, \ldots, T_m$ and $T_{m+1}, T_{m+2}, \ldots, T_n$. This statement is based on the inequality

$$\begin{aligned} &\left(\sum_{i=0}^{n-1}(T_{i+1} - T_i)(L_{tT_iT_{i+1}} - K)P_{tT_{i+1}}\right)_+ \\ &\quad \le \left(\sum_{i=0}^{m-1}(T_{i+1} - T_i)(L_{tT_iT_{i+1}} - K)P_{tT_{i+1}}\right)_+ \\ &\qquad + \left(\sum_{i=m}^{n-1}(T_{i+1} - T_i)(L_{tT_iT_{i+1}} - K)P_{tT_{i+1}}\right)_+ \end{aligned}$$

At the expiry $t = T_0$, the left-hand side is the price of the swaption with tenor structure $T_0, T_1, \ldots, T_n$ while the first term on the right-hand side is the price of the swaption with tenor structure $T_0, T_1, \ldots, T_m$. The second term on the right-hand side is the price of the swaption with tenor structure $T_{m+1}, T_{m+2}, \ldots, T_n$ when the exercise decision has to be taken at T_0. As it is suboptimal to exercise swaptions early, see Sect. 2.5, this swaption is worth less than the corresponding standard swaption. The statement is therefore true for $t = T_0$ and by the no-arbitrage principle also for general $t \leq T_0$.

13.4 Convexity Adjustment

We have explained how to price LIBOR and swap rate payments, and options on these, when the payments occur at the natural dates of the rates. There also exist products for which the rates are paid out at other dates. We here explain how the pricing can be done to a high degree of accuracy without resorting to models for the whole yield curve.

We saw in Sect. 13.2 that the time t value of a contract paying $(T_1 - T_0)L_{T_0T_1}$ at T_1 is equal to

$$(T_1 - T_0)L_{tT_0T_1}P_{tT_1} = P_{tT_0} - P_{tT_1}$$

This contract was particularly simple to price since it is equivalent to being long a bond maturing at T_0 and short a bond maturing at T_1. We now compute the price of a contract that pays $L_{T_0T_1}$ at a date different from T_1. As $L_{T_0T_1}$ is known at T_0, it is possible to pay the rate at any date after T_0. In fact, it is common that $L_{T_0T_1}$ is paid at its fixing date T_0, i.e. the date when the rate is read off from the market. We see below that it is then no longer possible to statically replicate the contract with zero-coupon bonds. Indeed, such a replication is only possible when the payment is made on the inherent payment date T_1 in the definition of the rate.

We consider in detail the pricing of a *LIBOR-in-arrears* payment, i.e. a cash flow that pays the LIBOR rate $L_{T_0T_1}$ at the fixing date T_0. As the discount factor between T_0 and T_1 can be expressed in terms of the LIBOR rate itself, the time T_0 payment of $L_{T_0T_1}$ is equivalent with the following payment at T_1:

$$L_{T_0T_1}(1 + (T_1 - T_0)L_{T_0T_1}) = L_{T_0T_1} + (T_1 - T_0)L^2_{T_0T_1}$$

The first term is nothing but a LIBOR payment at its natural payment date, for which the pricing is obvious. As $L_{T_0T_1}$ is typically in the order of magnitude of 5%, the second term is much smaller. Furthermore, as this term is a convex function of $L_{T_0T_1}$, it is referred to as a convexity adjustment. Thus, the value of a LIBOR-in-arrears payment is equal to an ordinary LIBOR payment plus the convexity adjustment.

As a non-linear expression of the underlying rate $L_{T_0T_1}$, the convexity adjustment is often priced with dynamic replication. The calculations are similar to the ones for caplet pricing in Sect. 13.3, with the terminal condition $L^2_{T_0T_1}$ instead of $(L_{T_0T_1} - K)_+$. As the standard assumption of a lognormal model disregards skew and smile, the static replication technique of Sect. 2.3 can be preferable. To understand the details of how static replication can be used, consider the pricing of the convexity adjustment and caplets within the fundamental theorem of asset pricing:

$$\begin{cases} V_t' = P_{tT_1} E[L_{T_0T_1}^2] = P_{tT_1} \int_0^\infty L_{T_0T_1}^2 p(L_{T_0T_1}) dL_{T_0T_1} \\ V_t = P_{tT_1} E[(L_{T_0T_1} - K)_+] = P_{tT_1} \int_0^\infty (L_{T_0T_1} - K)_+ p(L_{T_0T_1}) dL_{T_0T_1} \end{cases}$$

The PDF p can be solved from the second equation by taking the second derivative with respect to the strike K. Inserting the PDF in the first equation gives the statically replicated value

$$V_t' = \int_0^\infty L_{T_0T_1}^2 \frac{\partial^2 V_t}{\partial K^2}(L_{T_0T_1}) dL_{T_0T_1}$$

of the convexity adjustment.

The pricing of a contract paying $L_{T_0T_1}$ at an arbitrary time $T' > T_0$ is similar. Let us first assume that $T' \in (T_0, T_1)$. The time T' payment can then be discounted to T_1 by assuming a rate of $L_{T_0T_1}$ in the time period from T' to T_1. The equivalent payment at T_1 is then given by

$$L_{T_0T_1}(1 + (T_1 - T')L_{T_0T_1})$$

which can be priced in a similar way as the LIBOR-in-arrears contract. Because of the natural properties of the continuously compounded rate, see Sect. 13.1, it appears more natural to assume a constant continuously compounded rate r in the time period $[T_0, T_1]$ instead of a constant simple compounded rate. r is then defined by

$$1 + (T_1 - T_0)L_{T_0T_1} = e^{r(T_1 - T_0)}$$

which gives an inverse discount factor from T' to T_1 of

$$e^{r(T_1 - T')} = (1 + (T_1 - T_0)L_{T_0T_1})^{(T_1 - T')/(T_1 - T_0)}$$

Although this formula gives a more consistent price for the contract paying $L_{T_0T_1}$ at time T', the implementation can be a performance bottleneck for Monte Carlo pricing of exotic products due to the expensive power function.

For contracts paying $L_{T_0T_1}$ at time $T' > T_1$ it is again possible to use the rate $L_{T_0T_1}$ for the discounting to T'. Unfortunately, the approximation breaks down when $T' \gg T_1$ as it is then no longer possible to view the contract as depending on only a single rate $L_{T_0T_1}$ but it depends on the rate $L_{T_1T'}$ as well.

We would like to point out that LIBOR-in-arrears contracts are usually not traded by themselves. Instead, the computations above should be considered as a useful exercise for pricing more complex contracts that contain LIBOR-in-arrears or other convexity-adjusted payments. For instance, for *in-arrear swaps* the floating payments are on the fixing dates rather than on the natural payment dates of the underlying LIBOR rates. Another example is given by the situation when a swap has its own payment schedule, which is often close, but not identical to the tenor structure of the underlying LIBOR rates. The convexity adjustment is then quite small, but can anyway be important because of the tight bid-offer spreads in the swap market.

We have analyzed interest rate swaps for which counterparty A pays counterparty B LIBOR in return for fixed-rate payments. This is only one of many combinations of swap structures. For example, LIBOR payments of different frequencies can be swapped. We now consider *constant maturity swaps (CMSs)* for which the LIBOR payments made by counterparty A are replaced with swap-rate payments. The swap rate in the payments is not necessarily related to the underlying tenor structure of the swap. For instance, the swap could be 10 years long with tenor structure T_0, $T_1, \dots, T_{10}$, where $T_{i+1} - T_i = 1\text{Y}$ and such that A pays the 5 year swap rate at each payment date in return for a fixed rate payment, for example, 5%. The payments made by B are not necessarily restricted to a fixed rate, but can be a floating rate such as a LIBOR or swap rate.

The value of a CMS can be computed by decomposing it into a sum of its individual cash flows. The non-trivial component in a CMS is therefore the pricing of a payment of a swap rate at one particular date T. We have already seen in Sect. 13.3 that if the swap rate is paid out on its underlying tenor structure, the pricing is straightforward through static replication of zero-coupon bonds. The pricing of a CMS payment is more complicated because the swap rate is not paid at its natural payment dates, but rather at a fixed single date. This is analogous to LIBOR-in-arrears pricing for which a convexity adjustment is needed.

Let $V(t)$ be the time t value of the payment of a swap rate R_T at T. Following the technique of Sect. 13.3 for swaption pricing, we use the annuity A as numeraire. Then both R_t and $V(t)/A_t$ are martingales and

$$V(t) = A_t E[V(T)/A_T] = A_t E[R_T A_T^{-1}]$$

As R_t is a martingale, we can, for example, assume it to follow a driftless lognormal process. Since A has an interest rate dependence that cannot be expressed in terms of R, this is not sufficient information for computing the expectation. Using our experience from LIBOR-in-arrears pricing, we know that the problem can be solved by approximating A in terms of R. The approximation can be made in several different ways, see Hagan (2003), for details. A particularly popular approximation is the one used in the definition of cash-settled swaptions. Observe that the approximation might not be accurate if the swap rate has a long tenor. The approximation can then be improved by letting A depend on one more rate and using a 2-factor model.

As for LIBOR-in-arrears, dynamic replication with a lognormal process does not take into account the market skew and smile. It can therefore be preferable to statically replicate CMS payments with swaptions. This technique is similar to the way caplets were used to statically replicate LIBOR-in-arrear payments.

The value of the convexity adjustment is heavily dependent on the choice of model. In fact, a poor choice of model can even lead to infinite prices. Consider, for instance, the stochastic volatility model

$$\begin{aligned} dL_t &= \sigma_t L_t d W_t \\ d\sigma_t &= \epsilon \sigma_t d Z_t \\ d W_t d Z_t &= 0 \end{aligned}$$

The LIBOR-in-arrears convexity adjustment can be computed using conditional expectation as the two Brownian motions are independent:

$$\begin{aligned} E[L_T^2] &= E\left[E\left[L_T^2 | dL_t = \sigma_t L_t d W_t\right] | d\sigma_t = \epsilon_t \sigma_t d Z_t\right] \\ &= L_0^2 E\left[\exp\left(\int_0^T \sigma_t^2 dt\right) | \sigma_t = \sigma_0 \exp\left(\epsilon Z_t - \epsilon^2 t/2\right)\right] \end{aligned}$$

Since

$$E\left[\exp\left(\sigma_T^2 T\right) | \sigma_t = \sigma_0 \exp\left(\epsilon Z_t - \epsilon^2 t/2\right)\right] = \infty$$

it follows that $E[L_T^2]$ is divergent. It is straight forward to verify divergence also for non-zero correlation. Thus, LIBOR-in-arrears payments have infinite prices in models for which both the underlying and the volatility are lognormal. This is obviously also true for CMS payments. As CMS payments are more complicated to price we from now on focus on them.

The above model can be obtained by setting $\beta = 1$ in the SABR process. As this is the most popular model for caplets and swaptions, the above divergence is important not only in theory but also for practical modeling.

Using a model with CEV parameter $\beta < 1$ avoids the divergence but reveals other weaknesses in the modeling. The problem is that caplets and swaptions are often priced with perturbation techniques (such as the SABR model). These models are successful when the strike is close to the forward. However, if calibrating the models to liquid swaption quotes, CMS payments are typically mispriced as they depend heavily on out of the money swaptions.

One solution is to avoid perturbation techniques and use, for example, Fourier transform methods, see Sect. 7.4. Needless to say, it can be dangerous to use one model for the CMS book and another model for the swaption book. To avoid this inconsistency, non-perturbative methods need to be used also for swaptions. Unfortunately, this can lead to performance bottlenecks.

As SABR is so generally accepted, it is popular to use this model for swaptions but to modify it for high strikes to obtain CMS prices that are in agreement with the market. This can be done by introducing a cutoff strike K_{cutoff} and use SABR (or any other perturbative technique) when $K < K_{\text{cutoff}}$ and an alternative model for $K > K_{\text{cutoff}}$. Clearly, K_{cutoff} should be large enough so that liquid swaptions are priced with the perturbative model. We now discuss what possibilities there are to model swaptions beyond the cutoff.

A common way to extrapolate swaptions is to construct the implied volatility as a functional form $\sigma_{\text{imp}}(K)$, $K > K_{\text{cutoff}}$, that dampens the increasing behavior of the implied volatility which is common in perturbative models. This function must be patched together at the cutoff with the implied volatility from the perturbative model. As the PDF is obtained from the second derivative of the swaption price with respect to the strike, a discontinuity in the second derivative of σ_{imp} is equivalent with a discontinuity in the PDF itself. Even worse, a discontinuity in the first derivative of σ_{imp} implies a spike in the PDF. Thus, it is important that σ_{imp} is patched smooth enough at the cutoff point K_{cutoff}. Furthermore, the functional form of σ_{imp} has to be chosen carefully to avoid arbitrage. This is particularly difficult as the no-arbitrage condition is complex when expressed in terms of the implied volatility.

Due to the complexity in the construction of σ_{imp} beyond the cutoff, the extrapolation is sometimes done in terms of the PDF. This is clearly possible since the implied volatility can be obtained from the PDF and vice versa. There are advantages as well as disadvantages to doing the extrapolation in terms of the PDF. One advantage is that the smoothness is two degrees better: a discontinuity in the PDF cannot be seen directly on the volatility curve since it is only a discontinuity in the second derivative. Furthermore, as long as we choose the PDF to be positive, the model will be arbitrage free. To understand the disadvantages, recall that digitals swaptions and ordinary swaptions are priced via the formulae

$$\int_K^\infty p(R)dR$$

$$\int_K^\infty (R-K)p(R)dR$$

Therefore, in order to not change the prices of swaptions with strikes $K < K_{\text{cutoff}}$, the following conditions must be satisfied:

$$\int_{K_{\text{cutoff}}}^\infty p(K)dK = -\partial_K V(K_{\text{cutoff}})/A_t$$

$$\int_{K_{\text{cutoff}}}^\infty Kp(K)dK = (V - K\partial_K V)(K_{\text{cutoff}})/A_t$$

where $V(K)$ denotes the swaption price with strike K. We conclude that a change in σ_{imp} for $K > K_{\text{cutoff}}$ leads to a change in $p(K)$ only for $K > K_{\text{cutoff}}$, while a change in $p(K)$ for $K > K_{\text{cutoff}}$ leads to a change in σ_{imp} only for $K > K_{\text{cutoff}}$ conditional on that the two conditions above are satisfied.

An additional constraint on p is that it must give reasonable CMS prices. Finally, it is desirable to have a functional form of p that reduces to the lognormal PDF for the case when the original (perturbative) model is lognormal.

We have so far discussed the trouble caused by the asymptotics for high strikes when using perturbative techniques. In the same way, the small strike expansion is not in agreement with far from ATM swaption prices. The difference from the high strike case is that this effect does not show up as drastically in CMS prices as they depend mainly on high strikes. To obtain more consistent prices, it is common to introduce a cutoff for small strikes as well.

For the SABR model, the introduction of a low strike cutoff K_{low} is crucial. Indeed, assume that we are backing out the PDF $p(K_{\text{low}})$ for a low strike by taking the second derivative of swaption prices with respect to the strike. We compare the result with the integral over the PDF up to K_{low}. This value is obtained from the digital swaption price at K_{low}. For long maturities we often see the inequality

$$K_{\text{low}}\, p(K_{\text{low}}) < \int_0^{K_{\text{low}}} p(K)dK$$

for modest values of K_{low} when using the SABR model. For this inequality to be true, p must have a local maximum in the region $[0, K_{\text{low}}]$ of low strikes. This is clearly an unnatural feature of the PDF and it has its roots in the choice of SABR boundary conditions that was used in the original paper by Hagan et al. (2002). The boundary conditions are such that they support negative values of the underlying. Therefore, one solution to the above problem is to allow negative interest rates. However, this can be rather dangerous since the pricing of low strike swaptions will be inaccurate. A better alternative seems to be the use of a perturbative model with an appropriate boundary condition when the underlying is zero, see Sect. 7.2.

By using a perturbative model that contains more expansion terms than the lowest order, it is possible to push the cutoff points K_{high} (and K_{low}) further away from the ATM point. The result is CMS prices that are less dependent on the choice of extrapolation. For instance, using the technique of Sect. 7.2, it is possible to include arbitrary higher order contributions.

13.5 The Yield Curve

When referring to the *yield curve*, what is usually meant is the function $T \mapsto r(t,T)$, where the interest rate $r(t,T)$ is simple compounded for $T - t \leq 1$ and annually compounded for $T - t > 1$. The part of the yield curve for which T is close to today t is called the *short end* of the curve while the part where T is large is called the *long end*. The interest rate in the limit $T \to t$ is called the short rate while it is called the long rate in the limit $T \to \infty$.

To understand the shape of the yield curve, assume that we have a certain amount of cash and are choosing between the strategies of: 1. lending it up to time T, or 2. lending it to T/N and when the money is returned we lend it until $2T/N$ and so on up to T. If interest rates are equal and constant, the second strategy is preferable as we at $N-1$ points in time can choose to stop lending if there is something else we would like to do with the money. Should cash be needed in the first strategy, it is possible to enter the opposite position by taking a loan. This netting of positions leads to a loss if interest rates have increased unexpectedly since the first loan was issued. The conclusion is that the first strategy is riskier and should be rewarded with a liquidity premium. Note the analogy with the convenience yield in commodities markets: in this case there is a convenience of holding cash. The implication is that short rates should be lower than long rates, i.e. $r(t, T)$ should be an increasing function of T.

Allowing interest rates to be stochastic does not affect the first strategy as the loan is locked at a fixed rate until T. For the second strategy, on the other hand, the earnings until T are unknown. A premium compensation for the unknown earnings promotes a downwards sloping yield curve.

Another explanation for the shape of the yield curve is that the default risk of the counterparty increases with the lending time T. The credit premium to compensate for this risk has the effect of an increasing yield curve.

The above discussion, just as the discussion regarding convenience yield for commodities, is purely theoretical. In reality, most of the yield curve shape is determined by the expected future values of rates, political decisions and regulations, and by the supply and demand by market participants. For example, the short end is influenced by governments setting the short rate based on the state of the economy in the monetary region concerned. This can sometimes lead to an inverted curve, i.e. a yield curve that decreases with T. The curve can even take peculiar shapes such as having a hump.

For investment purposes it is possible to take advantage of the fact that the yield curve usually has a positive tilt. This can be done by today locking in an interest rate for a future period. By entering the opposite position at a later time it is possible to obtain a return that is positive on the average as the short rate most often is below the long rate. Observe that even if the yield curve stays non-inverted, an investor in such a strategy can still make a loss from a parallel move of the curve. This investment is similar to the roll-yield investment strategy for commodities.

The above strategy is an example of an investment in the tilt, or spread, of the yield curve. Many of the products that depend on the whole yield curve expose the investor to the spread in various ways. A disadvantage of the above strategy is that it depends on the level as well as on the spread. It is therefore popular to structure products, such as *CMS spreads*, that contain the opposite position of the level so that only a pure spread dependence remains.

As mentioned in the introduction, no account to credit and liquidity risk is taken in the first part of this chapter. It means that interest rates can be modelled by a single yield curve, which will be the topic of discussion up to Sect. 13.22, whereafter the generalization to post-2007 interest rate modeling will be considered.

13.6 Yield Curve Instruments

The yield curve in a classical pricing system is constructed from the most liquid products from which the prices of zero-coupon bonds can be backed out. Depending on the currency, either interbank deposits or LIBOR fixings are used for maturities up to 3 months. Thereafter, either futures or FRAs are used out to 2–5 years. Finally, swaps are used for the long end of the curve.

In order to build a yield curve consistently, it is necessary to know the exact market conventions for the product types used in the construction. Except for the futures contracts, these are all OTC contracts. Nevertheless, for each currency there are certain standard conventions and these are the products that can be seen quoted on Reuters and Bloomberg.

We start with the settlement lag which is equal to 2 for most currencies. The main exceptions are AUD, CAD and GBP for which the lag is zero. The start date of an interest rate contract is called the *effective date*. For standard products like those detailed in this section, the effective date is equal to the spot date.

As mentioned in Sect. 3.11, a forward date is determined by adding a tenor to the spot date. For example, a 2Y semi-annual swap has payment dates located 6M, 12M, 18M and 24M after the spot date. The holiday adjustment is in general made with the modified following convention. An exception is deposits with maturity less than one month for which the following convention is used. Furthermore, the end-of-month rule is often applied.

The interest rates are simple compounded and the day-count convention for the floating leg of a swap is the same as for LIBOR fixings, deposits, FRAs and futures. It is actual/365 for AUD, CAD and GBP while actual/360 is used for most other currencies. 30/360 is used for EUROLIBOR and EURIBOR. For instance, consider a 1M USD deposit traded on December 13, a Monday, with quote L and notional N. It has cash flows N on the spot date December 15, and $-\left(1+\frac{33}{360}L\right)N$ on January 17, a Monday.

There are also deposits for which the start date of the rate is different from the spot date. These are the *overnight (O/N) loan* and the *tomorrow next (T/N) loan*. For O/N, the start date is the trade date and the end date is the next business day. For T/N, the start date is the next business day and the end date is the spot date.

FRAs are usually written as SxE, where S is the number of months to the start date and E is the number of months to the end date. For instance, a 3x6 USD FRA traded on February 4, a Friday, has spot date February 8, a Tuesday. The interest rate period has start date May 9, a Monday, and end date August 8, also a Monday. The 3M rate L is fixed May 5, a Thursday, i.e. 2 business days before the start date. Observe that the rate L is defined with respect to the 3M period between May 9 and August 9. A payment of $\frac{91}{360}\frac{L-K}{100}N/\left(1+\frac{91}{360}\frac{L}{100}\right)$ is made at the start date, where K is the fixed par rate determined at the trade date.

For certain currencies it is popular to trade IMM FRAs. They work as ordinary FRAs with the difference that the interest rate period is between two adjacent IMM dates. The *International Monetary Market dates (IMM dates)* are defined as the third Wednesday is March, June, September and December. Just as for ordinary FRAs (and swaps), there can be a few days mismatch between the interest period of the contract and that of the underlying interest rate.

The day-count convention for the fixed leg of a swap is actual/365 for CAD, GBP and JPY, while USD use actual/360 and most European currencies use 30/360. The fixed-leg payments are semi-annual for CAD, GBP, JPY while they are annual for USD and most European currencies. These conventions are closely related to the corresponding bond markets, see Table 13.1.

The floating side of a swap is paid quarterly for GBP, SEK and USD, while it is semi-annual for EUR, CHF and JPY. The fixing of the floating leg is done at the date for which the spot date equals the start date of the accounting period. For example, assume that the floating payment dates of a USD swap have been rolled out and been holiday adjusted, and that two adjacent dates are July 11, a Wednesday, and October 11, a Thursday. The floating rate L is then fixed at July 9, and a payment of $\frac{92}{360}\frac{L}{100}N$ takes place in October 11, where N denotes the notional. Payments are netted if a fixed-leg payment occurs on the same date.

Futures contracts work as exchange-traded FRAs that are settled on a daily basis. The most common contracts have an underlying loan that starts on an IMM date. The last trade date is the Monday preceding the IMM date that is the start date of the loan. The quote is presented in a different way compared to the other instruments in this section: if V is the quote then the simple compounded interest rate for the period is given by $1 - V/100$.

Table 13.1 Market conventions for (government) bonds

Bond Type	Frequency	Day count
US treasury	Semi-annual	Actual/Actual
US corporate	Semi-annual	30/360
UK gilts	Semi-annual	Actual/Actual
Euro government (OATs, bunds)	Annual	Actual/Actual
Italian government	Semi-annual	Actual/Actual
Japanese government	Semi-annual	Actual/365
Canadian government	Semi-annual	Actual/365
Australian government	Semi-annual	Actual/Actual
Eurobonds	Annual	30/360

Table 13.2 Standard conventions for interest rate instruments in some major currencies

Ccy	Lag	Money market dc	Fixed leg dc	Floating freq	Fixed freq
EUR	2	Actual/360	30/360	Semi-annual	Annual
USD	2	Actual/360	Actual/360	Quarterly	Annual
GBP	0	Actual/365	Actual/365	Semi-annual	Semi-annual
JPY	2	Actual/360	Actual/365	Semi-annual	Semi-annual
CHF	2	Actual/360	30/360	Semi-annual	Annual
SEK	2	Actual/360	30/360	Quarterly	Annual

We summarize the interest rate market conventions for some of the major currencies in Table 13.2. Special cases are the 1Y EUR and 1Y CHF swaps which are quarterly on the floating leg and the 1Y GBP swap which has an annual fixed leg payment. USD swaps are often also semi-annual and 30/360 on the fixed leg and this is the standard convention used for USD swaptions.

13.7 Yield Curve Construction

To illustrate the practical issues involved in the construction of a yield curve, let us consider a specific case when the building blocks are deposits up to 3M, futures up to 2Y and then swaps. The short end of the curve is constructed from the liquid deposits that mature in 1M, 2M, 3M, and also from shorter maturities such as 1W and O/N. This curve is patched together with the quotes from eight futures. Then the liquid swaps are used with maturities of 3Y, 4Y, 5Y, . . . , 10Y, 12Y, 15Y, 20Y, . . .

The yield curve can be represented by various different choices of underlying variable. We here choose to define it from the continuous compounded interest rate starting at the spot date. It is straightforward to find the rates that correctly price the deposits. Through interpolation, the discount factor to the start date of the first future can be obtained. The futures quotes then determine forward interest rates from which it is possible to compute the rates all the way up to the end date of the last futures contract. Using the swap quotes it is possible to continue the bootstrapping to the long end. The final step is to pick an interpolation method of choice, e.g. a tension spline, in this way connecting the dots. As pointed out in Sect. 4.3, above choice of interpolation method is important as it determines the smoothness of the curve and the locality of the risk.

There are a couple of complications that one encounters in the above curve construction. First of all, the three instrument types that build up the curve have different features and it is therefore not always possible to patch them together to a global and smooth curve. For instance, as we discuss in Sect. 13.24, the instruments expose the holder to various degrees of credit risk.

To be able to patch futures quotes together with deposit and swap quotes, it is necessary to account for the effect of the daily settlement. As described in Sect. 3.10, this is done by adjusting for the convexity.

Care needs to be taken when including the swap quotes. To understand the issue, assume that the yield curve has been bootstrapped just beyond the 2Y point and that we are about to include a 3Y swap with semi-annual frequency for the fixed leg. The price of this swap involves the unknown discount factors at 30M as well as at 36M. The bootstrapping therefore involves the solution of a problem with two unknowns and only one known variable. Similar complications appear for caplet bootstrapping, see Sect. 13.3. It can be solved by imposing an additional constraint or by interpolating the swap quotes to intermediate tenors.

The difficulty with including swaps originates in the iterative procedure in the bootstrapping process. An alternative approach is to use an interpolation scheme for the yield curve with node points (the x-values) agreeing with the end points in time of the deposits, futures and swaps. The quotes of these instruments are then as many as there are unknowns (the y-values), which means that a root finder can be used. This approach is slower than bootstrapping but allows for a higher flexibility in the choice of interpolation technique. It is also possible to use this type of interpolation scheme with the prices of the underlying instruments only approximately matched. The prices can then be accounted for by the method of adjusters. The disadvantage of this approach is that only approximate values are obtained for the greeks.

The constructed yield curve often looks smooth when inspecting the discount factors or the zero-coupon interest rates. When viewed in terms of the forward rates, on the other hand, the shortcomings of the construction method can become apparent.

The overnight interest rate spikes at certain dates because of low liquidity in the market. The reason is that banks seek increased liquidity for their balance sheets. For instance, this effect is often visible the last working day of the year. It does not only impact overnight rates, but any loan that extends over such a date. For example, the turn of year effect is visible for 3M EURIBORs with spot dates between October 1 and December 31. Needless to say, the longer the tenor, the smaller the impact. It is therefore important to include the effect of these spikes in the construction of the short end of the curve.

The spikes can be estimated from a historical time-series analysis together with a view of the future. Alternatively, by comparing an interest rate covering a low liquidity date with the interpolated result from neighboring interest rates, an estimate of the effect can be obtained. The curve construction should be done with the spikes removed to ensure a smooth curve. Once the interpolation has been done, the spikes need to be superimposed on the curve. One complication with modeling the spikes is that a study of historical time series (in EUR) show that the overnight interest rates do not immediately return to more normal values after a jump. Instead, the decay can take a couple of days. A consistent model therefore needs to include such a decay profile.

Disregarding the spikes, the overnight interest rate can jump at monetary policy meeting dates and then stays relatively constant between two such dates. It can, however, deviate substantially from being constant when there is a shortage of liquidity because of market stress. Furthermore, there is also the possibility of an interest rate change at an unscheduled meeting, but this happens rarely.

13.8 Yield Curve Modeling

For the rest of this chapter we focus on interest rate products that have payments at several points in time and cannot be priced with as simple models as in the previous sections, i.e. they depend on more than one rate. An example of such a product is a callable swap where one of the counterparties can terminate the deal at any of the tenor dates. If the termination is made at T_i, the resulting cash flows are identical to a swap with tenor dates $T_0, \dots, T_i$. It means that the price of the callable swap depends on the n swap rates with tenor dates $T_0, \dots, T_i$, $i = 1, \dots, n$. Furthermore, it also depends on their volatilities and correlations. It is clearly not possible to use the methods of the previous sections where only a single rate is modeled. Instead, a model for the full yield curve $\{P_{tT}\}_T$ is necessary. Models of this type are called *yield curve models*.

The volatility in a yield curve model is typically calibrated to caplets and swaptions. As we saw in Sect. 13.3, caplet pricing is best done in the forward measure, i.e. the measure corresponding to $P_{tT_{i+1}}$ as a numeraire. As the index i depends on the specific caplet, different measures are needed for the calibration of the different rates. When the calibration has been done, it is necessary to transform back to a common measure for which the pricing can be done. The technique of changing measures therefore takes a central role for yield curve models. This complication was avoided for commodities as we assumed deterministic interest rates. This is for obvious reasons not possible here.

There are four main types of yield curve models: short-rate models, HJM models, LMM models and Markov-functional models. Most publications cover these separately as if they were completely different model types. We use a different approach and highlight the connection between the models. This can be done by focusing on a particular model that has an interpretation within all these models types. The chosen model is then used as a point of reference through the remainder of the chapter.

A yield curve derivatives model should, of course, satisfy the usual demands e.g. be possible to calibrate to a rich set of quotes, have appropriate dynamics and be of high performance. Such models are often formulated in terms of the forward rates f_{tT}, $t \le T$ or the LIBOR rates $L_{tTT'}$, $t \le T < T'$, but can also be formulated with the zero-coupon bonds P_{tT}, $t \le T$. As the bonds are tradables (and not quotients of tradables), it is necessary to find an appropriate numeraire N_t and consider $\bar{P}_{tT} = P_{tT}/N_t$. Once $\bar{P}_{tT}$ has been computed through the model (for example by simulating it as a martingale), it is possible to obtain the value of the numeraire $N_t = \bar{P}_{tt}^{-1}$ and the zero-coupon bonds $P_{tT} = \bar{P}_{tT}/\bar{P}_{tt}$.

As all three variable types are popular to use for yield curve derivatives modeling, we derive the relations between them. To obtain a simple expression for the discounted bond prices in terms of the forward rates, we use the money market account $B_t = \exp(\int_0^t r_s ds) = \exp(\int_0^t f_{ss} ds)$ as a numeraire. We then obtain

$$\bar{P}_{tT} = \exp\left(-\int_t^T f_{ts} ds - \int_0^t f_{ss} ds\right)$$

The rates can be obtained from the discounted bond prices, or from each other, by

$$L_{tTT'} = \frac{1}{\delta}\left(\frac{P_{tT}}{P_{tT'}} - 1\right) = \frac{1}{\delta}\left(\frac{\bar{P}_{tT}}{\bar{P}_{tT'}} - 1\right) = \frac{1}{\delta}\left(\exp\left(\int_T^{T'} f_{ts} ds\right) - 1\right)$$

$$f_{tT} = -\frac{d}{dT} \ln P_{tT} = -\frac{d}{dT} \ln \bar{P}_{tT}$$

where $\delta = T' - T$. Thus, if one of $\{\bar{P}_{tT}\}_{t \leq T}$, $\{f_{tT}\}_{t \leq T}$ or $\{L_{tTT'}\}_{t \leq T < T'}$ is known, the other two can be computed.

13.9 The Gaussian Model

We focus on the model for which the forward rates satisfy a normal SDE

$$df_{tT} = \alpha_{tT}dt + \sigma_{tT}dW_t$$

under the risk-neutral measure, i.e. the martingale measure corresponding to the numeraire B_t. Our motivation for using this particular model is that it can be formulated within several of the most popular yield curve model types including HJM models, short-rate models and LMM models. We are then able to give an introduction to these model types with a specific example in mind.

Another reason for using this particular model is that it is analytically solvable, i.e. closed-form expressions exist for the future probability distributions. In fact, f_{tT} is Gaussian distributed which motivates the name Gaussian model. Analytical formulae are useful as they help us to understand the basics of yield curve modeling before moving on to more complex models. The advantages of using closed-form expressions should not be underestimated as even a simple model such as the Gaussian can be difficult to understand fully. The reason is that the products to be priced can have complicated payoff structures, which means that it might be necessary to use advanced numerical schemes even though the model is analytic. We therefore recommend developers of derivatives pricing software to initially implement an analytical model for which the results can be better understood and analyzed. The model can also guide us in the extension to more sophisticated models.

Analytically solvable models are often used because closed-form expressions lead to high-performing implementations. This type of models can also be used as a component in more advanced models. For instance, consider a product that depends on an FX rate but also has a weak dependence on the level of the rates in the two currencies. Because of performance, it can be a good idea to use an analytic model for the rates.

The empirical study by Rebonato and de Guillaume (2010) show that interest rates behave as lognormal when they are small enough (less than about 1.5%) or high enough (larger than about 5%) while they are Gaussian in between. The result is consistent across currencies and means that our choice of model is a good representation of reality in common market scenarios.

The shortcomings of the Gaussian model are that there is only a single driver, negative rates are supported and there is no flexibility in the skew and smile. Furthermore, the dynamics cannot be controlled. The use of a single driver means that options on the difference between two rates are mispriced. The existence of negative rates does usually not cause any problem for short-dated products as the probability of this happening is small. Long-dated products, on the other hand, can be mispriced. Fortunately, the Gaussian model can easily be extended or modified to more advanced models. We show how it is possible to only allow positive rates, how the skew and the smile can be controlled and how the dynamics can be changed. Just as for the commodity futures modeling in Sect. 12.3, it is possible to generalize to multiple drivers for decorrelation and to decompose the volatility in product form to boost the performance.

The LIBOR rates obey

$$(1 + (T' - T)L_{tTT'})(1 + (T'' - T')L_{tT'T''}) = (1 + (T'' - T)L_{tTT''})$$

This cocycle relation puts a restriction on the possible evolutions of the rates. This is similar to commodities for which there is a corresponding condition. We later show that the LIBOR rates follow

$$dL_{tTT'} = \frac{1}{T' - T}(1 + (T' - T)L_{tTT'})\eta_{tT} dW_t$$

in the Gaussian model. Using Ito's lemma on the cocycle relation, we see that if this type of dynamics is imposed on $L_{tTT'}$ and $L_{tT'T''}$, then $L_{tTT''}$ follows the same SDE, at least when the drift is disregarded. However, we show later in this chapter that the drift is unimportant as it is determined from a no-arbitrage condition and must be zero in the forward measure of the rate. The Gaussian model therefore has the attractive theoretical property that it is closed under the cocycle relation. In practical modeling it is not necessary to limit oneself to SDEs closed under the cocycle relation, but one should be aware of the pitfalls that can follow from violating the relation. It is straightforward to verify that the class of lognormal SDEs is an example of processes that are not closed under the cocycle relation.

13.10 Derivation of the Pricing Formula

We now solve the Gaussian model, i.e. we derive expressions for $\bar{P}_{tT}$, f_{tT} or $L_{tTT'}$. From the SDE, we immediately see that

$$f_{tT} = f_{0T} + \int_0^t \alpha_{sT} ds + \int_0^t \sigma_{sT} dW_s$$

from which it follows that

$$\begin{aligned}
\ln \bar{P}_{tT} &= -\int_t^T f_{ts} ds - \int_0^t f_{ss} ds \\
&= -\int_t^T \left(f_{0s} + \int_0^t \alpha_{us} du + \int_0^t \sigma_{us} dW_u \right) ds \\
&\quad - \int_0^t \left(f_{0s} + \int_0^s \alpha_{us} du + \int_0^s \sigma_{us} dW_u \right) ds
\end{aligned}$$

$$
\begin{aligned}
&= -\int_t^T f_{0s}ds - \int_0^t du \int_t^T ds\alpha_{us} - \int_0^t dW_u \int_t^T ds\sigma_{us} \\
&\quad - \int_0^t f_{0s}ds - \int_0^t du \int_u^t ds\alpha_{us} - \int_0^t dW_u \int_u^t ds\sigma_{us} \\
&= -\int_0^T f_{0s}ds - \int_0^t du \int_u^T ds\alpha_{us} - \int_0^t dW_u \int_u^T ds\sigma_{us} \\
&= \ln \bar{P}_{0T} + \int_0^t A_{sT}ds + \int_0^t \Psi_{sT}dW_s,
\end{aligned}
$$

$$
\begin{cases}
A_{tT} = -\int_t^T \alpha_{ts}ds \\
\Psi_{tT} = -\int_t^T \sigma_{ts}ds
\end{cases}
$$

$$
\Leftrightarrow \bar{P}_{tT} = \bar{P}_{0T} \exp\left(\int_0^t A_{sT}ds + \int_0^t \Psi_{sT}dW_s\right)
$$

and

$$
\begin{aligned}
L_{tTT'} &= \frac{1}{\delta}\left(\frac{\bar{P}_{tT}}{\bar{P}_{tT'}} - 1\right) \\
&= \frac{1}{\delta}\left((1+\delta L_{0TT'})\exp\left(\int_0^t (A_{sT} - A_{sT'})ds + \int_0^t (\Psi_{sT} - \Psi_{sT'})dW_s\right) - 1\right)
\end{aligned}
$$

From these formulae, a straightforward application of Ito's lemma gives the corresponding SDEs:

$$
\begin{aligned}
d\bar{P}_{tT} &= \bar{P}_{tT}\left(\left(A_{tT} + \frac{1}{2}\Psi_{tT}^2\right)dt + \Psi_{tT}dW_t\right) \\
dL_{tTT'} &= \frac{1}{\delta}(1+\delta L_{0TT'})\exp(...)\frac{d\exp(...)}{\exp(...)} \\
&= \frac{1}{\delta}(1+\delta L_{tTT'})\left(\left(A_{tT} - A_{tT'} + \frac{1}{2}(\Psi_{tT} - \Psi_{tT'})^2\right)dt + (\Psi_{tT} - \Psi_{tT'})dW_t\right)
\end{aligned}
$$

As $\bar{P}_{tT}$ is a quotient of a tradable and the numeraire, it must have zero drift. The drift and the volatility must therefore be related by

$$
A_{tT} + \frac{1}{2}\Psi_{tT}^2 = 0
$$

Thus, when constructing a model in terms of the forward rates, it is enough to specify either the drift or the volatility. The other parameter can be computed from the no-arbitrage relation. The standard approach is to specify the volatility and from that derive the drift. The reason is that it is often straightforward to calibrate the volatility to market quotes.

Observe that the computations do not depend on the Gaussian assumption, i.e. they hold when the drift and the volatility are allowed to depend on the forward rates. The relation between the drift and the volatility is therefore a general result. Thus, models based on forward rates incorporate the no-arbitrage condition in a particularly simple way and have as a consequence become popular. They are referred to as HJM models in honor of their discoverers (Heath et al. (1992)).

In terms of the original variables, the HJM condition takes the form

$$\alpha_{tT} = \sigma_{tT} \int_t^T \sigma_{ts} ds$$

Using the HJM condition, the equations can be rewritten as

$$\begin{cases} f_{tT} = f_{0T} + \int_0^t \sigma_{sT} \left(\int_s^T \sigma_{su} du \right) ds + \int_0^t \sigma_{sT} dW_s \\ \bar{P}_{tT} = \bar{P}_{0T} \exp\left(-\frac{1}{2} \int_0^t \Psi_{sT}^2 ds + \int_0^t \Psi_{sT} dW_s \right) \\ L_{tTT'} = \frac{1}{\delta} \Big((1 + \delta L_{0TT'}) \exp\Big(-\frac{1}{2} \int_0^t (\Psi_{sT}^2 - \Psi_{sT'}^2) ds \\ \qquad + \int_0^t (\Psi_{sT} - \Psi_{sT'}) dW_s \Big) - 1 \Big) \end{cases}$$

and the SDEs as

$$\begin{cases} df_{tT} = \sigma_{tT} \left(\int_t^T \sigma_{ts} ds \right) dt + \sigma_{tT} dW_t \\ d\bar{P}_{tT} = \bar{P}_{tT} \Psi_{tT} dW_t \\ dL_{tTT'} = \frac{1}{\delta} (1 + \delta L_{tTT'}) \left(\Psi_{tT'} \left(\Psi_{tT'} - \Psi_{tT} \right) dt + (\Psi_{tT} - \Psi_{tT'}) dW_t \right) \end{cases}$$

The equations are straightforward to solve as long as σ_{tT} (and therefore Ψ_{tT}) does not depend on the forward rates.

Denote $\int_0^t \Psi_{sT}^2 ds$ by ω_{tT} and observe that $W_{\omega_{tT}}$ describes the same process as $\int_0^t \Psi_{sT} dW_s$, see Sect. 5.7. The Gaussian model can then be formulated in terms of $\bar{P}_{tT}$ and $L_{tTT'}$ according to

$$\bar{P}_{tT} = \bar{P}_{0T} \exp\left(-\frac{1}{2}\omega_{tT} + W_{\omega_{tT}}\right)$$
$$L_{tTT'} = \frac{1}{\delta}\left((1+\delta L_{0TT'})\exp\left(-\frac{1}{2}(\omega_{tT}-\omega_{tT'}) + (W_{\omega_{tT}} - W_{\omega_{tT'}})\right) - 1\right)$$

We see that under the Gaussian model, the logarithm of the discounted bond price is nothing but a time-scaled Brownian motion with drift determined by the HJM condition. This is actually obvious as the starting point of the computations was a formulation of the forward rates that can be interpreted as a time-scaled Brownian motion with appropriate drift.

For an example of how the equations can be applied, we compute the futures rate, which according to Sect. 3.10 is equal to the expectation of $L_{TTT'}$ under the risk-neutral measure:

$$\begin{aligned} E\left[L_{TTT'}\right] &= \frac{1}{\delta}\left((1+\delta L_{0TT'})\exp\left(-\frac{1}{2}\int_0^T(\Psi_{sT}^2 - \Psi_{sT'}^2)ds\right)\right. \\ &\qquad \left. E\left[\exp\left(\int_0^T(\Psi_{sT}-\Psi_{sT'})dW_s\right)\right] - 1\right) \\ &= \frac{1}{\delta}\left((1+\delta L_{0TT'})\exp\left(\int_0^T(\Psi_{sT'}(\Psi_{sT'}-\Psi_{sT}))ds\right) - 1\right) \end{aligned}$$

Subtracting the forward rate $L_{0TT'}$ gives the well-known convexity adjustment between futures and FRAs:

$$\frac{1}{\delta}(1+\delta L_{0TT'})\left(\exp\left(\int_0^T(\Psi_{sT'}(\Psi_{sT'}-\Psi_{sT}))ds\right) - 1\right)$$

To obtain an expression that is easy to analyze, we assume a time-independent volatility $\sigma_{tT} = \sigma$. It gives $\Psi_{tT} = -(T-t)\sigma$ and the convexity adjustment

$$\begin{aligned} &\frac{1}{\delta}(1+\delta L_{0TT'})\left(\exp\left(\sigma^2\int_0^T(T'-s)(T'-T)ds\right) - 1\right) \\ &= \frac{1}{\delta}(1+\delta L_{0TT'})\left(\exp\left(\frac{1}{2}\sigma^2\delta T(T+2\delta)\right) - 1\right) \end{aligned}$$

which for small volatilities and interest rates can be approximated by

$$\frac{1}{2}\sigma^2 T(T+2\delta)$$

or

$$\frac{1}{2}\sigma_{\text{LN}}^2 L_{0TT'}^2 T(T+2\delta)$$

if using a lognormal volatility.

The convexity adjustment is small for low interest rates and the Gaussian model then gives a good lowest order approximation. For higher values of interest rates it is necessary to account for the non-perfect correlations within the yield curve. That can be achieved by adding more drivers (one more appears to be sufficient) to the SDE. Furthermore, in Piterbarg and Renedo (2004) it was found that the impact of the skew and smile on the convexity adjustment cannot be disregarded.

13.11 Change of Measure

Observe that the equations of the previous section look simpler for $\bar{P}_{tT}$ than for f_{tT} and $L_{tTT'}$. This should come as no surprise since $\bar{P}_{tT}$ is a quotient of a tradable P_{tT} and the numeraire B_t and must be a martingale. It therefore makes sense to work with $\bar{P}_{tT}$ if formulating the model in the risk-neutral measure. Unfortunately, it is not straightforward to calibrate the model in this measure.

The calibration of the volatility information Ψ_{tT} should be done to match the most liquid volatility-dependent products in the market, which are the caps and swaptions. We restrict ourselves to the calibration to caplets; the calibration to swaptions can then be done as in Sect. 13.20. Recall that caplet pricing is simplest in the forward measure of the underlying LIBOR rate. In this measure, the LIBOR rate is a martingale and has a particularly simple expression. However, as we now show, $\bar{P}_{tT}$ looks more complicated as it is no longer a martingale. Observe that a yield curve model contains several LIBOR rates $L_{tTT'}$ and each of them is typically calibrated in its own natural measure with numeraire being the zero coupon bond maturing at T'. A model is therefore usually calibrated in several different measures. Once a model has been calibrated, it has to be reformulated in the pricing measure, which could, for instance, be the risk-neutral measure. This measure is chosen to obtain pricing computations that are as simple as possible.

We have formulated the Gaussian model in the risk-neutral measure and for calibration purposes we now show how the model looks when working in an arbitrary forward measure with maturity $\check{T}$. Choosing $P_{t\check{T}}$ as a numeraire, it follows from the Appendix that the Radon-Nikodym derivative is given by $M_t = B_0 P_{t\check{T}} / B_t P_{0\check{T}} = \bar{P}_{t\check{T}} / \bar{P}_{0\check{T}}$. Using the results of the previous section, we obtain

$$M_t = \exp\left(-\frac{1}{2}\int_0^t \Psi^2_{s\check{T}} ds + \int_0^t \Psi_{s\check{T}} d W_s\right)$$

Girsanov's theorem implies that

$$W_t^{\check{T}} = W_t - \int_0^t \Psi_{s\check{T}} ds$$

is a Brownian motion in the $\check{T}$-forward measure. The results of the previous section can then be expressed in the $\check{T}$-forward measure, i.e. with respect to the Brownian motion $W^{\check{T}}$:

$$\left\{\begin{aligned} f_{tT} &= f_{0T} + \int_0^t \sigma_{sT}\left(\int_{\check{T}}^{T} \sigma_{su} du\right) ds + \int_0^t \sigma_{sT}\, dW_s^{\check{T}} \\ \bar{P}_{tT} &= \bar{P}_{0T} \exp\left(-\frac{1}{2}\int_0^t \Psi_{sT}(\Psi_{sT} - 2\Psi_{s\check{T}}) ds + \int_0^t \Psi_{sT}\, dW_s^{\check{T}}\right) \\ L_{tTT'} &= \frac{1}{\delta}\Bigg((1 + \delta L_{0TT'}) \exp\Bigg(-\frac{1}{2}\int_0^t (\Psi_{sT'} - \Psi_{sT}) \\ &\qquad (2\Psi_{s\check{T}} - \Psi_{sT'} - \Psi_{sT}) ds + \int_0^t (\Psi_{sT} - \Psi_{sT'}) dW_s^{\check{T}}\Bigg) - 1\Bigg) \end{aligned}\right.$$

The corresponding SDEs are

$$\left\{\begin{aligned} df_{tT} &= \sigma_{tT}\left(\int_{\check{T}}^{T} \sigma_{ts} ds\right) dt + \sigma_{tT}\, dW_t^{\check{T}} \\ d\bar{P}_{tT} &= \bar{P}_{tT}(\Psi_{tT}\Psi_{t\check{T}} dt + \Psi_{tT}\, dW_t^{\check{T}}) \\ dL_{tTT'} &= \frac{1}{\delta}(1 + \delta L_{tTT'})\big((\Psi_{tT'} - \Psi_{tT})(\Psi_{tT'} - \Psi_{t\check{T}}) dt \\ &\quad + (\Psi_{tT} - \Psi_{tT'}) dW_t^{\check{T}}\big) \end{aligned}\right.$$

As expected, the choice $\check{T} = T$ makes f_{tT} a martingale and is particularly simple to handle. This follows as f_{tT} can be obtained from the quotient $(P_{t,T-\delta} - P_{tT})/\delta P_{tT}$ in the limit $\delta \to 0$, where the numeraire is in the denominator. Similarly, the expression for $L_{tTT'}$ simplifies for $\check{T} = T'$. For $\bar{P}_{tT}$, on the other hand, there is no choice of $\check{T}$ that simplifies the expression as it is not a martingale in any forward measure.

These measure transformations make it possible to calibrate the model. We discuss the details in Sect. 13.20 and from now on we assume that the model is calibrated, meaning that $\{\Psi_{tT}\}$ are known. We instead focus on the properties of the model and on its generalizations.

13.12 Local Volatility

Recall that the Gaussian model supports negative rates and lacks control of the skew and smile. These problems can be avoided by allowing the volatility to be state dependent, i.e. to depend on the rates. For instance, it is popular to assume lognormality,

$$df_{tT} = \sigma_{tT} f_{tT} \left(\int_{\check{T}}^{T} \sigma_{ts} f_{ts} ds \right) dt + \sigma_{tT} f_{tT} dW_t^{\check{T}}$$

in the $\check{T}$-forward measure. This SDE looks particularly simple in the natural measure of f_{tT} where $\check{T} = T$. But this is not the situation for arbitrary $\check{T}$, and as we discussed earlier, it is often necessary to work in different measures for yield curve models. If f_{tT} needs to be simulated between two times t_i and t_{i+1} under the lognormal model, the rates $\{f_{tT'}\}_{\check{T} \leq T' \leq T}$ are needed. As the rates are stochastic, the evaluation from one time step t_i to the next t_{i+1} has to be done by an approximate scheme such as the predictor-corrector method. This introduces numerical errors in the simulations and it is no longer be possible to take large simulation steps, as can be done for the Gaussian model. This has a significant effect on the performance.

Just as for commodities, the dimensionality of the implementation can be reduced by assuming a separable volatility, see Sect. 13.15 for a further discussion. Unfortunately, this approach is not possible if the volatility depends on the rate f_{tT}. One way to circumvent the problem is to use a method proposed in Cheyette (1992) and Cheyette (1996). To illustrate the technique, assume a separable volatility $\sigma_{tT} = \sigma_t \Lambda_T$. Let T' be the payment date of the last rate $f_{tT'}$ used in the pricing and assume that f_{tT} depends locally on this rate instead of on itself:

$$df_{tT} = \Lambda_T f_{tT'}^2 \sigma_t^2 \left(\int_{\check{T}}^{T} \Lambda_s ds \right) dt + \Lambda_T f_{tT'} \sigma_t dW_t^{\check{T}}$$

When the pricing is done in the *terminal measure*, i.e. the forward measure of the last rate, we obtain

$$f_{tT} = f_{0T} + \Lambda_T \left(\int_{T'}^{T} \Lambda_u du \right) \int_0^t f_{sT'}^2 \sigma_s^2 ds + \Lambda_T \int_0^t f_{sT'} \sigma_s dW_s^{T'}$$

We see that all the forward rates can be computed by only storing the two stochastic variables $\int_0^t f_{sT'} \sigma_s dW_s^{T'}$ and $\int_0^t f_{sT'}^2 \sigma_s^2 ds$. By setting $\check{T} = T$ in the SDE, however, we see that the calibration of Cheyette type of models is not as straightforward as, for example, the Gaussian model or the model defined from a lognormal SDE.

13.13 Stochastic Volatility

The skew and smile can be controlled through stochastic volatility. We consider the following set of SDEs:

$$\begin{cases} df_{tT} = \sigma_{tT} \left(\int_t^T \sigma_{ts} ds \right) dt + \sigma_{tT} dW_t \\ d\sigma_{tT} = \gamma_{tT} dt + \epsilon_{tT} dZ_t \end{cases}$$

where γ_{tT} and ϵ_{tT} are allowed to depend on σ_{tT}. The Radon-Nikodym derivative $M_t = \bar{P}_{tT}/\bar{P}_{0T}$ for the change to the T-forward measure satisfies

$$dM_t = M_t \Psi_{tT} dW_t$$

Girsanov's theorem implies that

$$\begin{cases} W_t^T = W_t - \displaystyle\int_0^t \Psi_{sT} ds \\ Z_t^T = Z_t - \rho \displaystyle\int_0^t \Psi_{sT} ds \end{cases}$$

are Brownian motions in the T-forward measure, where ρ is the correlation between W and Z. The SDEs in the forward measure take the form

$$\begin{cases} df_{tT} = \sigma_{tT} dW_t^T \\ d\sigma_{tT} = (\gamma_{tT} + \rho\epsilon_{tT}\Psi_{tT})dt + \epsilon_{tT} dZ_t^T \end{cases}$$

We would like to investigate under which assumptions we can retain the attractive features of the Gaussian model, namely that of analytic solvability and a low-dimensional implementation.

Observe that the volatility SDE depends on $\{\sigma_{ts}\}_{t<s<T}$ through Ψ_{tT}, which means that a high-dimensional implementation is needed. This problem can be avoided by setting $\rho = 0$. Whether this is a reasonable assumption depends on the reason why the correlation was included in the model: was it to match the implied skew or to obtain realistic dynamics? In the former case, many interest rate markets are observed to only have a weak skew meaning that the assumption can be acceptable. In the latter case, we note that there is no strong fundamental reason why interest rates should be correlated with their volatility. This is in contrast to equities and commodities for which such reasons can be found. In fact, studies of historical data (Chen and Scott (2001)) reveal that the correlation is indistinguishable from zero.

With the correlation set to zero, the volatility SDE is unaffected by transformations between the risk-neutral measure and the forward measures. Furthermore, unlike the SDE for the underlying, there does not exist any no-arbitrage condition for the volatility SDE. It means that we have no restrictions in our choice of either ϵ_{tT} or γ_{tT}. As the SDE for f_{tT} is analytic, it makes sense to choose ϵ_{tT} and γ_{tT} so that the volatility SDE becomes analytically solvable as well.

Recall that a low-dimension tree implementation of the underlying process is possible if the volatility is of product form. By the same token, the volatility of volatility must be of product form $\epsilon_{tT} = \eta_T \epsilon_t$ for a low-dimensional implementation of a stochastic volatility model. No such restriction exists for γ_{tT}.

In Sect. 13.15 we argue for the obvious fact that a single driver for the underlying is insufficient for accurate pricing of correlation-dependent products. Instead, an extension to multiple drivers of the underlying is necessary. It is then an interesting question whether it is enough to have a single driver for the volatility or if this process also has to be decorrelated? According to Piterbarg (2005), the volatility decorrelation effect is important and needs to be accounted for.

As we argued above, the form of the SDE for the underlying has been fixed to obtain analytical properties and the only remaining degrees of freedom live in the volatility SDE. Thus, it is the volatility process that needs to be calibrated to the caplets and swaptions skew and smile. The details are somewhat similar to those of Dupire in Sect. 6.2, but for the volatility instead of the underlying. For the sake of argument, we focus on calibration to caplets and use the SDE for a LIBOR rate in the forward measure

$$dL_{tTT'} = \frac{1}{\delta}(1 + \delta L_{tTT'})\sigma_t dW_t^{T'}$$

where we have used the notation $\sigma_t = \Psi_{tT} - \Psi_{tT'}$ for the volatility. For deterministic volatility, we follow the logic in Sect. 13.3 to arrive at an expression for the caplet price:

$$C(t) = \delta P_{tT'}\left(\left(L_{tTT'} + \frac{1}{\delta}\right)N(d_+) - \left(K + \frac{1}{\delta}\right)N(d_-)\right)$$

$$d_\pm = \frac{\ln\left(\left(L_{tTT'} + \frac{1}{\delta}\right) / \left(K + \frac{1}{\delta}\right)\right)}{\sqrt{\int_t^T \sigma_u^2 du}} \pm \frac{1}{2}\sqrt{\int_t^T \sigma_u^2 du}$$

If an uncorrelated volatility process is included, the caplet price can be computed by conditional expectation. The result is

$$C(t) = \delta P_{tT'} E\left[\left(L_{tTT'} + \frac{1}{\delta}\right)N(d_+) - \left(K + \frac{1}{\delta}\right)N(d_-)\right]$$

where the expectation is over the volatility process. It appears difficult to infer the volatility distribution from this expression. Instead, we consider payoff functions $f(L_{TT'})$ for which it is easier to back out the distribution. It is then possible to use the fact that the present value of products with arbitrary fixed time payoffs can be obtained from static replication of caplets.

Using ideas similar to those in Carr and Lee (2008), we find that shifted power functions

$$f(L_{TT'}) = \left(\frac{L_{TT'} + \frac{1}{\delta}}{L_{tTT'} + \frac{1}{\delta}}\right)^q$$

are particularly suitable for our purpose. It is straightforward to compute the price for deterministic volatility:

$$V(t) = \delta P_{tT'} \exp\left(\frac{1}{2}(q^2 - q)\int_t^T \sigma_u^2 du\right)$$

With $v_T = \int_t^T \sigma_u^2 du$ it follows that the stochastic-volatility price is given by

$$V(t) = \delta P_{tT'} \int_0^\infty \exp\left(\frac{1}{2}(q^2 - q)v_T\right) \rho(v_T) dv_T$$

It means that the Laplace transformation of the variance probability density function gives the prices of the shifted power options. Through Laplace inversion it is then possible to back out the PDF for the variance and in that way calibrating the volatility process.

13.14 Inclusion of Jumps

We extend the computations of Sects. 13.10 and 13.11 to include jumps. For this purpose, we consider the SDE

$$df_{tT} = \alpha_{tT} dt + \sigma_{tT} dW_t + \int_{\mathbb{R}} \delta_{tTx} \tilde{\mu}(dt, dx)$$

where we for simplicity have assumed appropriate regularity conditions to be fulfilled to avoid the cutoff in the final term. The solution is given by

$$f_{tT} = f_{0T} + \int_0^t \alpha_{sT} ds + \int_0^t \sigma_{sT} dW_s + \int_0^t \int_{\mathbb{R}} \delta_{sTx} \tilde{\mu}(ds, dx)$$

The discounted bond prices can be computed in a similar way as when no jumps were present:

$$\begin{aligned}
\ln \bar{P}_{tT} &= -\int_t^T f_{ts} ds - \int_0^t f_{ss} ds \\
&= -\int_t^T \left(f_{0s} + \int_0^t \alpha_{us} du + \int_0^t \sigma_{us} dW_u + \int_0^t \int_{\mathbb{R}} \delta_{usx} \tilde{\mu}(du, dx) \right) ds \\
&\quad - \int_0^t \left(f_{0s} + \int_0^s \alpha_{us} du + \int_0^s \sigma_{us} dW_u + \int_0^s \int_{\mathbb{R}} \delta_{usx} \tilde{\mu}(du, dx) \right) ds \\
&= -\int_t^T f_{0s} ds - \int_0^t du \int_t^T ds \alpha_{us} - \int_0^t dW_u \int_t^T ds \sigma_{us} \\
&\quad - \int_0^t \int_t^T ds \int_{\mathbb{R}} \delta_{usx} \tilde{\mu}(du, dx) \\
&\quad - \int_0^t f_{0s} ds - \int_0^t du \int_u^t ds \alpha_{us} - \int_0^t dW_u \int_u^t ds \sigma_{us} \\
&\quad - \int_0^t \int_u^t ds \int_{\mathbb{R}} \delta_{usx} \tilde{\mu}(du, dx)
\end{aligned}$$

$$
\begin{aligned}
&= -\int_0^T f_{0s} ds - \int_0^t du \int_u^T ds \alpha_{us} - \int_0^t dW_u \int_u^T ds \sigma_{us} \\
&\quad - \int_0^t \int_u^T ds \int_{\mathbb{R}} \delta_{usx} \tilde{\mu}(du, dx) \\
&= \ln \bar{P}_{0T} + \int_0^t A_{sT} ds + \int_0^t \Psi_{sT} dW_s + \int_0^t \int_{\mathbb{R}} D_{sTx} \tilde{\mu}(ds, dx),
\end{aligned}
$$

$$
\begin{cases}
A_{tT} = -\int_t^T \alpha_{ts} ds \\
\Psi_{tT} = -\int_t^T \sigma_{ts} ds \\
D_{tTx} = -\int_t^T \delta_{tsx} ds
\end{cases}
$$

$$
\Leftrightarrow \bar{P}_{tT} = \bar{P}_{0T} \exp\left(\int_0^t A_{sT} ds + \int_0^t \Psi_{sT} dW_s + \int_0^t \int_{\mathbb{R}} D_{sTx} \tilde{\mu}(ds, dx)\right)
$$

which gives the following expression for the LIBOR rates:

$$
\begin{aligned}
L_{tTT'} &= \frac{1}{\delta}\left(\frac{\bar{P}_{tT}}{\bar{P}_{tT'}} - 1\right) \\
&= \frac{1}{\delta}\Bigg((1 + \delta L_{0TT'}) \exp\Bigg(\int_0^t (A_{sT} - A_{sT'}) ds + \int_0^t (\Psi_{sT} - \Psi_{sT'}) dW_s \\
&\qquad + \int_0^t \int_{\mathbb{R}} (D_{sTx} - D_{sT'x}) \tilde{\mu}(ds, dx)\Bigg) - 1\Bigg)
\end{aligned}
$$

A straightforward application of Ito's lemma gives the SDEs

$$
\begin{aligned}
d\bar{P}_{tT} &= \bar{P}_{tT}\Bigg(\left(A_{tT} + \frac{1}{2}\Psi_{tT}^2\right) dt + \int_{\mathbb{R}} (e^{D_{tTx}} - 1 - D_{tTx}) \nu(dt, dx) \\
&\qquad + \Psi_{tT} dW_t + \int_{\mathbb{R}} (e^{D_{tTx}} - 1) \tilde{\mu}(dt, dx)\Bigg) \\
dL_{tTT'} &= \frac{1}{\delta}(1 + \delta L_{0TT'}) \exp(\cdots) \frac{d \exp(\cdots)}{\exp(\cdots)} \\
&= \frac{1}{\delta}(1 + \delta L_{tTT'})\Bigg(\left(A_{tT} - A_{tT'} + \frac{1}{2}(\Psi_{tT} - \Psi_{tT'})^2\right) dt \\
&\qquad + \int_{\mathbb{R}} (e^{D_{tTx} - D_{tT'x}} - 1 - D_{tTx} + D_{tT'x}) \nu(dt, dx) \\
&\qquad + (\Psi_{tT} - \Psi_{tT'}) dW_t + \int_{\mathbb{R}} (e^{D_{tTx} - D_{tT'x}} - 1) \tilde{\mu}(dt, dx)\Bigg)
\end{aligned}
$$

The HJM condition reads

$$A_{tT} + \frac{1}{2}\Psi_{tT}^2 + \int_{\mathbb{R}} (e^{D_{tTx}} - 1 - D_{tTx})\nu(\cdot, dx) = 0$$

where $\nu(\cdot, dx) = \nu(dt, dx)/dt$.

Writing the HJM condition as

$$\alpha_{tT} = \sigma_{tT} \int_t^T \sigma_{ts} ds + \int_{\mathbb{R}} \delta_{tTx}(1 - e^{-\int_t^T \delta_{tsx} ds})\nu(\cdot, dx)$$

the equations and SDEs can be summarized as

$$\begin{cases} \begin{aligned} f_{tT} &= f_{0T} + \int_0^t \sigma_{sT} \left(\int_s^T \sigma_{su} du \right) ds + \int_0^t \int_{\mathbb{R}} \delta_{sTx}(1 - e^{-\int_s^T \delta_{sux} du})\nu(ds, dx) \\ &\quad + \int_0^t \sigma_{sT} dW_s + \int_0^t \int_{\mathbb{R}} \delta_{sTx} \tilde{\mu}(ds, dx) \\ \bar{P}_{tT} &= \bar{P}_{0T} \exp\Bigg(-\frac{1}{2} \int_0^t \Psi_{sT}^2 ds - \int_0^t \int_{\mathbb{R}} (e^{D_{sTx}} - 1 - D_{sTx})\nu(ds, dx) \\ &\quad + \int_0^t \Psi_{sT} dW_s + \int_0^t \int_{\mathbb{R}} D_{sTx} \tilde{\mu}(ds, dx) \Bigg) \\ L_{tTT'} &= \frac{1}{\delta} \Bigg((1 + \delta L_{0TT'}) \exp\Bigg(-\frac{1}{2} \int_0^t (\Psi_{sT}^2 - \Psi_{sT'}^2) ds \\ &\quad - \int_0^t \int_{\mathbb{R}} (e^{D_{sTx}} - e^{D_{sT'x}} - D_{sTx} + D_{sT'x})\nu(ds, dx) \\ &\quad + \int_0^t (\Psi_{sT} - \Psi_{sT'}) dW_s + \int_0^t \int_{\mathbb{R}} (D_{sTx} - D_{sT'x}) \tilde{\mu}(ds, dx) \Bigg) - 1 \Bigg) \end{aligned} \end{cases}$$

$$\begin{cases} \begin{aligned} df_{tT} &= \sigma_{tT} \left(\int_t^T \sigma_{ts} ds \right) dt + \textstyle\int_{\mathbb{R}} \delta_{tTx}(1 - e^{-\int_t^T \delta_{tsx} ds})\nu(dt, dx) \\ &\quad + \sigma_{tT} dW_t + \int_{\mathbb{R}} \delta_{tTx} \tilde{\mu}(dt, dx) \\ d\bar{P}_{tT} &= \bar{P}_{tT}(\Psi_{tT} dW_t + \int_{\mathbb{R}} (e^{D_{tTx}} - 1)\tilde{\mu}(dt, dx)) \\ dL_{tTT'} &= \frac{1}{\delta}(1 + \delta L_{tTT'}) \Bigg(\Psi_{tT'}(\Psi_{tT'} - \Psi_{tT}) dt \\ &\quad + \int_{\mathbb{R}} \left(e^{D_{tTx} - D_{tT'x}} - 1 - e^{D_{tTx}} + e^{D_{tT'x}} \right) \nu(dt, dx) \\ &\quad + (\Psi_{tT} - \Psi_{tT'}) dW_t + \int_{\mathbb{R}} (e^{D_{tTx} - D_{tT'x}} - 1)\tilde{\mu}(dt, dx) \Bigg) \end{aligned} \end{cases}$$

The next step is to investigate how the equations look in the $\check{T}$-forward measure. The Radon-Nikodym derivative $M_t = \bar{P}_{t\check{T}}/\bar{P}_{0\check{T}}$ satisfies

$$dM_t = M_t \left(\Psi_{t\check{T}} dW_t + \int_{\mathbb{R}} (e^{D_{t\check{T}x}} - 1)\tilde{\mu}(dt, dx) \right)$$

It follows from Sect. 8.3 that

$$W_t^{\check{T}} = W_t - \int_0^t \Psi_{s\check{T}} ds$$

is a Brownian motion in the $\check{T}$-forward measure and that

$$\nu^{\check{T}}(dt, dx) = e^{D_{t\check{T}x}} \nu(dt, dx)$$

is the Lévy measure. We obtain the following equations and SDEs in the $\check{T}$-forward measure:

$$\begin{cases}
\begin{aligned}
f_{tT} &= f_{0T} + \int_0^t \sigma_{sT} \left(\int_{\check{T}}^T \sigma_{su} du \right) ds \\
&\quad + \int_0^t \int_{\mathbb{R}} \delta_{sTx} (1 - e^{-\int_{\check{T}}^T \delta_{sux} du}) \nu^{\check{T}}(ds, dx) \\
&\quad + \int_0^t \sigma_{sT} dW_s^{\check{T}} + \int_0^t \int_{\mathbb{R}} \delta_{sTx} \tilde{\mu}^{\check{T}}(ds, dx) \\
\bar{P}_{tT} &= \bar{P}_{0T} \exp\left(-\frac{1}{2} \int_0^t \Psi_{sT}(\Psi_{sT} - 2\Psi_{s\check{T}}) ds \right. \\
&\quad - \int_0^t \int_{\mathbb{R}} (e^{D_{sTx} - D_{s\check{T}x}} - e^{D_{s\check{T}x}} + D_{sTx}) \nu^{\check{T}}(ds, dx) \\
&\quad \left. + \int_0^t \Psi_{sT} dW_s^{\check{T}} + \int_0^t \int_{\mathbb{R}} D_{sTx} \tilde{\mu}^{\check{T}}(ds, dx) \right) \\
L_{tTT'} &= \frac{1}{\delta} \left((1 + \delta L_{0TT'}) \exp\left(-\frac{1}{2} \int_0^t (\Psi_{sT'} - \Psi_{sT})(2\Psi_{s\check{T}} - \Psi_{sT'} - \Psi_{sT}) ds \right. \right. \\
&\quad - \int_0^t \int_{\mathbb{R}} (e^{-D_{sT'x}}(e^{D_{sTx}} - e^{D_{sT'x}}) - D_{sTx} + D_{sT'x}) \nu^{\check{T}}(ds, dx) \\
&\quad \left. \left. + \int_0^t (\Psi_{sT} - \Psi_{sT'}) dW_s^{\check{T}} + \int_0^t \int_{\mathbb{R}} (D_{sTx} - D_{sT'x}) \tilde{\mu}^{\check{T}}(ds, dx) \right) - 1 \right)
\end{aligned}
\end{cases}$$

$$\begin{cases} df_{tT} = \sigma_{tT}\left(\int_{\check{T}}^{T} \sigma_{ts} ds\right) dt + \int_{\mathbb{R}} \delta_{tTx}(1 - e^{-\int_{\check{T}}^{T} \delta_{tsx} ds})\nu^{\check{T}}(dt, dx) \\ \qquad + \sigma_{tT} dW_t^{\check{T}} + \int_{\mathbb{R}} \delta_{tTx} \tilde{\mu}^{\check{T}}(dt, dx) \\ d\bar{P}_{tT} = \bar{P}_{tT}\left(\Psi_{tT}\Psi_{t\check{T}} dt - \int_{\mathbb{R}} (e^{D_{tTx} - D_{t\check{T}x}} - e^{D_{t\check{T}x}} - e^{D_{tTx}} + 1)\right. \\ \qquad \left. + \Psi_{tT} dW_t^{\check{T}} + \int_{\mathbb{R}} (e^{D_{tTx}} - 1)\tilde{\mu}^{\check{T}}(dt, dx)\right) \\ dL_{tTT'} = \frac{1}{\delta}(1 + \delta L_{tTT'})\Bigg((\Psi_{tT'} - \Psi_{tT})(\Psi_{tT'} - \Psi_{t\check{T}})dt \\ \qquad + \int_{\mathbb{R}} \left(e^{D_{tT'x} - D_{t\check{T}x}} - e^{D_{tTx} - D_{t\check{T}x}} + e^{D_{tTx} - D_{tT'x}} - 1\right) \nu^{\check{T}}(dt, dx) \\ \qquad + (\Psi_{tT} - \Psi_{tT'}) dW_t^{\check{T}} + \int_{\mathbb{R}} (e^{D_{tTx} - D_{tT'x}} - 1)\tilde{\mu}^{\check{T}}(dt, dx)\Bigg) \end{cases}$$

As before, the SDE for f_{tT} simplifies for $\check{T} = T$ and the SDE for $L_{tTT'}$ simplifies for $\check{T} = T'$.

13.15 The Hull-White Model

The Gaussian model has been formulated in terms of σ_{tT} for the forward rates and in terms of ω_{tT} for the discounted bond prices and the LIBOR rates. As σ_{tT} and ω_{tT} are 2-dimensional functions of time, it is popular to reduce them to a 1-dimensional dependence.

We consider the performance-boosting assumption that the volatility is separable: $\sigma_{tT} = \Lambda_T \sigma_t$. We obtain

$$f_{tT} = f_{0T} + \Lambda_T \int_0^t \sigma_s^2 \left(\int_s^T \Lambda_u du\right) ds + \Lambda_T \int_0^t \sigma_s dW_s$$

and

$$\begin{aligned} dr_t = f_{t+dt,t+dt} - f_{tt} &= f_{0,t+dt} + \Lambda_{t+dt} \int_0^{t+dt} \sigma_s^2 \left(\int_s^{t+dt} \Lambda_u du\right) ds \\ &\quad + \Lambda_{t+dt} \int_0^{t+dt} \sigma_s dW_s - f_{0t} - \Lambda_t \int_0^t \sigma_s^2 \left(\int_s^t \Lambda_u du\right) ds - \Lambda_t \int_0^t \sigma_s dW_s \\ &= \left(\partial_t f_{0t} + \frac{\partial_t \Lambda_t}{\Lambda_t}(f_{tt} - f_{0t}) + \Lambda_t^2 \int_0^t \sigma_s^2 ds\right) dt + \Lambda_t \sigma_t dW_t \\ &= \lambda_t(\tilde{r}_t - r_t)dt + \tilde{\sigma}_t dW_t \end{aligned}$$

which means that the short rate follows an Ornstein-Uhlenbeck process with volatility $\tilde{\sigma}_t = \Lambda_t \sigma_t$, mean-reversion factor $\lambda_t = -(\partial_t \Lambda_t)/\Lambda_t$ and mean-reversion level

$$\tilde{r}_t = f_{0t} + \partial_t f_{0t}/\lambda_t + \frac{\Lambda_t^2}{\lambda_t}\int_0^t \sigma_s^2 ds$$

A model for which the short rate follows an Ornstein-Uhlenbeck process is said to be a Hull-White model (Hull and White (1990)). Thus, we conclude that normally distributed forward rates with separable volatilities $\sigma_{tT} = \Lambda_T \sigma_t$ give a Hull-White model.

To show that the two model types are equivalent, let us now start with the Hull-White model

$$dr_r = \lambda_t(\tilde{r}_t - r_t)dt + \tilde{\sigma}_t dW_t$$

and derive forward rates that are normally distributed with separable volatilities. The pricing is straightforward as the money market account $B_t = \exp\left(\int_0^t r_s ds\right)$ is expressed in terms of the short rate and

$$\bar{P}_{tT} = E_t[B_T^{-1}]$$

To simplify the computations, we only compute the stochastic part of $\bar{P}_{tT}$. The reason is that the deterministic part can be computed by using the fact that $\bar{P}_{tT}$ is a martingale. Letting "$\sim$" denote equality up to a deterministic part, we have

$$r_t \sim \Lambda_t \int_0^t \sigma_s dW_s, \quad \Lambda_t = e^{-\int_0^t \lambda_s ds}, \quad \sigma_t = \Lambda_t^{-1}\tilde{\sigma}_t$$

according to Sect. 5.6. With $\phi_t = -\int_0^t \Lambda_s ds$, we obtain

$$\begin{aligned}
\bar{P}_{tT} &\sim E_t\left[\exp\left(-\int_0^T \Lambda_s\left(\int_0^s \sigma_u dW_u\right)ds\right)\right] \\
&= E_t\left[\exp\left(-\int_0^T \sigma_u\left(\int_u^T \Lambda_s ds\right)dW_u\right)\right] \\
&= E_t\left[\exp\left(\int_0^T \sigma_u(\phi_T - \phi_u)dW_u\right)\right] \\
&\sim E_t\left[\exp\left(\int_0^t \sigma_u(\phi_T - \phi_u)dW_u\right)\right] \\
&= \exp\left(\int_0^t \sigma_u(\phi_T - \phi_u)dW_u\right)
\end{aligned}$$

which gives

$$\bar{P}_{tT} = \bar{P}_{0T} \exp\left(-\frac{1}{2}\int_0^t \sigma_s^2(\phi_T - \phi_s)^2 ds + \int_0^t \sigma_s(\phi_T - \phi_s) dW_s\right)$$

To simplify the expression, we introduce a new measure Q by

$$\left(\frac{dQ}{dP}\right)_t = M_t = \exp\left(-\frac{1}{2}\int_0^t \sigma_s^2 \phi_s^2 ds - \int_0^t \sigma_s \phi_s dW_s\right)$$

with corresponding numeraire $D_t = M_t B_t$. It gives

$$D_t^{-1} P_{tT} = P_{0T} \exp\left(-\frac{1}{2}\phi_T^2 \int_0^t \sigma_s^2 ds + \phi_T \int_0^t \sigma_s^2 \phi_s ds + \phi_T \int_0^t \sigma_s dW_s\right)$$

which can be written as

$$D_t^{-1} P_{tT} = P_{0T} \exp\left(-\frac{1}{2}\phi_T^2 \int_0^t \sigma_s^2 ds + \phi_T \int_0^t \sigma_s d\check{W}_s\right)$$

in terms of the Brownian motion $\check{W}_t = W_t + \int_0^t \sigma_s \phi_s ds$ in the measure Q. Finally, with $\eta_t = \int_0^t \sigma_s^2 ds$, we obtain

$$D_t^{-1} P_{tT} = P_{0T} \exp\left(-\frac{1}{2}\phi_T^2 \eta_t + \phi_T \check{W}_{\eta_t}\right)$$

The pricing equation for the Hull-White model is similar to what would have obtained from the Gaussian model if letting ω_{tT} be of product form $\phi_T^2 \eta_t$:

$$B_t^{-1} P_{tT} = P_{0T} \exp\left(-\frac{1}{2}\phi_T^2 \eta_t + \phi_T W_{\eta_t}\right)$$

The difference is that the pricing equations are formulated in different measures. The forward rates can be obtained from

$$\begin{aligned} f_{tT} &= -\frac{d}{dT} \ln P_{tT} = -\frac{d}{dT} \ln B_t^{-1} P_{tT} \\ &= f_{0T} + (\partial_T \phi_T) \int_0^t \sigma_s^2 (\phi_T - \phi_s) ds - (\partial_T \phi_T) \int_0^t \sigma_s dW_s \\ &= f_{0T} + \Lambda_T \int_0^t \sigma_s^2 \left(\int_s^T \Lambda_u du\right) ds + \Lambda_T \int_0^t \sigma_s dW_s \end{aligned}$$

which is identical to the starting equation and proves that the Hull-White model indeed is equivalent to the Gaussian model with separable volatility.

Let us briefly discuss the calibration of the model. As mentioned before, the calibration of the volatilities can be done, for example, by formulating the model in terms of the LIBOR rates. It is also necessary to calibrate the model to the yield curve. This calibration is straightforward if using the formulae for f_{tT}, $\bar{P}_{tT}$ and $L_{tTT'}$ in Sect. 13.10 as they contain the information f_{0T}, $\bar{P}_{0T}$ and $L_{0TT'}$ about today's yield curve. The calibration is actually also simple when the drift and the volatility are allowed to depend on the rate, which means that explicit expressions for f_{tT}, $\bar{P}_{tT}$ and $L_{tTT'}$ might not be found. The reason is that the model can be formulated in terms of SDEs and the current yield curve is the initial condition for these equations. We therefore see that an advantage of formulating a yield curve model as a stochastic model for f_{tT}, $\bar{P}_{tT}$ or $L_{tTT'}$ is that the calibration to today's yield curve is automatic. The calibration to the yield curve for *short-rate models*, i.e. models formulated in terms of the short rate, is not as straightforward. A notable exception is the Hull-White model. Indeed, we saw above that the mean-reversion level is determined from today's yield curve if the other model parameters are known. It is interesting to observe how the yield curve calibration is transformed from an initial condition to a mean-reversion level through the equivalence relation between the Gaussian model with separable volatility and the Hull-White model.

When calibrated to the yield curve, the free parameters in the Hull-White model are the mean-reversion factor and the volatility. We saw above how these parameters are in one-to-one correspondence with the separable volatility $\Lambda_T \sigma_t$:

$$\begin{cases} \tilde{\sigma}_t &= \Lambda_t \sigma_t \\ \lambda_t &= -(\partial_t \Lambda_t)/\Lambda_t \end{cases} \Leftrightarrow \begin{cases} \sigma_t &= \tilde{\sigma}_t / \exp\left(-\int_0^t \lambda_s ds\right) \\ \Lambda_t &= \exp\left(-\int_0^t \lambda_s ds\right) \end{cases}$$

where Λ_T and σ_t have been scaled so that $\Lambda_0 = 1$.

When simulating the forward rates f_{tT}, the first variable t is evolved while the second variable T is kept fixed. This is in contrast to the short rate $r_t = f_{tt}$ for which both variables are evolved. This motivates us to study the rolling forward rates $f_{t,t+T}$ that describe the instantaneous forward rate at a certain time period T into the future instead of at a fixed time as was done for f_{tT}. The short rate is then obtained in the limit $T \to 0$. The evolution of the rolling rates for the Gaussian model with separable volatility can be derived in the same way as was done for the short rate:

$$\begin{aligned} f_{t+dt,t+dt+T} - f_{t,t+T} &= f_{0,t+dt+T} + \Lambda_{t+dt+T} \int_0^{t+dt} \sigma_s^2 \left(\int_s^{t+dt+T} \Lambda_u du \right) ds \\ &+ \Lambda_{t+dt+T} \int_0^{t+dt} \sigma_s dW_s - f_{0,t+T} - \Lambda_{t+T} \int_0^t \sigma_s^2 \left(\int_s^{t+T} \Lambda_u du \right) ds \\ &- \Lambda_{t+T} \int_0^t \sigma_s dW_s \end{aligned}$$

$$= \left(\partial_t f_{0,t+T} + \frac{\partial_t \Lambda_{t+T}}{\Lambda_{t+T}} (f_{t,t+T} - f_{0,t+T}) + \Lambda_{t+T} \sigma_t^2 \int_t^{t+T} \Lambda_s ds \right.$$
$$\left. + \Lambda_{t+T}^2 \int_0^t \sigma_s^2 ds \right) dt + \Lambda_{t+T} \sigma_t dW_t$$

from which it follows that the rolling rates are mean reverting just as the short rate.

We conclude that forward rates (with a fixed payment date) of separable volatility are equivalent with mean-reverting rolling rates. We have also seen that when specifying an Ornstein-Uhlenbeck process for the short rate, the evolution of all the forward rates can be derived. A consequence is that when specifying an Ornstein-Uhlenbeck process for the short rate, the rolling forward rates must all follow Ornstein-Uhlenbeck processes with parameters determined from the short-rate process. It means that there is no degree of freedom left for $f_{t,t+T}$ once an Ornstein-Uhlenbeck process has been specified for the special case $T = 0$.

Recall that the no-arbitrage condition takes the form of the HJM condition for the forward rates. We now see that in terms of the rolling rates, it takes the form that all rates are determined from the short rate. This result is general and not restricted to the Hull-White model. Indeed, any model for the short rate predicts the evolution of the whole yield curve. This can be understood from the relation $\bar{P}_{tT} = E_t \left(\exp \left(-\int_0^t r_s ds \right) \right)$. The reverse is not true, i.e. an arbitrary stochastic evolution of the yield curve may not necessarily be described by a short-rate model. To obtain richer dynamics of the yield curve it has therefore become popular to formulate models directly in terms of forward rates or LIBOR rates.

In parallel to the modeling of a commodity futures curve, a separable volatility implies a highly correlated curve. The Hull-White model should therefore not be used to price products that have a strong dependence on the correlation between rates. Instead, correlation-dependent products are best handled with higher-dimensional models, see Sect. 12.3.

A whole myriad of short-rate models have been suggested in the literature, see Brigo and Mercurio (2006) for an overview. Most of the models are not analytically solvable and are cumbersome to calibrate and to generalize to price correlation-dependent products. Our personal view is that models based on the theory in this chapter are preferable and for this reason we only consider the Hull-White model of all the short-rate models. Furthermore, short-rate models are inconsistent if allowing the short rate to be lognormal. Indeed, disregarding the contribution from the drift gives $r_t = r_0 \exp(\sigma W_t)$, where we have assumed a constant volatility. The expected value of the money market account is then given by

$$E[B_t] \sim E\left[\exp\left(\int_0^t r_u du\right)\right] \approx E\left[\exp\left(\frac{1}{2} t \left(r_0 + r_t\right)\right)\right]$$

where the integral has been approximated by the trapezoidal method. As $E\left[\exp\left(e^{\eta}\right)\right]$ is divergent for $\eta \sim \mathcal{N}(0,1)$, this heuristic argument shows that the expected value of the money market account is divergent for lognormal short rates.

Fortunately, the divergence disappears if the model is implemented numerically on a lattice, which means that the model can anyway be used in practice. It is then important to be aware of the fact that the price depends on the choice of lattice and no convergence is obtained in the limit when the distance between the points tends to zero.

13.16 Markov-Functional Models

Markov-functional models, developed in Hunt et al. (2000), are based on the fact that an arbitrary interest rate product can be priced if the discounted zero-coupon bond prices $N_t^{-1} P_{tT}$ are known for arbitrary t and $T > t$, where N_t is the numeraire. This fact follows from Sect. 13.8. As

$$N_t^{-1} P_{tT} = E_t \left[N_T^{-1} \right]$$

all that we need to know is $E_t \left[N_T^{-1} \right]$. To compute this conditional expectation, we assume that the numeraire can be written as a function of a simple process, chosen such that the computations are straightforward. We typically, but not necessarily, assume that the numeraire is a function of a Brownian motion: $N_T = \theta_T \Phi_T (W_T)^{-1}$. θ_T is a scalar and Φ_T is such that $E[\Phi_T(W_T)] = 1$, where the expectation is conditional on today's state. θ_T can then be calibrated to today's yield curve according to

$$P_{0T} = E \left[N_T^{-1} \right] = \theta_T^{-1} E \left[\Phi_T(W_T) \right] = \theta_T^{-1}$$

where we have assumed that $N_0 = 1$. The Markov-functional approach therefore amounts to finding suitable functions $\Phi_T(W_T)$ with mean equal to 1. An example of a Markov-functional model was given in Sect. 13.10 where

$$N_t = B_t = \bar{P}_{tt}^{-1} = \bar{P}_{0t}^{-1} \exp \left(\frac{1}{2} \int_0^t \Psi_{st}^2 ds - \int_0^t \Psi_{st} d W_s \right)$$

The main characteristic feature of Markov-functional models is the flexibility in the choice of functions $\Phi_t(W_t)$ and therefore in the matching of skew and smile. To illustrate how an appropriate function Φ can be constructed, let $W_{\omega_{tt}} = \int_0^t \Psi_{st} d W_s$ in the Gaussian model be replaced with $g(W_{\omega_{tt}})$ for some suitable function g. An example of such a function is

$$g(x) = x + \frac{1}{a + bx + cx^2}$$

which can be used to match the skew and the smile. Note that Φ needs to be rescaled to preserve the identity $E[\Phi_t] = 1$.

Markov-functional models are usually implemented on a discrete tenor structure $\{T_i\}$. To simplify notation, we set $P_{ij} = P_{T_i T_j}$ and $N_i = N_{T_i}$. $\{N_i^{-1} P_{ij}\}_{i \leq j}$ can then be computed from $\{N_i\}$ by the pricing formula.

It is useful to formulate the model in terms of LIBOR rates instead of the numeraire when calibrating to caplets. The numeraire and the LIBOR rates are related by

$$P_{i,i+1}/N_i = E_i\left[N_{i+1}^{-1}\right] \Leftrightarrow N_i^{-1} = (1 + \delta_i L_i(T_i)) E_i\left[N_{i+1}^{-1}\right]$$

If the numeraire is chosen as the terminal bond, $N_i = P_{in}$, then $N_n = 1$ and the above formula can be used to recursively determine the numeraire from the distribution of the LIBOR rates L_i at their fixing dates T_i. The distribution of the LIBOR rate can be chosen to match today's yield curve in a similar way to what was done for the numeraire. Indeed, setting $L_i = \theta_i \Phi(W_i)$, the scalar is determined from

$$\frac{P_{0i} - P_{0,i+1}}{P_{0n}} = E_0\left[\frac{1 - P_{i,i+1}}{P_{in}}\right] = E_0\left[\frac{\delta_i L_i(T_i) P_{i,i+1}}{N_i}\right] = \delta_i \theta_i E_0\left[\Phi_i E_i[N_{i+1}^{-1}]\right]$$

$$\Leftrightarrow \theta_i = \frac{P_{0i} - P_{0,i+1}}{\delta_i P_{0n} E_0\left[\Phi_i E_i[N_{i+1}^{-1}]\right]}$$

The above formulae can appear quite abstract. For this reason, we illustrate with a tree implementation. We assume that W_i can only attain certain discrete values $\{\sqrt{T_i} a_k\}_k$, with $a_0 = 0$. The probabilities

$$\eta_{lk}^{ji} = P(W_j = \sqrt{T_j} a_l | W_i = \sqrt{T_i} a_k)$$

are then assumed to be known. Regarding the values of the distribution, we use the notation $\Phi_{il} = \Phi_i(\sqrt{T_i} a_l)$. It follows that

$$\theta_{n-1} = \frac{P_{0,n-1} - P_{0,n}}{\delta_{n-1} P_{0n} \sum_l \eta_{l0}^{n-1,0} \Phi_{n-1,l}}$$

$$\Rightarrow N_{n-1,k}^{-1} = N_{n-1}^{-1}|_{W(T_{n-1}) = \sqrt{T_{n-1}} a_k} = 1 + \delta_{n-1} \theta_{n-1} \Phi_{n-1,k}$$

$$\Rightarrow \theta_{n-2} = \cdots$$

$$\Rightarrow \cdots$$

which calibrates the model to today's yield curve through the choice of $\{\theta_i\}$ and determines the numeraire values $\{N_i\}$ used in the pricing.

The calibration to volatility-dependent products can be complex for Markov-functional models and often needs to be done numerically. This is fortunately not always as bad as it first appears. For instance, we saw in Chap. 4 that there are several techniques by which a numerical calibration can be improved performance-wise. Furthermore, for Markov-functional models it is possible to choose the evaluation equations in the pricing measure to avoid complicated drift computations which can be a performance bottleneck for interest rate models, see next section. The processes can also be chosen to be analytic which means that a simulation does not have to be limited to small time steps. Instead, the rates can be evaluated in one go to the time points where the pricer needs them. Thus, even though calibration by simulation appears performance-challenged from the outset, this speed-up can make it feasible. However, bear in mind that for Markov-functional models it is necessary to calibrate to the yield curve as well as to the volatilities while for the other models in this chapter, the yield curve calibration is automatic.

When implementing Markov-functional models on a discrete tenor structure, the values $\{N_i^{-1}P_{ij}\}_{i\leq j}$ are obtained. For dates not aligned with the tenor dates, $N_t^{-1}P_{tT}$ can be obtained from interpolation. Unfortunately, interpolation leads to mispricing of volatility-dependent products when more than one driving factor is used. The reason is that account is not taken to the stochastic nature of the processes between the tenor dates. To give an example that illustrates the effect, assume that L_1 and L_2 follow normal processes with volatility σ and correlation ρ. The linearly interpolated mid rate follows

$$d\frac{1}{2}(L_1 + L_2) = \frac{1}{2}\sigma(dW_1 + dW_2) = \frac{\sqrt{1+\rho}}{\sqrt{2}}\sigma dW$$

and has a volatility less than σ unless L_1 and L_2 are perfectly correlated.

Observe that the reduced volatility due to interpolation is not limited to Markov-functional models but applies to any model implemented on a discrete tenor structure. It therefore also occurs for LMMs, SMMs and for models in other asset classes as well. The problem can be avoided by not doing an interpolation but rather using a Brownian bridge process to generate values on dates not lying on the simulation dates.

13.17 LIBOR Market Models

The Gaussian model gives the following SDEs for the LIBOR rates in their natural (T'-forward) measure:

$$dL_{tTT'} = \frac{1}{\delta}(1 + \delta L_{tTT'})(\Psi_{tT} - \Psi_{tT'})dW_t^{T'}$$

We focus on LIBOR rates $L_{ti} = L_{tT_iT_{i+1}}$ defined on a discrete tenor structure T_0, $T_1, \ldots, T_n$. The SDE can then be written as

$$dL_{ti} = (1 + \delta_i L_{ti})\sigma_{ti}dW_{ti}$$

where $W_{ti} = W_t^{T_i+1}$ and $\sigma_{ti} = (\Psi_{tT_i} - \Psi_{tT_{i+1}})/\delta_i$. Models formulated as SDEs of discrete LIBOR rates are called *LIBOR market models (LMMs)* and were developed independently by Brace et al. (1997) and Miltersen et al. (1997). Needless to say, these models are simple to calibrate to caplets.

As we use a discrete tenor structure, the continuous-time numeraire (the money market account) cannot be used. It is instead popular to work in the terminal (T_n-forward) measure. The Radon-Nikodym derivative is then given by

$$M_t = \frac{P_{tn}/P_{0n}}{P_{ti}/P_{0i}} = \prod_{j=i+1}^{n} \left(\frac{1+\delta_j L_{tj}}{1+\delta_j L_{0j}} \right)^{-1}$$

where $P_{ti} = P_{tT_i}$. It gives the expression for W_i^n, the Brownian driver of L_i in the terminal measure:

$$dW_{ti} = dW_{ti}^n + \langle dW_{ti}, d\ln M_t \rangle = dW_{ti}^n - \sum_{j=i+1}^{n} \langle dW_{ti}, d\ln(1+\delta_j L_{tj}) \rangle$$

$$= dW_{ti}^n - \sum_{j=i+1}^{n} \frac{\delta_j}{1+\delta_j L_{tj}} \langle dW_{ti}, dL_{tj} \rangle = dW_{ti}^n - \sum_{j=i+1}^{n} \delta_j \rho_{ij} \sigma_j$$

The resulting SDE is

$$dL_{ti} = (1+\delta_i L_{ti})\sigma_{ti} \left(dW_{ti}^n - \sum_{j=i+1}^{n} \delta_j \rho_{tij} \sigma_{tj} \, dt \right)$$

Observe that we without any additional complications have allowed the Brownian drivers to be correlated.

The LIBOR rates satisfy a shifted lognormal SDE that is close to the normal SDE as $\delta_i L_i$ is in general much smaller than 1. The reason why the LIBOR rates follow a more complicated SDE than f_{tT} is that the simple compounding "disrupts" the process. Indeed, if instead working with continuous compounded forward rates $f_i = \frac{1}{\delta_i} \ln(1+\delta_i L_i)$, a Gaussian behavior is obtained:

$$df_{ti} = \frac{1}{1+\delta_i L_{ti}} dL_{ti} - \frac{1}{2} \frac{\delta_i}{(1+\delta_i L_{ti})^2} (dL_{ti})^2$$

$$= \sigma_{ti} \, dW_{ti}^n - \left(\sigma_{ti} \sum_{j=i+1}^{n} \delta_j \rho_{tij} \sigma_{tj} + \frac{1}{2} \delta_i \sigma_{ti}^2 \right) dt$$

The discrete-time version of the money market account is given by a constant reinvestment at the LIBOR rate for each tenor date. The time t value of this strategy is given by

$$P_{t\theta_t} \prod_{j=0}^{\theta_t} (1 + \delta_j L_j)$$

where $\theta_t = i$ if $t \in (T_{i-1}, T_i]$ and T_{-1} is today's date. The corresponding measure is called the *spot measure* and is, together with the terminal measure, the most popular pricing measures for LMM.

Just as in Sect. 13.12, it is common to sacrifice the analytical properties of the Gaussian model for a better match of the skew and smile by allowing the volatility to depend on the rates. For instance, the most popular type of LMMs is the lognormal LIBOR market model:

$$dL_{ti} = \sigma_{ti} L_{ti} dW_{ti}$$

As we work on a discrete tenor structure, the model avoids the divergence that otherwise occurs for a lognormal short rate. The model has the form

$$dL_{ti} = \sigma_{ti} L_{ti} \left(dW_{ti}^n - \sum_{j=i+1}^{n} \rho_{tij} \frac{\delta_j}{1 + \delta_j L_{ti}} \sigma_{tj} L_{tj} \right)$$

in the terminal measure. The calibration is trivial as $\{\sigma_i\}$ are the lognormal caplet volatilities. Because of the simple conversion between ATM volatilities for the lognormal and the shifted lognormal model, as was described in Sect. 5.4, the calibration to ATM caplets is simple for the Gaussian model as well. The skew of the LMM can be controlled by using a CEV process or a shifted lognormal process:

$$dL_{ti} = \sigma_{ti} \left(L_{0i} + \beta_i (L_{ti} - L_{0i}) \right) dW_{ti}$$

Finally, we would like to point out that it is straightforward to extend the LMM to include stochastic volatility and jumps in parallel to what was done in Sects. 13.13 and 13.14.

13.18 Swap Market Models

The main reason for using the LMM is the simple caplet calibration. Unfortunately, the swaption calibration is less straightforward for these types of models, as we show in Sect. 13.20. An alternative approach is to formulate the model in terms of swap rates on a discrete tenor structure instead of in terms of LIBOR rates. Such models are called *swap market models (SMM)* and they are characterized by their simple calibration to swaptions. The caplet calibration unfortunately needs to be done using techniques similar to how swaption calibration is done for LMMs.

The choice between the LMM and the SMM depends whether the product to be priced is more sensitive to the correctness of the caplet quotes or swaption quotes. However, the LMM is generally viewed as the simpler model to work with and it is often the model type of choice.

Just as for LMM, we work with rates defined on a tenor structure $T_0, T_1, \ldots, T_n$. We base the model on swap rates with a common final payment date T_n, i.e. $R_i = (P_i - P_n)/A_i$, where $A_i = \sum_{j=i+1}^{n} \delta_j P_j$ is the annuity. This particular class of models is called coterminal SMMs. Other choices are also possible, e.g. all swap rates can have the same start date, see Galluccio et al. (2007) for a discussion of various types of swap market models.

We initially formulate the model in the measure belonging to the numeraire A_i. This makes the swaption pricing and calibration straightforward. The swap rate is a martingale in this measure and we assume an SDE of the form

$$dR_{ti} = \sigma_{ti}\, d\tilde{W}_{ti}$$

where $\tilde{W}_i$ is a Brownian motion in the measure corresponding to A_i and the volatility σ_i is allowed to depend on R_i.

We choose the terminal measure as the common measure in which the pricing is done. The transformation to this measure is given by the Radon-Nikodym derivative

$$M_t = \frac{P_{tn}/P_{0n}}{A_{ti}/A_{0i}}$$

and Girsanov's theorem

$$d\tilde{W}_{ti} = dW_{ti}^{n} + \langle d\tilde{W}_{ti}, d \ln M_t \rangle$$

To simplify the computations, we set

$$\omega_{ij} = \sum_{k=j}^{n-1} \delta_{k+1} \prod_{l=i+1}^{k} (1 + \delta_l R_l), \quad i \leq j$$

and $\omega_i = \omega_{ii}$. As

$$A_{n-1} = \delta_n P_n = \omega_{n-1} P_n$$

it follows by induction that

$$\begin{aligned}
\omega_i P_n &= (\delta_{i+1} + (1 + \delta_{i+1} R_{i+1})\omega_{i+1})\, P_n \\
&= \delta_{i+1} P_n + (1 + \delta_{i+1} R_{i+1}) A_{i+1}) = \delta_{i+1} P_n + A_{i+1} + \delta_{i+1}(P_{i+1} - P_n) \\
&= A_i
\end{aligned}$$

We then obtain

$$\begin{aligned}
d\ln M_t \sim d\ln \omega_{ti} &\sim -\frac{1}{\omega_{ti}}\sum_{k=i}^{n-1}\delta_{k+1} d\prod_{l=i+1}^{k}(1+\delta_l R_{tl}) \\
&= -\frac{1}{\omega_{ti}}\sum_{k=i+1}^{n-1}\sum_{j=i+1}^{n-1}\delta_{k+1}\left(\prod_{l=i+1,l\neq j}^{k}(1+\delta_l R_{tl})\right)\delta_j\sigma_{tj}\,d\tilde{W}_{tj} \\
&= -\frac{1}{\omega_{ti}}\sum_{j=i+1}^{n-1}\sum_{k=j}^{n-1}\delta_{k+1}\left(\prod_{l=i+1,l\neq j}^{k}(1+\delta_l R_{tl})\right)\delta_j\sigma_{tj}\,d\tilde{W}_{tj} \\
&= -\frac{1}{\omega_{ti}}\sum_{j=i+1}^{n-1}\frac{\delta_j\sigma_{tj}}{1+\delta_j R_{tj}}\omega_{tij}\,d\tilde{W}_{tj}
\end{aligned}$$

which gives the SDE for the swap rate in the forward measure:

$$dR_{ti} = \sigma_{ti}\,dW^n_{ti} - \frac{\sigma_{ti}}{\omega_{ti}}\sum_{j=i+1}^{n-1}\frac{\delta_j\sigma_{tj}}{1+\delta_j R_{tj}}\omega_{tij}\rho_{tij}\,dt$$

The choice of letting σ_i be proportional to the rate R_i is common. This is the lognormal swap market model. Observe that since ω_{ij} depends on $\{R_k\}_{i+1\leq k\leq n-1}$, unlike the LMM, there is no functional form of σ_i that makes the drift and the diffusion state-independent.

13.19 Including Adjusters

When using a Gaussian interest rate model

$$dL_{ti} = (1+\delta_i L_{ti})\sigma_{ti}\,dW_{ti}$$

we obtain a Markovian model. It means that a tree implementation is possible and that simulations can be done efficiently using large time steps. Unfortunately, with the exception of adding an uncorrelated volatility process, there are no remaining degrees of freedom for fitting the skew and the smile of the implied volatility surface. Indeed, if the skew and the smile are prioritized and we choose a non-Gaussian process for the underlying, e.g. the lognormal LMM, the Markovian property is destroyed and the implementation is more complex and performance demanding.

This dilemma of yield curve modeling can be solved with the method of adjusters. It is then possible to use the Gaussian model for the evolution of the curve while a skew and smile model can be used for the adjustment. In that way we obtain a high-performing version of the LMM that can match the skew and the smile.

13.20 Volatilities and Correlations

The modeling of interest rates depends on a curve, the yield curve. It is therefore similar to the modeling of commodities that depend on the futures curve. As we now show, the parallel is particularly strong regarding the choice of volatilities and correlations. Much of the results in Sect. 12.4 apply here as well and we therefore choose to focus on the differences.

Let us start with the calibration of the volatility surface σ_{tT} to the liquid volatility-dependent products for interest rates, namely caplets (caps) and swaptions. We show below that the calibration is more complex than for commodities as the two time variables t and T become mixed. We therefore discuss the volatility calibration in more detail.

Consider the calibration to caplet and swaption quotes for a given tenor structure $T_0, \dots, T_N$. We assume that the model is formulated in terms of SDEs for the LIBOR rates $L_i = L_{tT_iT_{i+1}}$ in their natural measure. The calibration to caplets is then straightforward. For the calibration to swaptions, we express the swap rate R_{ij} with tenor structure $T_i, \dots, T_j$ in terms of the underlying LIBOR rates:

$$R_{ij} = \frac{1 - \prod_{l=i+1}^{j} \frac{1}{1+\delta_l L_l}}{\sum_{k=i+1}^{j} \delta_k \prod_{l=i+1}^{k} \frac{1}{1+\delta_l L_l}}$$

As R_{ij} depends on several LIBOR rates, we can use a higher-dimensional technique of Chap. 10 to obtain an approximate distribution. For example, when the LIBOR rates are lognormal it is popular to write the stochastic part of $d \ln R_{ij}$ as $\tilde{\sigma}_{ij}\, dW_{tij}$ and freeze the $\{L_k\}$-dependent volatilities at today's values $\{L_k(0)\}$. This method by Hull and White (2000) is popular and the accuracy can be verified numerically by computing the ATM swaption prices with the SDEs for the LIBOR rates. The SDEs for the swap rates are commonly approximated by the same type of SDE followed by the LIBOR rates. For instance, if the LIBOR rates satisfy a lognormal, shifted lognormal, CEV or SABR process, the swap rates are approximated with the same type of process, see, for example, Hagan and Lesniewski (2008) for the SABR evaluation. This is a natural approach as the LIBOR rates are the 1-period special case of swap rates.

As an illustrating example for the calibration, we use our benchmark model

$$dL_{ti} = \sigma(1 + \delta_i L_{ti})dW_{ti}$$

We start by describing how the calibration can be done for piece-wise constant volatilities

$$\sigma_{tT} = \sigma_{kl}, \quad t \in [T_{k-1}, T_k], T \in [T_{l-1}, T_l]$$

where T_{-1} is today's date. Observe that $0 \le k < l$ as the volatility of a rate is not defined after its fixing date. A swaption with tenor structure $T_i, \dots, T_j$ then depends on the volatilities with $l = i + 1, \ldots, j$. These volatilities need to be evolved up to the fixing T_i of the underlying swap rate, which means that $k = 0, 1, \ldots, i$. For example, the swaption with tenor structure (T_1, T_2, T_3) depends on σ_{02}, σ_{03}, σ_{12} and σ_{13}.

To see how the volatilities σ_{kl} can be bootstrapped from the swaption quotes on the tenor structure, we first focus on the (T_0, T_1) swaption (which is actually a caplet). This swaption only depends on σ_{01} and can therefore be determined from the market quote of the swaption (caplet). Next, consider the (T_0, T_1, T_2) swaption, which depends on σ_{01} and σ_{02}. As σ_{01} has been computed, σ_{02} can be backed out. By the same token, all σ_{0l} can be determined for arbitrary positive integers l. For $k = 1$, the (T_1, T_2) swaption (caplet) depends on σ_{02} and σ_{12}. The first volatility is known, so the latter can be backed out. Repeating this procedure, we can determine all volatilities for $k = 1$ and then in the same way for $k > 1$.

For N LIBOR rates, there are $N(N + 1)/2$ parameters and as many calibration instruments. This large number of parameters can lead to an overspecification of the model. An alternative is to reduce the number of parameters by assuming a product form of the volatility. This approach is similar to what we did for commodities in Sect. 12.4.

The shape of the volatility surface is fundamentally different from that found for commodities markets. For interest rates, the short end of the yield curve is influenced by political choices. This has a damping effect which means that the volatility often increases initially as a function of the maturity T, for fixed t. After a certain time period out on the curve, the volatility behaves in a more normal way and starts to decay. Thus, the volatility is often found to have a humped shape. A popular choice of parametric form is therefore

$$\sigma_{tT} = \alpha + (\beta + \gamma(T - t))e^{-\delta(T-t)}, \quad \delta > 0$$

This can obviously be generalized by allowing the parameters to depend on t.

The correlation modeling for interest rates and commodities have many parallels. For example, with only some minor modifications it is possible to use the parametric models in Sect. 12.4. It is popular to use a parametric correlation surface, often based on historical data, and after that calibrate the volatility. One difference from commodities is that there are liquid calibration instruments (the swaptions) that depend on the intracorrelation. As they have a volatility dependence that cannot be directly inferred from elsewhere, however, it can be hard to extract the correlation component. The correlation can also be obtained from other, less liquid, products such as CMS spreads. When it comes to the intercorrelation, i.e. the correlation between interest rates in different currencies, there are few market instruments that can be used for the calibration.

Recall the full-rank intracorrelation parametrization in Sect. 12.4:

$$\rho(T, T') = \exp(\phi_T - \phi_{T'}) \quad T \le T'$$

where ϕ_T is a concave and increasing function. It implies that

$$\frac{d^2}{dT^2}\rho(T_0,T) = \exp(\phi_0 - \phi_T)\left(\left(\frac{d\phi_T}{dT}\right)^2 - \frac{d^2\phi_T}{dT^2}\right) \geq 0$$

so $\rho(T_0,T)$ is convex. This is in disagreement with empirical studies where $\rho(T_0,T)$ is found to be concave for T close to T_0 and convex for larger T, see Rebonato (2002) and references therein. Furthermore, the 2-factor model in Sect. 12.4 also changes sign of the convexity and the model we are looking for should preferably be a higher-order correction of the 2-factor model and should therefore not have fundamental different properties. Thus, we have good reasons to require the model to have a convexity that changes sign.

Although the 2-factor parametrization shows good agreement with reality when it comes to the convexity, it cannot reproduce the decorrelation observed in the market. For instance, studies of historical yield curves movements report that the largest eigenvalue often accounts for only 80–90% of the dynamics, while for 2-factor models it is virtually impossible to get this number below 90%. Therefore, if choosing between the 2-factor parametrization and the full-rank parametrization above, the modeler has to decide which is more important: the convexity or the decorrelation.

An alternative approach is to use a new type of parametrization or by modifying one of the existing ones. For example, the proof of the convexity of the full-rank parametrization only used the cocycle condition and condition 6 of Sect. 12.4. Therefore, concavity can be obtained for T close to T_0 by relaxing one of these conditions. For instance, consider the situation when condition 6 is relaxed and ϕ_T is no longer concave everywhere. It is possible to obtain a correlation with changing sign of the convexity. This approach makes some sense for interest rates as condition 6 is not specially important. Indeed, we know from the discussion of volatilities that most of the dynamics of the yield curves do not lie in the very front (which is the case for commodities) and it is therefore not unreasonable to assume that the rates are decorrelated the most at the place of the curve where the volatility hump is found.

The choice of driving factors is important for correct hedging and pricing. Unlike commodities it is not always clear beforehand how interest rate products depend on the correlation. The reason is that the interest rate calibration instruments themselves depend on the correlation. For a concrete example, we follow Andersen and Andreasen (2001) and consider the pricing of a Bermudan swaption in the LMM. This model is usually calibrated to swaptions (and caplets) which, just as the Bermudan swaption, depend on the correlation between the LIBOR rates. It means that if the correlation structure is changed, it is not always obvious where the effect is stronger: for the calibration instruments or for the product to be priced. In order to retain the swaption market prices, the change in their prices from the correlation needs to be compensated by changing the volatilities. It is then not clear if the end result is an increase or a decrease of the Bermudan swaption price.

13.21 Interest Rate Effects on Other Asset Classes

So far in this chapter, we have been concerned with pure interest rate modeling. We now study the impact of the yield curve evaluation on other asset classes. For a concrete example, we consider the pricing of an exotic derivative that depends on the value of a tradable S. We use a lognormal process

$$dS_t = \mu_t S_t dt + \sigma_t S_t dW_t$$

formulated in the risk-neutral measure. As S_t/B_t is a martingale, it follows that

$$\left(0 = d\frac{S_t}{B_t}\right)_{\text{drift}} = (\mu_t - r_t)\frac{S_t}{B_t}dt \Leftrightarrow \mu_t = r_t$$

The volatility σ_t is usually determined through calibration to vanilla option prices. Consider, for example, the pricing of a call option paying $(S_T - K)_+ = (F_T - K)_+$ at T, where $F_t = P_{tT}^{-1} S_t$ is the forward. As usual, the option price depends on the volatility of the martingale F_t. We see from the relation between the underlying and the forward that their volatilities are only equal for deterministic interest rates. To obtain a more general result, assume that the interest rates satisfy

$$df_{tT} \sim \sigma_t^{\text{IR}} f_{tT} dW_t^{\text{IR}}$$

where the drift has been omitted as it does not affect the volatility computations. We have assumed a lognormal process for the interest rates since it makes the analysis easier.

Using

$$dP_{tT} = d\exp\left(-\int_t^T f_{tT'}dT'\right) \sim -\sigma_t^{\text{IR}} P_{tT}\left(\int_t^T f_{tT'}dT'\right)dW_t^{\text{IR}}$$

we obtain

$$d\left(P_{tT}^{-1}S_t\right) \sim P_{tT}^{-1}S_t\left(\sigma_t dW_t + \sigma_t^{\text{IR}} P_{tT}\left(\int_t^T f_{tT'}dT'\right)dW_t^{\text{IR}}\right)$$

which gives the volatility

$$\tilde{\sigma} = \sigma\left(1 + 2\rho\frac{\sigma_t^{\text{IR}}}{\sigma_t}\int_t^T f_{tT'}dT' + \left(\frac{\sigma_t^{\text{IR}}}{\sigma_t}\int_t^T f_{tT'}dT'\right)^2\right)^{1/2}$$

for the forward. To understand the interest rate effect on the volatility more easily, assume that the rates and the volatilities are constant: $f_{tT} = f_t$, $\sigma_t = \sigma$ and $\sigma_t^{\text{IR}} = \sigma^{\text{IR}}$. As rates are typically in the order of magnitude of 5%, it follows that

$\int_t^T f_{tT'} dT' = f_t(T-t) \ll 1$ if the maturity is not too long. We can then omit the last term to obtain

$$\tilde{\sigma} \approx \sigma \left(1 + \rho \frac{\sigma^{\mathrm{IR}}}{\sigma} f_t(T-t) \right)$$

We conclude that the volatility of the forward is approximately equal to the volatility of the underlying for short maturities. The two volatilities then deviate linearly with the maturity. If the correlation is zero, on the other hand, the relation is purely quadratic for small maturities. By the same token, it can be shown that the skews and smiles are approximately equal for the forward and the underlying when the maturity is small. For longer maturities, however, they deviate in general.

The effect that we have described here is not only important for the calibration of exotics models, but also for long-dated vanilla options. Indeed, it follows from the above how such options need to be dynamically hedged not only with respect to the underlying but also by the interest rate components.

13.22 Overnight Index Swaps

We have defined P_{tT}^{-1} as the fair amount to be paid back at T for a loan of \$1 at t. In a practical trading situation, the value of P_{tT}^{-1} depends on the credit worthiness of the borrower because a default means that the lender will not necessarily be repaid. Thus, the value of P_{tT}^{-1} is high if the loan taker has a poor credit rating. The lowest possible value of P_{tT}^{-1} corresponds to loan takers with zero probability of default. These risk-free loans can to a good approximation be taken by stable governments and certain major corporations.

The credit exposure depends on the tenor $T - t$ of the loan. The longer the tenor, the larger the risk of the loan taker to default during the period of the loan. It means that the interest rate for overnight loans taken by highly rated corporations is close to risk free. Money can therefore be lent risk free for longer time periods by repeatedly entering overnight contracts with counterparties that are considered safe. The money repaid from one overnight loan is used for a new loan expiring the following day.

One problem with the strategy of repeated overnight loans is that the future overnight interest rates are unknown as of today. For instance, we do not know in advance the overnight interest rate that will apply between the date one month from now and the day after that. It means that the future earnings of the strategy are unknown.

Overnight index swaps (OISs) can be used to overcome the above problem. In these products, the earnings obtained up to a certain time in the overnight lending strategy are exchanged for a fixed-rate cash flow. For a more detailed description, consider the OISs that are actively traded in the Euro zone. They are based on the *euro overnight index average (EONIA)* which is a weighted average of overnight interest rates quoted by a selected panel of major banks. The swaps themselves are called *EONIA swaps*. They use the actual/360 day count convention and the payment takes place one day after the end date. For instance, a 1M EONIA swap with 4% fixed rate traded at August 10, a Wednesday, uses a floating rate r defined by

$$1+\frac{31}{360}r=\left(1+\frac{3}{360}r_{\mathrm{A12}}\right)\left(1+\frac{1}{360}r_{\mathrm{A15}}\right)\cdots\left(1+\frac{1}{360}r_{\mathrm{S8}}\right)\left(1+\frac{3}{360}r_{\mathrm{S9}}\right)$$

where r_{A12} is the overnight interest rate fixed at August 12, and so on. Observe that because of non-business days, some of the rates are used for more than 1 day (e.g. 3 days over a weekend). The cash flow resulting from the swap equals

$$N\frac{31}{360}(r-4\%)$$

where N is the notional. The payment takes place one day after the end of the swap, in this instance September 13. EONIA swaps with maturities less than 1Y have only a single payment at the end of the swap. For longer maturities, the payments are annual.

OISs in other currencies are very similar, such as the GBP OISs, which are based on the *sterling overnight index average (SONIA)* and the USD OISs, which are based on the *federal funds rate*. For example, the only difference between USD OISs and EONIA swaps is that the payment takes place 2 instead of 1 day after the end of the swap. As usual, there are exceptions to every rule. For instance, ZAR OISs use a different compounding convention. Furthermore, for some currencies there exist overnight index swaps that are forward starting and defined for periods between monetary policy meeting dates.

SONIA, EONIA, TONAR (JPY) and TOIS (CHF) swaps are quoted out to 30Y while USD OISs can be found out to 2Y. OIS quotes for other major currencies (e.g. AUD, NZD, CAD, RUB, SEK) can be found out to a year while for some currencies (e.g. NOK), they are not traded at all.

Combining the strategy of repeated overnight loans with an OIS results in the floating swap leg remaining the loans and only the fixed leg remaining. It means that as of today, the earnings obtained at the end date of the swap are known. Assuming a maturity of less than a year, so that the OIS only has a single payment, an investment of \$1 accumulates to $\$(1+r\cdot d/360)$ at maturity, where d is the number of days to maturity and r is the OIS rate. Thus, the OIS quotes represent the rates at which risk-free investments can be made.

By bootstrapping OIS quotes and using appropriate interpolation and extrapolation techniques, risk-free rates can be obtained to arbitrary maturities. The resulting curve can be used for discounting future risk-free cash flows. Discount curves for specific counterparties, or classes of counterparties, can be constructed by adding spreads determined by the probability of default.

13.23 Collateral

To avoid credit risk for OTC contracts, *collateral* can be posted between the counterparties. A contract value that becomes negative for one of the counterparties then has to be compensated by paying the corresponding amount to the other counterparty. The terms of the collateralization are most often given by the *Credit Support Annex (CSA)* which is a part of the *ISDA master agreement*. The latter is a standardized contract created by *International Swaps and Derivatives Association (ISDA)* regarding the trading terms between the counterparties of a derivatives deal. Collateralized products are valued on a regular basis, often daily or when a certain threshold has been reached, and collateral is transferred between the counterparties. At the end of the deal, the collateral is paid back in return for the the contract cash flows.

The practice of using collateral reduces the credit risk but does not eliminate it completely. Indeed, there is still the possibility of a default between two collateral posting dates, referred to as the *gap risk*. The gap risk can be non-negligible if the contract value is correlated with the risk of the counterparty defaulting. Furthermore, the situation can get even worse if the value of the collateral is correlated with these quantities. Because of the initial margin, the gap risk is much smaller for exchange-traded product.

In the case of a default, the non-defaulting counterparty typically needs to enter a new contract. The associated risk is called *replacement risk*. Due to the time interval between the default and the writing of the new contract, the market will have moved meaning that the replacement cost can be positive as well as negative. Such a market move can be expensive if the default is correlated with the market move. If the default happens during volatile times, the replacement risk can be particularly costly. For instance, there can be several market participants that all need replacement, leading to a shift in market price. The bid-offer spreads also tend to be large in this scenario.

The terms of the collateral posting can be unilateral as well as bilateral. In the former case, only one of the counterparties is required to pay collateral. The counterparty that does not pay collateral typically belong to a group of market participants commonly referred to as SSA (sovereign, supernatural and agency). We choose to only cover bilateral collateral agreements, for which both counterparties are required to post collateral.

The collateral is usually represented by cash, but other assets are possible. We initially consider single-currency products and assumed that the collateral is in the same currency. The multi-currency extension is dealt with in Sect. 14.11. Furthermore, we assume that the value of the assets used for collateral is uncorrelated with the contract value.

The collateral receiver has to pay interest rate on the collateral. We assume that the overnight interest rate is paid on the collateral as this is the most common situation (when the collateral is cash). To explain the impact of collateralization on derivatives pricing, consider the simple example when two counterparties agree to exchange a fixed cash flow for a floating at a future time T. Assume that the contract is valued to par at deal time but immediately thereafter changes due to a market move. We then assume the market to remain at its level until maturity. The day after the deal, collateral needs to be posted, on which the overnight interest rate should be paid until T. By entering an OIS swap, the daily overnight payments can be seen to be equivalent with a single payment of the OIS rate at maturity. The correct amount of collateral has been posted if this payment together with the collateral amount equals the deal value at T. This can only be true if the required collateral was calculated by discounting the time T cash flow using the OIS rate. We conclude that collateralized derivatives should be priced by discounting future cash flows using the OIS curve. If a different interest rate is paid on the collateral, such as a repo rate or LIBOR rate, then the corresponding curve should be used for discounting.

The choice of discounting curve has a minor effect on products that values to par. The difference in price is due to a mismatch between the dates where the cash flows are positive respective negative. For products with a non-zero present value, on the other hand, the impact can be substantial.

13.24 Credit and Liquidity Risk

We have previously discussed how credit exposure can be handled with CVA and DVA. We now discuss another effect of credit risk that was largely ignored before 2007. Since then, it has been important to take into account credit exposure for market participants that previously were considered as safe. It means that LIBOR rates, which are lending rates between banks, show signs of this effect. This is important as many market instruments are defined in terms of these rates and are therefore affected. Thus, credit exposure needs to be taken into account even when trading simple products such as swaps between two counterparties for which there is negligible default risk.

When pricing cash flows like those from a corporate bond, the discounting can be done by using the risk-free rate together with a spread obtained from credit default models. For major corporations and governments, the input to the models can be the rating from credit agencies such as Standard & Poor's, Moody's and Fitch Ratings. For the computation of the credit spread between LIBOR rates, more sophisticated models for credit defaults are necessary as the set of LIBOR-rated banks can vary with time. To understand how this impacts LIBOR spreads, compare a 6M LIBOR loan with two consecutive 3M LIBOR loans. Both strategies lead to a loss if the counterparty LIBOR-rated bank should default. However, consider the scenario when the LIBOR-rated bank, to which the 6M loan and the first 3M loans are given, becomes downgraded within 3M and then defaults between 3M and 6M. This would lead to a loss for the 6M loan but not for the 3M loans as the second 3M loan was by assumption taken by a LIBOR-rated bank and can therefore not coincide with the defaulting bank. We conclude that if credit models are used to explain the spread between LIBOR rates, they need to model not only the probability of default but also the probability of a downgrade followed by a default. We do not pursue this path but rather take a more direct approach for modeling LIBOR spreads.

We have simplified the terminology by referring to the credit exposure as the sole cause of the difference between an interest rate and a compounding of interest rates with shorter tenors that together span the same time period. As pointed out in Sect. 13.5, other causes are supply and demand, liquidity and premium compensation for unknown future rates. These effects are not necessarily minor compared to the credit exposure. For instance, research (Michaud and Upper (2008)) indicates that credit exposure only explains a relatively small part of the spread between LIBORs and OISs during the credit crunch that started in the summer of 2007. Another possible contributor to rate spreads is that some interest rate quotes, e.g. LIBOR, are only fixings and the banks that quote them are not committed to trade at these levels. This opens up for strategic misrepresentations. To reduce such effects in the BBA LIBOR rates, the upper and lower quartiles of the 8, 12 or 16 banks (depending on the currency) are removed before computing the average.

Rates with certain tenors can be particularly unattractive (or attractive). For instance, the bank sector in a country might have a lot of 1M payments coming in from retail while they prefer a different tenor, for example 3M, that better matches their other cash flows. Such a supply and demand situation also contributes to the spread between LIBOR rates. Furthermore, for each currency there is typically one tenor that is particularly liquid, see the column for the floating swap frequency in Table 13.2, which also affects the spread.

13.25 Interest Rate Surface Construction

The curve construction in Sect. 13.7 is obviously not consistent if credit and liquidity risk is taken into account: the underlying instruments have different features and cannot be combined in a single curve. Instead, a surface has to be constructed that depends on, for example, the start date and the tenor of the interest rate.

We start by considering a cash-collateralized swap for which the floating leg pays 6M LIBOR. It follows from Sect. 13.2 that the swap rate can be written as

$$\frac{\sum_{i=0}^{n-1}(T_{i+1}-T_i)L_{tT_iT_{i+1}}P_{tT_{i+1}}}{\sum_{i=0}^{n-1}(T_{i+1}-T_i)P_{tT_{i+1}}}$$

Observe that $L_{tT_iT_{i+1}}(T_{i+1}-T_i)P_{tT_{i+1}} \neq P_{tT_i} - P_{tT_{i+1}}$ as the LIBOR rates contain 6M credit exposure while the discount factors are risk free because of the use of collateral. When taking the credit exposure into account it is therefore no longer possible to further simplify the swap rate expression as was done in Sect. 13.2.

We instead define the forward LIBOR rates $L_{tT_iT_{i+1}}$ from the condition that $L_{tT_iT_{i+1}}P_{tT_{i+1}}$ is today's value of a LIBOR payment taking place at $P_{tT_{i+1}}$.

Assuming the discount curve has been backed out from OIS quotes, it is possible to determine a curve for 6M forward rates from swap quotes using the above formula. By including 6M FRAs and the 6M deposit rate, the curve can be made granular for shorter maturities. In a similar manner, curves for forward rates with any tenor (for which there are liquid market quotes) can be constructed.

When constructing the OIS curve, it is necessary to allow for jumps at monetary policy meeting dates and to make the curve as flat as possible (while still calibrated to liquid market data) between two such dates. For this purpose, it is useful to calibrate the curve to the forward OISs (between consequtive monetary policy meeting dates) that are traded for some currencies. This effect is only important for the short end of the curve where the market can have a view on the future rates policy. For this segment of the curve it is also important to include the spikes that are caused by low liquidity at specific dates. When it comes to the long end, for most currencies there is a lack of liquidity. The OIS curve can then be extrapolated by defining the risk-free interest rates as a constant spread subtracted from, for example, the 3M or 6M LIBOR curve, or by extrapolating the OIS swap quotes by subtracting a spread from LIBOR swap quotes. Alternatively, the risk-free rates can be determined from an extrapolation of LIBOR curves, for instance, by using the 1M and 3M tenor curves.

One possible approach to obtain an interest rate surface is by first constructing a set of curves for fixed tenors as detailed above. The forward rate for an arbitrary tenor can then be found by interpolation or extrapolation. The discount curve is obtained from the special case when the tenor is equal to one day. Using the interest rate surface, it is possible to price forward starting swaps, swaps with stub periods, exotic interest rate products, etc.

With the above approach it is also possible to price, or calibrate to, *basis swaps*. These are swaps for which both legs are floating, for example, 3M LIBOR against 6M LIBOR. The spread that needs to be added to one of the legs for the price to be at par is called the *basis spread*. When the two legs are in different currencies, the basis swap is a floating vs floating cross currency swap, described in detail in Sect. 14.11.

13.26 Caps, Floors and Swaptions Revisited

We turn our attention to the pricing of simple volatility products in the presence of credit and liquidity spreads. We restrict the discussion to caplets and swaptions. For the purpose of pricing caplets, consider first FRAs, which can be viewed as paying $L_{T_0T_1} - K$ at T_1, where the notional and the day count fraction have been omitted. Using the T_1 forward measure, the present value of the contract equals

$P_{tT_1}E[L_{T_0T_1}-K]$. The FRA rate $L_{tT_0T_1}$ is by definition the value of K such that the contract values to par:

$$L_{tT_0T_1}=E[L_{T_0T_1}]$$

Observe that because of the credit exposure, it is not possible to explicitly compute the expectation as was done in Sect. 13.2.

Following the same path for caplets, the price can be written as $E[(L_{T_0T_1}-K)_+]$. Observe that $L_{tT_0T_1}$ is a martingale as it can be written as an expectation. Caplets can therefore be priced in the usual way by assuming a driftless lognormal (or SABR, etc.) process for the FRA rate $L_{tT_0T_1}$.

In practice, the payment of a FRA is slightly different from the above, namely

$$\frac{L_{T_0T_1}-K}{1+(T_1-T_0)L_{T_0T_1}}$$

at T_0. As the denominator contains a certain amount of credit exposure, it does not exactly represent the risk-free discount factor to T_1. For the purpose of caplet and FRA pricing, however, it can be shown that this effect is negligible (assuming the credit spread has a low volatility).

Just as in Sect. 13.3, it is possible to use the annuity as a numeraire for swaption pricing and arrive at the expression

$$V(t)=A_tE[(R_{T_0}-R_{t=0})_+]$$

The only difference when taking credit risk into account is that the swap rate has a slightly different expression, see Sect. 13.25. Nevertheless, it is still possible to assume driftless lognormal (or SABR, etc.) dynamics to derive an expression for the price.

13.27 Interest Rate Surface Modeling

We here discuss the pricing of products that depend on the whole yield curve. Assume that we have already chosen our favorite interest rate model for the risk-free rate. The simplest extension to account for credit and liquidity effects is to assume deterministic basis spreads added to the risk-free curve. Equivalently, the LIBOR curve, with tenor such that the most liquid volatility instruments can be found, can be used as a base curve from which the deterministic basis spreads are defined. Such models are easy to implement but they misprice products that depend on basis spread volatilities.

A more general approach is to evolve the risk-free rate and the LIBOR rates separate, see Mercurio (2009). In this instance we recommend the Gaussian model to avoid the complication of a state-dependent drift. The calibration is straight forward but the basis spreads can be negative. An alternative approach is to directly model the basis spreads. Using an appropriate process, e.g. the lognormal, means that the spreads remain positive. The drawback is that the calibration becomes more complicated, see Mercurio (2010) for details.

Bibliography

Andersen L, Andreasen J. (2001) Factor dependence of bermudan swaption prices: fact or fiction? J Financial Econ 62:3–37

Brace A, Gatarek D, Musiela M (1997) The market model of interest rate dynamics. Math Finance 7:127–155

Brigo D, Mercurio F (2006) Interest rate models – theory and practice with smile, inflation and credit. Springer Finance, Berlin

Carr P, Lee RW (2008) Robust replication of volatility derivatives. Working paper, Bloomberg LP and University of Chicago

Chen RR, Scott L (2001) Stochastic volatility and jumps in interest rates: an empirical analysis. Working paper, Rutgers University and Morgan Stanley

Cheyette O (1992) Term structure dynamics and mortgage valuation. J Fixed Income 1:28–41

Cheyette O (1996) Representation of the Heath-Jarrow-Morton Model. Working paper, BARRA

Galluccio S, Huang Z, Ly J-M, Scaillet O (2007) Theory and calibration of swap market models. Math Finance 17:111–141

Hagan P (2003) Convexity conundrums: pricing CMS swaps, caps, and floors. WILMOTT Magazine March:38–44

Hagan P, Kumar D, Lesniewski A, Woodward D (2002) Managing smile risk. WILMOTT Magazine, September:84–108

Hagan P, Lesniewski A (2008) LIBOR market model with SABR style stochastic volatility. http://www.lesniewski.us/papers/working/SABRLMM.pdf. Accessed 16 May 2011

Heath D, Jarrow R, Morton A (1992) Bond princing and the term structure of interest rates: a new methodology for contingent claim valuation. Econometrica 60:77–105

Hull J, White A (1990) Pricing interest rate derivative securities. Rev Financial Stud 3:573–592

Hull J, White A (2000) Forward rate volatilities, swap rate volatilities, and the implementation of the LIBOR market model. J Fixed Income 10:46–62

Hunt PJ, Kennedy JE, Pelsser AAJ (2000) Markov-functional interest rate models. Finance Stochast 4(4):391–408

Mercurio F (2008) Cash-settled swaptions and no-arbitrage. Risk February:96–98

Mercurio F (2009) Interest rates and the credit crunch: new formulas and market models. Social Science Research Network. http://papers.ssrn.com/sol3/papers.cfm?abstract_id = 1332205. Accessed 16 May 2011

Mercurio F (2010) A Libor market model with a stochastic basis. Risk December:84–89

Michaud FL, Upper C (2008) What drives interbank rates? Evidence from the Libor panel. BIS Q Rev, March

Miltersen KR, Sandmann K, Sondermann D (1997) Closed form solutions for term structure derivatives with log-normal interest rates. J Finance 52:409–430

Piterbarg V (2005) Is CMS spread volatility sold too cheaply? Presented at II Fixed Income Conference, Prague

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (2002) Numerical recipes in C++: the art of scientific computing. Cambridge University Press, Cambridge

Rebonato R (2002) Modern pricing of interest rate derivatives: the LIBOR Market Model and beyond. Princeton University Press, Princeton

Rebonato R, de Guillaume N (2010) A universal feature of interest rates: the CEV exponent, and it relevance for hedging. Conference presentation, Global Derivatives & Risk Management

Chapter 14
Foreign Exchange

Up to this point, we have only considered cash flows that depend on a single currency. We now cover the multi-currency extension. The fundamental building stones in such a theory are the FX rates X_t^{ij} which represent the time t value in currency j of one unit of currency i. By construction, the FX rates satisfy the inverse relation $(X_t^{ji})^{-1} = X_t^{ij}$ and the cocycle relation $X_t^{ij} X_t^{jk} = X_t^{ik}$.

We single out one of the currencies and use it to define the numeraire and the pricing measure. We call it the *domestic currency* while the others currencies are referred to as *foreign currencies*. A measure defined from a numeraire quoted in the domestic (foreign) currency is called a *domestic (foreign) measure*.

Derivatives depending on foreign exchange can be classified into two types. The first type consists of products that only depend on the values of the FX rates, e.g. FX forwards and FX options. The challenge of modeling this product type is to find a suitable equation for the evaluation of the FX rates. The second product type consists of multi-currency extensions of the products discussed in previous chapters. These products depend on underlyings in the domestic and the foreign currencies in such a way that it is not possible to decompose them into a sum of single currency products. An example is given by a contract that pays a certain cash amount if two underlyings (e.g. interest rates) quoted in different currencies simultaneously exceed some predefined levels. The evolving equation for a foreign underlying is typically assumed to have been defined and calibrated in the corresponding foreign measure. For a consistent pricing with several currencies, it is then necessary to transfer these equations to the (domestic) pricing measure. The methods developed in this chapter can also be used to model combinations of the two product types, i.e. products that depend explicitly on FX rates at the same time as they depend on underlyings in several currencies.

Some countries have chosen to *peg* their currencies. It means that the value of their currency closely (but not exactly) follows another, more major, currency. Examples include the pegging of the Chinese renminbi to the US dollar, the Danish krone to the euro and the historical practice of pegging currencies to gold or silver. An additional example is given by the Chinese currency, which exists in two versions: in the domestic market (CNY) and in the international market (CNH),

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2_14,

with slightly different values. As the dynamics of exchange rates between pegged currency pairs depend on political decisions, we choose not to cover this special case.

We initially focus on forward contracts and static replication. We then turn our attention to dynamic replication techniques and price FX options. Exotic derivatives are analyzed with a focus on foreign underlyings and quantos. We end the chapter by summarizing the conventions used in the FX markets and describing the impact of credit and liquidity risk.

14.1 Static Replication

An FX forward is an agreement to exchange a certain amount K of the domestic currency at maturity T in return for one unit of the foreign currency. With X the domestic-foreign FX rate, the payment can be written as $V_T = X_T - K$ in the domestic currency. It follows that $V_t = P_{tT}^{\text{for}} X_t - P_{tT}^{\text{dom}} K$ since if we have this amount at t, we can sell K domestic bonds and purchase one foreign bond to obtain V_T at T. The cash amount K such that the forward contract is worth zero is called the forward F_t and is equal to $(P_{tT}^{\text{dom}})^{-1} P_{tT}^{\text{for}} X_t$. This relation between domestic and foreign interest rates, the FX spot and the FX forward is called *covered interest rate parity*.

An FX call option gives the holder the right to buy one unit of the foreign currency for the domestic strike amount K at maturity T. The payment can be written as $(X_T - K)_+$ in the domestic currency. Using $(X_T - K)_+ = X_T K(1/K - 1/X_T)_+$ we see that this is equivalent with a payment of $K(1/K - 1/X_T)_+$ in the foreign currency. As $1/X$ is the inverse FX rate, the same payment can be obtained by holding K number of FX put options in the foreign currency. Through a static replication argument, it follows that today's value of K FX put options $V_{1/K}^{\text{P}}(1/X)$ in the foreign currency equals an FX call option $V_K^{\text{C}}(X)$ in the domestic currency. By converting to the domestic currency we conclude that for FX options we not only have the standard put-call parity, but also the parity relation

$$V_K^{\text{C}}(X) = KXV_{1/K}^{\text{P}}(1/X)$$

Using put-call parity, this relation can be rewritten as

$$V_K^{\text{C}}(X) - KXV_{1/K}^{\text{C}}(1/X) = P_{tT}^{\text{for}} X - P_{tT}^{\text{dom}} K$$

If allowed to exercise early, it is clear that holders of the above call and put options would exercise simultaneously. The parity relation between puts and calls described above therefore holds for American options as well. This is in contrast to the standard put-call parity relation which does not hold for American options, see Sect. 2.5.

For another example of an FX product for which static replication is important, consider a digital FX option that pays one unit of the domestic currency if the foreign currency is worth more than the domestic strike K at maturity. Viewed in terms of the domestic currency, the payment can be written as $V_T = \theta(X_T - K)$. This contract can be priced as in Sect. 14.2. Here we are instead interested in the twist when the payment $\theta(X_T - K)$ is made in the foreign currency. This is an example of a quanto product, i.e. the payment is made in a non-natural currency for the product. Writing the payment in the domestic currency $V_T = X_T\theta(X_T - K) = K\theta(X_T - K) + (X_T - K)_+$, we see that it can be statically replicated with K non-quanto FX digitals and one FX call option. In a similar way, following the discussion in Sect. 2.3, we see that $X_T(X_T - K)_+ = 2\int_K^\infty (X_T - K')_+ dK' + K(X_T - K)_+$ which shows how an FX call quanto can be statically replicated with ordinary calls.

Please note that there is an alternative and simpler approach to price digital FX quanto options. Indeed, the payment can be written as $\theta(X_T - K) = \theta((1/K) - (1/X_T))$ in the foreign currency. Being a non-quanto digital FX option in the foreign currency it can be priced with the methods in Sect. 14.2 and then converted to the domestic currency by using today's FX rate. Combining the above results, we conclude that a digital FX option in one currency is equal to a sum of an FX digital and an FX call option in the other currency. It follows that an FX call (or put) option is equal to the difference of non-quanto digital FX options in the respective currencies of the FX rate. Indeed, a digital FX put option in the foreign currency with payment $\theta((1/K) - (1/X_T)) = \theta(X_T - K)$ is equivalent to the payment $X_T\theta(X_T - K)$ in the domestic currency. Adding $-K$ domestic FX digital options then gives the payment of an FX call option.

14.2 FX Options

An FX call (put) option is typically modeled in the domestic T-forward measure with numeraire P_{tT}^{dom}. The fundamental theorem of asset pricing states that the time t value of the option is given by

$$V_t = P_{tT}^{\text{dom}} E[(X_T - K)_+] = P_{tT}^{\text{dom}} E[(F_T - K)_+]$$

where we have used $P_{TT}^{\text{dom}} = 1 = P_{TT}^{\text{for}}$. As $F_t = P_{tT}^{\text{for}} X_t / P_{tT}^{\text{dom}}$ is the quotient of a domestic tradable (a product of a foreign tradable and the FX rate) and the numeraire, it must be a martingale. We therefore see that FX options can be priced with the methods used previously in the book. For instance, it is possible to use a lognormal process

$$dF_t = \sigma_t F_t dW_t$$

or any other driftless SDE to describe the evolution of the forward in the domestic T-forward measure.

For a consistent modeling of FX rates, the class of SDEs describing their evolution must be closed under the inverse relation and the cocycle relation. We now show that lognormal SDEs satisfy these constrains.

As X_t^{-1} is the inverse FX rate, $P_{tT}^{\text{dom}} X_t^{-1} / P_{tT}^{\text{for}} = F_t^{-1}$ is the forward of the domestic currency when priced in the foreign currency. It means that the forwards also satisfy the inverse relation. To verify that lognormal SDEs are closed under the inverse relation, we therefore need to prove that F_t^{-1} follows a lognormal process. Using Ito's lemma, we obtain

$$dF_t^{-1} = -\sigma_t F_t^{-1} dW_t + \sigma_t^2 F_t^{-1} dt$$

We would now like to transform this SDE, formulated in the domestic T-forward measure, to the foreign T-forward measure. As F_t^{-1} is a quotient of a foreign tradable and the foreign zero-coupon bond maturing at T, we actually know that the drift must be zero, resulting in

$$dF_t^{-1} = -\sigma_t F_t^{-1} d W_t^{\text{for}}$$

where W_t^{for} is a standard Brownian motion in the foreign T-forward measure. However, for the reader to gain familiarity with the techniques of measure changes in FX modeling, we perform the explicit computations.

By multiplying all tradables in the domestic currency with the FX rate X_t^{-1}, they can be viewed as tradables in the foreign currency. This multiplication is irrelevant for the modeling as we are only interested in the quotient with the numeraire which itself is a domestic tradable and has therefore also been multiplied by the FX rate. Thus, the FX multiplication cancels out in the quotient. After the multiplication, the numeraire has changed from P_{tT}^{dom} to $P_{tT}^{\text{dom}} X_t^{-1}$. The transformation to the measure with numeraire P_{tT}^{for} can then be done by using the Radon-Nikodym derivative

$$M_t = \frac{P_{tT}^{\text{for}}}{X_t^{-1} P_{tT}^{\text{dom}}} \frac{X_0^{-1} P_{0T}^{\text{dom}}}{P_{0T}^{\text{for}}} = F_t / F_0$$

according to the Appendix. Using Girsanov's theorem, it follows that

$$d W_t^{\text{for}} = d W_t - \langle d W_t, d \ln M_t \rangle = d W_t - \sigma dt$$

from which we obtain the driftless SDE for F_t^{-1} in the foreign T-forward measure.

We have proven that if F_t satisfies a lognormal SDE in the domestic T-forward measure, F_t^{-1} also satisfies a lognormal SDE, but in the foreign T-forward measure. It means that the class of driftless lognormal SDEs is closed under the inverse relation. To show that this class is closed under the cocycle relation as well, observe that

$$F^{ik} = P^i X^{ik} / P^k = P^i X^{ij} X^{jk} / P^k = (P^i X^{ij} / P^j)(P^j X^{jk} / P^k) = F^{ij} F^{jk}$$

which proves that the forwards fulfill the cocycle relation. Assuming lognormal SDEs for the forwards on the right-hand side:

$$\begin{cases} dF_t^{ij} &= \sigma_{ij} F_t^{ij} dW_t^{ij} \\ dF_t^{jk} &= \sigma_{jk} F_t^{jk} dW_t^{jk} \end{cases}$$

we obtain from Ito's lemma that

$$dF_t^{ik} = \sigma_{ik} F_t^{ik} dW_t^{ik}, \quad \sigma_{ik}^2 = \sigma_{ij}^2 + \sigma_{jk}^2 + 2\rho_{ij,jk}\sigma_{ij}\sigma_{jk}$$

We have chosen not to include the explicit computations for the drift as it is obviously zero in the natural measure of the forward. This can also be shown with a calculation similar to the one for the inverse relation. We have thereby concluded the proof that the class of lognormal SDEs is closed under both the inverse relation and the cocycle relation.

Observe that the FX parity relation $V_K^{\mathrm{C}}(X) = KXV_{1/K}^{\mathrm{P}}(1/X)$ is satisfied by the Black–Scholes FX model, i.e. when the forward is assumed to be lognormal. With similar arguments as in Sect. 3.6, we conclude that the implied volatility $\sigma_{\mathrm{imp}}(T, K)$ is equal to $\sigma_{\mathrm{imp}}(T, 1/K)$. We thereby see how volatility surfaces of reciprocal currency pairs are related.

14.3 Stochastic Volatility

For a consistent construction of a stochastic volatility model in FX, the inverse relation and the cocycle relation have to be fulfilled. We here give a simple example of a model that satisfies these constraints. The model we have in mind has the form

$$\begin{aligned} dF_t^{ij} &= \sigma_{ij} F_t^{ij} dW_t^{ij} \\ d\sigma_{ij} &= \epsilon\sigma_{ij} dZ_t \end{aligned}$$

which means that all the volatilities have the same driver and volatility of volatility. We furthermore assume that Z is uncorrelated to the W^{ij}s which means that we do not have any control over the skew.

The process is clearly closed under the inverse relation so we focus on the cocycle relation. As F_t^{ik} follows a lognormal process with volatility

$$\sigma_{ik} = \sqrt{\sigma_{ij}^2 + \sigma_{jk}^2 + 2\rho_{ij,jk}\sigma_{ij}\sigma_{jk}}$$

it remains to show that σ_{ik} follows the correct process. Applying the differential operator to both sides of the equation gives

$$
\begin{aligned}
d\sigma_{ik} &= \frac{\partial \sigma_{ik}}{\partial \sigma_{ij}} d\sigma_{ij} + \frac{\partial \sigma_{ik}}{\partial \sigma_{jk}} d\sigma_{jk} + \frac{1}{2}\frac{\partial^2 \sigma_{ik}}{\partial \sigma_{ij}^2}(d\sigma_{ij})^2 \\
&\quad + \frac{1}{2}\frac{\partial^2 \sigma_{ik}}{\partial \sigma_{jk}^2}(d\sigma_{jk})^2 + \frac{\partial^2 \sigma_{ik}}{\partial \sigma_{ij}\partial \sigma_{jk}} d\sigma_{ij} d\sigma_{jk} \\
&= \epsilon \left(\frac{\sigma_{ij} + \rho_{ij,jk}\sigma_{jk}}{\sigma_{ik}}\sigma_{ij} + \frac{\sigma_{jk} + \rho_{ij,jk}\sigma_{ij}}{\sigma_{ik}}\sigma_{jk} \right) dZ_t \\
&\quad + \frac{1}{2}\epsilon^2 \left(-\frac{(\sigma_{ij} + \rho_{ij,jk}\sigma_{jk})^2}{\sigma_{ik}^3}\sigma_{ij}^2 + \frac{\sigma_{ij}^2}{\sigma_{ik}} - \frac{(\sigma_{jk} + \rho_{ij,jk}\sigma_{ij})^2}{\sigma_{ik}^3}\sigma_{jk}^2 + \frac{\sigma_{jk}^2}{\sigma_{ik}} \right. \\
&\quad \left. - 2\frac{(\sigma_{ij} + \rho_{ij,jk}\sigma_{jk})(\sigma_{jk} + \rho_{ij,jk}\sigma_{ij})}{\sigma_{ik}^3}\sigma_{ij}\sigma_{jk} + 2\frac{\rho_{ij,jk}\sigma_{ij}\sigma_{jk}}{\sigma_{ik}} \right) dt \\
&= \epsilon \sigma_{ik} dZ_t
\end{aligned}
$$

14.4 Exotics

When pricing exotic options, it is popular to use the domestic terminal measure or the domestic risk-neutral measure. We choose to work in the latter and for concreteness assume a lognormal SDE for the FX rate:

$$dX_t = \mu_t^X X_t dt + \sigma_t^X X_t dW_t^X$$

As $B_t^{\text{for}} X_t / B_t^{\text{dom}}$ is a quotient of a domestic tradable and the numeraire, it must be a martingale. It therefore follows that

$$
\begin{aligned}
0 &= \left(d\frac{B_t^{\text{for}} X_t}{B_t^{\text{dom}}} \right)_{\text{drift}} = \left(r_t^{\text{for}} + \mu_t^X - r_t^{\text{dom}} \right) \frac{B_t^{\text{for}} X_t}{B_t^{\text{dom}}} dt \\
\Leftrightarrow \mu_t^X &= r_t^{\text{dom}} - r_t^{\text{for}}
\end{aligned}
$$

Thus, the no-arbitrage condition determines as usual the drift of the SDE. Observe that FX modeling is similar to equity modeling if we interpret the foreign interest rate as a continuous dividend payment.

14.5 Modeling Foreign Underlyings

A call option on a foreign underlying S gives the holder the right to purchase the underlying at maturity T for the domestic strike amount K. The payment at T can be written as $(S_T X_T - K)_+ = ((P_{TT}^{\text{dom}})^{-1} S_T X_T - K)_+$ in the domestic currency.

As $S_t X_t$ is a domestic tradable, we conclude that $(P_{tT}^{\text{dom}})^{-1} S_t X_t$ is a martingale in the domestic forward measure. This expression equals $(P_{tT}^{\text{dom}})^{-1} P_{tT}^{\text{for}} X_t (P_{tT}^{\text{for}})^{-1} S_t$ which is the product of the FX forward and the forward of the foreign underlying. Since the volatilities of these forwards can be derived from vanilla option quotes, we obtain the volatility of $(P_{tT}^{\text{dom}})^{-1} S_t X_t$ from which the option can be priced.

To model exotics, assume for simplicity that the foreign tradable S follows a lognormal SDE in the domestic risk-neutral measure:

$$dS_t = \mu_t^S S_t dt + \sigma_t^S S_t dW_t^S$$

The calibration of the volatility is a single currency affair as it can be done in a measure of the foreign currency. To determine the drift, use the fact that $S_t X_t / B_t^{\text{dom}}$ is a martingale in the risk-neutral measure. With a lognormal process for the FX rate, we obtain

$$0 = \left(d \frac{S_t X_t}{B_t^{\text{dom}}} \right)_{\text{drift}} = (\mu_t^S + r_t^{\text{dom}} - r_t^{\text{for}} + \rho_t \sigma_t^S \sigma_t^X - r_t^{\text{dom}}) \frac{S_t X_t}{B_t^{\text{dom}}} dt$$
$$\Leftrightarrow \mu_t^S = r_t^{\text{for}} - \rho \sigma_t^S \sigma_t^X$$

The SDE for the foreign underlying is then known from which exotic underlyings can be priced, for example, by simulation.

Let us now focus on non-tradable foreign underlyings. Because of their importance, we consider interest rates. For illustration, we assume that the foreign interest rates follow the SDE

$$df_{tT}^{\text{for}} = \sigma_{tT} dW_t^{\text{for}}$$

where W^{for} is a Brownian motion in the foreign T-forward measure. This process was analyzed in detail in Chap. 13. Recall that the drift must vanish as we have assumed that we are working in the natural measure of the rate.

By multiplying all the foreign tradables with the FX rate, we can view them as valued in the domestic currency. The above SDE is then in the measure of the numeraire $X_t P_{tT}^{\text{for}}$. The Radon-Nikodym derivative for the transformation to the domestic risk-neutral measure is given by

$$M_t = \frac{B_t^{\text{dom}}}{X_t P_{tT}^{\text{for}} / X_0 P_{0T}^{\text{for}}}$$

Disregarding the contribution from the drift gives

$$d \ln M_t \sim -d \ln P_{tT}^{\text{for}} - d \ln X_t \sim \left(\int_t^T \sigma_{ts} ds \right) dW_t^{\text{for}} - \sigma_t^X dW_t^{\text{X}}$$

We then obtain

$$dW_t^{\text{for}} = dW_t + \langle dW_t^{\text{for}}, d \ln M_t \rangle = dW_t + \left(\int_t^T \sigma_{ts} ds \right) dt - \rho_t \sigma_t^X dt$$

which gives the SDE in the domestic risk-neutral measure:

$$df_{tT}^{\text{for}} = \sigma_{tT} dW_t + \left(\int_t^T \sigma_{ts} ds - \rho_t \sigma_t^X \right) \sigma_{ts} dt$$

This equation is valid for rate-dependent volatility as well and it is relatively straightforward to extend it to multiple drivers and to simple compounded rates as in the LIBOR market model.

We end this section by considering a commodity futures contract quoted in a foreign currency. Assuming a lognormal model, we have

$$dF_{tT} = \sigma_{tT} F_{tT} dW_t^{\text{for}}$$

where W_t^{for} now denotes a Brownian motion in the foreign risk-neutral measure. Viewed in the domestic currency, the numeraire is $X_t B_t^{\text{for}}$ which gives the Radon-Nikodym derivative

$$M_t = \frac{B_t^{\text{dom}}}{X_t B_t^{\text{for}} / X_0}$$

for the transformation to the domestic risk-neutral measure. We then obtain the following SDE for the futures values

$$dF_{tT} = \sigma_{tT} F_{tT} dW_t^{\text{for}} - \rho_t \sigma_{tT} \sigma_t^X F_{tT} dt$$

14.6 Quantos

A *quanto* contains a payment in a non-natural currency for the product. An example is given by a product that pays the quote of a foreign stock in the domestic currency instead of in the foreign currency. This allows an investor to take a view on the level of the stock value without additional FX exposure. For example, it enables a European speculator to follow trading advice of American media without being exposed to fluctuations in EURUSD. We do not present an extensive list of possible quanto products, but rather consider two examples after which we believe the reader is able to model the most general quanto.

For the first example, let S be a domestic tradable and let the quanto have the payment S_T in a foreign currency at time T. Viewed in terms of the domestic currency, the time T value of the quanto is $V_T = S_T X_T$. Using lognormal dynamics of the underlying and the FX rate, the dynamics of $S_t X_t$ are given by

$$\begin{aligned} d(S_t X_t) &= \left(2r_t^{\text{dom}} - r_t^{\text{for}} + \rho \sigma_t^S \sigma_t^X\right) S_t X_t dt \\ &\quad + \left(\left(\sigma_t^S\right)^2 + \left(\sigma_t^X\right)^2 + 2\rho \sigma_t^S \sigma_t^X \right)^{1/2} S_t X_t dW_t \end{aligned}$$

from which the quanto can be priced.

For the second example, consider a quanto caplet paying $(L^{\text{for}}_{tTT'} - K)_+$ at maturity T', where the strike K and the payment are in the domestic currency. This product can be priced by using the evaluation of $L^{\text{for}}_{tTT'}$ in the domestic T'-forward measure which can be obtained from the SDE for df^{for}_{tT} that was derived in the previous section.

14.7 Volatilities and Correlations

Just as for equities, the volatility of the forward can be backed out from quotes on European call options. The forward volatility can then be converted to a volatility for the underlying X by using the relation $F_t = (P^{\text{dom}}_{tT})^{-1} P^{\text{for}}_{tT} X_t$.

When pricing higher-dimensional derivatives it is sometimes necessary to include the correlation $\rho_{ij,jk}$ between FX rates X^{ij} and X^{jk}. Using lognormal dynamics, we derived the relation

$$\sigma^2_{ik} = \sigma^2_{ij} + \sigma^2_{jk} + 2\rho_{ij,jk}\sigma_{ij}\sigma_{jk}$$

in Sect. 14.2. As the volatilities can be calibrated to liquid market quotes, the value of the correlation can be backed out. This value is referred to as the *implied correlation*.

By doing some algebraic manipulations, we find that the correlations are constrained by

$$\begin{aligned}&\sqrt{1-\rho^2_{ij,jk}}\sqrt{1-\rho^2_{jk,ki}}\sqrt{1-\rho^2_{ki,ij}}\\&\quad= \rho_{ij,jk}\rho_{jk,ki}\sqrt{1-\rho^2_{ki,ij}} + \rho_{jk,ki}\rho_{ki,ij}\sqrt{1-\rho^2_{ij,jk}} + \rho_{ki,ij}\rho_{ij,jk}\sqrt{1-\rho^2_{jk,ki}}\end{aligned}$$

To find the implied correlations $\rho_{ij,kl}$ for $j \neq k$, we use the identity $F^{kl} = F^{kj}F^{jl}$ to conclude that $\sigma_{kl}dW^{kl}$ is equal to $\sigma_{kj}dW^{kj} + \sigma_{jl}dW^{jl}$ up to a drift term. We arrive at

$$\begin{aligned}\rho_{ij,kl}\sigma_{kl} &= \rho_{ij,kj}\sigma_{kj} + \rho_{ij,jl}\sigma_{jl}\\ \Leftrightarrow\ 2\rho_{ij,kl}\sigma_{ij}\sigma_{kl} &= -2\rho_{ij,jk}\sigma_{ij}\sigma_{jk} + 2\rho_{ij,jl}\sigma_{ij}\sigma_{jl}\\ &= -\sigma^2_{ik} + \sigma^2_{ij} + \sigma^2_{jk} + \sigma^2_{il} - \sigma^2_{ij} - \sigma^2_{jl}\\ &= \sigma^2_{il} - \sigma^2_{ik} + \sigma^2_{jk} - \sigma^2_{jl}\end{aligned}$$

In this case it can be shown that the correlations are constrained by the relation

$$\begin{aligned}&\rho_{ij,kl}\rho_{ik,lj}\rho_{il,jk} + \rho_{ik,lj}\rho_{il,lk}\rho_{ij,jk} + \rho_{il,jk}\rho_{ij,jl}\rho_{ik,kl} + \rho_{ij,kl}\rho_{ik,kj}\rho_{il,lj}\\&\quad - \rho_{ik,kj}\rho_{il,lk}\rho_{ij,jl} + \rho_{ij,jk}\rho_{ik,kl}\rho_{il,lj} = 0\end{aligned}$$

14.8 Volatility Interpolation

The FX market is global and trading takes place 24 hours a day. However, when a local center, e.g. Tokyo for trades in yen, is closed, the trading activity decreases for that particular currency. The outcome is that the local volatility decreases by a certain factor on such days. To be able to calibrate to a given implied volatility, the local volatility therefore has to be increased on business days. This effect is important as options on currency pairs are often found to be liquid with tight bid-offer spreads. The impact is particularly visible for short-dated options.

The factor by which the volatility should be multiplied depends on whether the local trading center is closed because of a weekend or a holiday. It also depends on if any of the major trading centers London, New York and Tokyo are closed. Certain days need to be weighted with a factor larger than 1, indicating a higher trading activity than normal. An example is FX options involving SEK that had a period extending over September 14, 2003, the day of the Swedish referendum on the euro.

14.9 Numeraire

When dealing with foreign exchange, there are $N(N-1)/2$ currency pairs to keep track of for N currencies. For N large, this number can get uncontrollable in practice as we need to keep track of the spot, the volatility and other variables for all currency pairs. An alternative is to use the currencies themselves as fundamental objects rather than the currency pairs. For this approach to succeed, it is necessary to find something that the currencies can be valued against, i.e. to find a suitable numeraire.

The numeraire should be a tradable product for which everyone in the world agrees upon the price, e.g. gold. A counterexample is given by natural gas or electricity which are expensive to transport and therefore differ in price between different regions of the world. The numeraire does not have to be a commodity but could, for example, be a US treasury bond. The beautiful thing about using a numeraire for currencies is that it cancels out when computing quotients to obtain currency pairs. The choice of numeraire is therefore irrelevant and we can assume it to have the most well-behaved properties.

Let X^i denote the value of currency i in terms of the numeraire. As the currency pairs are obtained from $X^{ij} = X^i/X^j$, it is sufficient to keep track of the N variables $\{X^i\}$ instead of the $N(N-1)/2$ variables $\{X^{ij}\}_{i>j}$. Furthermore, the inverse relation and the cocycle relation are automatically satisfied as

$$X^{ji} = X^j/X^i = \left(X^i/X^j\right)^{-1} = \left(X^{ij}\right)^{-1}$$

$$X^{ij}X^{jk} = X^i/X^j \cdot X^j/X^k = X^i/X^k = X^{ik}$$

It means that we no longer need to keep track of any constraints.

The use of a numeraire can be helpful in constructing consistent FX models. Indeed, if we succeed in finding models for the X^is such that the process for X^{ij} does not depend on X^i or X^j separately, but only on the combination X^i/X^j, the model is closed under the inverse relation and the cocycle relation. Furthermore, the volatilities and correlations for $\{X^i\}$ can be chosen freely and are not subject to any constraints as is the case for $\{X^{ij}\}$, see Sect. 14.7.

In real life FX trading, the (money market account in) dollar can be considered as a numeraire. The reason is that for most currencies, the most liquid currency pair is the one with the dollar as the second currency. Quotes on currency pairs not involving the dollar are obtained by using the dollar as an intermediary. The dollar is then the domestic currency according to the convention used in the introduction to this chapter.

From a modeling point of view it might not be the best idea to use the dollar as a numeraire as this amounts to singling out one of the currencies. The advantages of using a currency numeraire, however, are that it is not necessary to involve any external asset class (such as commodities) and that there are now only $N-1$ instead of N variables to model.

14.10 Conventions

As the conventions used in foreign exchange markets can be confusing, we find it worth to summarizing them. The rules which we give below are valid in most cases and in particular for the major currencies. However, the reader should be aware of the fact that there exist many exceptions.

An FX rate is written as CCY1CCY2 or CCY1/CCY2, stating the price of one unit CCY1 in terms of CCY2. CCY1 is called the *base currency* and CCY2 the *terms currency* or *quoting currency*. The quotation is usually such that CCY2 is worth the least of the currencies, resulting in a rate larger than 1. The main exceptions are EUR, GBP, AUD, NZD, FJD, TOP, WST, PGK, BWP, SBD, USD, where the earliest currency in the list is CCY1. For example, a sterling-dollar quote (called the *cable* by practitioners) can have the form GBPUSD $= 1.56269$. There are many exceptions to this rule such as SEKPLN which currently is less than 1. These conventions apply in the *interbank market*. In local markets, on the other hand, the retail and commercial market participants prefer to see the domestic currency always as CCY1 or CCY2.

Most of the liquidly traded FX rates have the dollar as one of the currencies. When neither of the two currencies are the dollar, the FX is said to be a cross rate and the quote is often determined indirectly via the dollar. An example is given by JPYINR, for which the bid price can be found by dividing the bid price of USDINR with the offer price of USDJPY. Because of the particular importance of the euro, rates such as EURGBP are often not considered to be cross rates. Along the same lines, GBPCHF is often considered to be a euro cross rather than a dollar cross because of the high liquidity of EURCHF and EURGBP.

In FX markets, the date when cash changes hands is called the *value date*. The standard is that the value date is equal to the spot date, but it can be both before and after. Trades with value date before the spot date are most often done as *cash deals*, where the settlement date coincides with the trade date. Trades settled after the spot date are forward contracts.

The spot date is determined from the trade date in the following way: use a settlement lag of 2 business days and compute two spot dates using the calendars of the trading centers belonging to the constituting currencies. If one of the currencies is the dollar, use the non-dollar date, otherwise use the later of the dates. If the resulting date is a business day in both trading centers as well as a New York business day, the spot date has been found. If not, the spot date is the next day that is a business day in New York as well as in the trading centers belonging to the constituting currencies. For instance, the GBPUSD spot date is the business day in both New York and London that is as few London business days as possible (but not less than 2) after the trade date. Exceptions to this rule include USDCAD, USDRUB, USDTRY, EURRUB, EURTRY, CADRUB, CADTRY, TRYRUB, which have a settlement lag of 1 business day, and certain Middle Eastern and Latin American currencies. Furthermore, it is not necessary to require the spot date to be a New York business day for currency pairs that are not dollar crosses, e.g. EURSEK.

When defining payment dates for option and forward contracts in FX markets, the standard is to use the modified following holiday adjustment. The adjustment is made so that the resulting payment date is a business day in New York as well as in the trading centers of the constituting currencies. Furthermore, the end-of-month rule is used. The expiry and delivery dates are determined as in Sect. 3.11 when the tenor is expressed as a number of months or years. For tenors which are written as a number of days or weeks, on the other hand, the expiry is obtained by adding the tenor to today's date. If the date obtained this way is a holiday, the following day adjustment is used conditioned on that the expiry should be a business day in both currencies. Observe that in this instance we do not require the expiry to be a US business day for cross rates. The delivery date is the spot date calculated from the expiry.

The remainder of the section surveys the various option conventions used in FX markets. For this purpose, we recall the Black–Scholes formula and the delta for a call option:

$$V_t = P_{tT}^{\mathrm{for}} X_t N(d_+) - P_{tT}^{\mathrm{dom}} K N(d_-)\,, \quad d_\pm = \frac{\ln(P_{tT}^{\mathrm{for}} X_t/(P_{tT}^{\mathrm{dom}} K))}{\sigma\sqrt{T-t}} \pm \frac{1}{2}\sigma\sqrt{T-t}$$

$$\Delta^{\mathrm{C}} = P_{tT}^{\mathrm{for}} N\left(d_+\right)$$

Because of the 1-1 correspondence between the strike and the delta, it is possible to use the latter to define the moneyness of an option. Indeed, it is standard to quote implied volatilities in terms of deltas for FX markets, as opposed to equities, interest rates and commodities markets for which strikes are used. FX options are quoted as the ATM implied volatility and the 25% risk reversal and strangle. For options on liquid currency pairs, the 10% risk reversal and strangle are also given. They give the traders access to the tail behavior of the probability distribution.

The 25% *risk reversal* is related to implied volatilities via the relation

$$\sigma\left(\Delta^{\mathrm{C}}=25\%\right)-\sigma\left(\Delta^{\mathrm{P}}=-25\%\right)=\mathrm{RR}_{25\%}$$

The 25% *strangle*, on the other hand, is related to option prices through

$$\begin{aligned}&V^{\mathrm{C}}\left(\tilde{\Delta}^{\mathrm{C}}=25\%\right)+V^{\mathrm{P}}\left(\tilde{\Delta}^{\mathrm{P}}=-25\%\right)\\&\quad=\mathrm{BS}^{\mathrm{C}}\left(\tilde{\Delta}^{\mathrm{C}}=25\%,\sigma_{\mathrm{ATM}}+\mathrm{STR}_{25\%}\right)+\mathrm{BS}^{\mathrm{P}}\left(\tilde{\Delta}^{\mathrm{P}}=-25\%,\sigma_{\mathrm{ATM}}+\mathrm{STR}_{25\%}\right)\end{aligned}$$

where the Black–Scholes formula has been used on the right-hand side for the computation of the option price. $\tilde{\Delta}$ is related to the strike through the Black–Scholes formula using a volatility of $\sigma_{\mathrm{ATM}}+\mathrm{STR}_{25\%}$. This is fundamentally different from Δ that is related to the strike via the market volatility. For instance, the strike corresponding to $\Delta^{\mathrm{C}}=25\%$ is determined through the Black–Scholes formula using the volatility $\sigma\left(\Delta^{\mathrm{C}}=25\%\right)$. The same relations apply to the 10% risk reversal and strangle.

Because two different types of deltas are used, the volatility smile needs to be backed out numerically from the above relations. Fortunately, it can be shown (Reiswich and Wystup (2009)) that for small values of the risk reversal, the following approximation works well:

$$\frac{1}{2}\left(\sigma\left(\Delta^{\mathrm{C}}=25\%\right)+\sigma\left(\Delta^{\mathrm{P}}=-25\%\right)\right)=\sigma_{\mathrm{ATM}}+\mathrm{STR}_{25\%}$$

from which it is possible to back out the implied volatilities according to

$$\sigma\left(\Delta^{\mathrm{C}}=25\%\right)=\sigma_{\mathrm{ATM}}+\frac{1}{2}\mathrm{RR}_{25\%}+\mathrm{STR}_{25\%}$$

$$\sigma\left(\Delta^{\mathrm{P}}=25\%\right)=\sigma_{\mathrm{ATM}}-\frac{1}{2}\mathrm{RR}_{25\%}+\mathrm{STR}_{25\%}$$

The strangle usually only varies slowly with time while the risk reversal is stochastic and has a relatively high correlation with the spot. The 10% risk reversal and strangle are in general tightly linked to their 25% counterparts.

The deltas appearing above are not necessarily equal to the spot delta $\frac{dV}{dX}$, but several different versions are used depending on the currency pair and on the tenor of the option. For instance, it is common to consider the *forward delta*, defined as the number of forward contracts necessary to hedge an option position:

$$\Delta_f^{\mathrm{C}}=\frac{dV}{d(X_tP_{tT}^{\mathrm{for}}-KP_{tT}^{\mathrm{dom}})}=\left(P_{tT}^{\mathrm{for}}\right)^{-1}\Delta^{\mathrm{C}}=N(d_+)$$

Black–Scholes formula gives the option price in the domestic currency. It is also common to state the price in the foreign currency. It is then equal to V/S. The currency with respect to which the price is measured is called the *premium currency*.

Recall that the spot delta states the amount of the foreign currency that needs to be purchased to hedge the option position. If the option price is paid in the foreign currency, we already have the amount V/X in that currency. The remaining amount $\Delta_{\text{adj}} = \Delta - V/X$ that needs to be hedged is called the *premium adjusted delta*. We also see that the number of forwards contracts necessary for the hedge should be reduced by $\left(P_{tT}^{\text{for}}\right)^{-1} V/X$, resulting in $\Delta_{\text{f, adj}} = \left(P_{tT}^{\text{for}}\right)^{-1} (\Delta - V/X)$.

The premium currency coincides most often with the base currency. The main exception is USD which is always the premium currency. It means, for example, that adjusted deltas are used for EURGBP but not for EURUSD.

The forward delta is useful when taking the interest rate risk into account in the hedge. It is therefore used for long maturities or for large interest rate differentials between the currencies. As a general rule, the spot delta is only used when both currencies belong to USD, EUR, JPY, GBP, AUD, NZD, CAD, CHF, NOK, SEK, DKK and the maturity is less than or equal to 1Y. The forward delta is used otherwise.

The delta convention is not only used in the definition of the risk reversal and the strangle, but also for the ATM point. Indeed, the ATM definition $\Delta^{\text{C}} = -\Delta^{\text{P}}$ is often used for short-dated FX options on liquidly traded currency pairs. The ATM definition $K = F$ is used otherwise.

14.11 FX Swaps and Cross Currency Swaps

As financial institutions find it easier and cheaper to borrow in the domestic market than in foreign markets, they often follow a 2-step procedure to raise foreign cash. A loan is first taken in the domestic currency after which it is converted to a synthetic loan in a foreign currency by using an *FX swap*. An FX swap consists of a spot exchange of currencies followed by a forward exchange in the opposite direction. The spot and the forward amount are equal for one of the currencies. The spot amount for the other currency is obtained from the spot FX rate while the forward amount is such that the swap prices at par, which means that the amount can be computed from the forward FX rate.

FX swaps, or equivalently, FX forwards, are usually liquidly traded out to 1Y. The natural extension to longer maturities is *cross currency swaps*. Because of their long maturities, the interest rate effect is more important than the FX effect for these products. Cross currency swaps are for this reason often handled by the IR desk and not by the FX desk. The most popular type of cross currency swaps exchanges 3M USD LIBOR for 3M LIBOR in another currency. A spread is added to the non-USD leg so that the swap is valued to par. At the start of the swap the notionals of the legs are related through the FX rate and are exchanged.

At each coupon payment date the notional of the leg paying LIBOR flat is reset according to the prevailing FX rate. For instance, assume that counterparty A enters a cross currency swap with counterparty B where JPY 3M LIBOR – 20 bp is exchanged for USD 3M LIBOR and 1 *basis point* (bp) is defined as 0.01%. Counterparty B then has to pay counterparty A the JPY notional N^{JPY} at spot in return for the dollar amount $N_0^{\text{USD}} = N^{\text{JPY}}/\text{FX}_0$, where FX_0 is the USDJPY spot FX rate. After 3M time counterparty A pays (JPY 3M LIBOR - 20bp)$\delta^{\text{JPY}} N^{\text{JPY}}$ and receives (USD 3M LIBOR)$\delta^{\text{USD}} N_0^{\text{USD}}$. Furthermore, A receives $N_0^{\text{USD}} - N_{3\text{M}}^{\text{USD}} = N_0^{\text{USD}} - N^{\text{JPY}}/\text{FX}_{3\text{M}}$ (or has to make a payment if the amount is negative). The same procedure repeats itself after 6M, 9M, etc.

The financial literature is almost exclusively concerned with a different type of cross currency swap for which the notionals are only exchanged at initialization and at the final payment date. As all our arguments apply to these products as well and because they are not as commonly traded as the cross currency swaps that we described above, we omit any further discussion of this type of cross currency swaps.

14.12 Credit and Liquidity Risk

We describe how certain FX relations break down in the presence of credit risk, supply and demand, liquidity, etc. The discussions can be viewed as a continuation of the end of Chap. 13, in which the corresponding pure interest rate effect was considered. We start by considering collateralized FX forwards and revisit covered interest rate parity. With the new assumptions, the replication of forwards does not succeed as there are credit and liquidity premia embedded into the loans.

Attempts at improving the replication are unsuccessful as well. For instance, instead of taking loans covering the full tenor of the forward contract, it is possible to reinvest using overnight loans. The proceeds can be locked in using OIS. There are several reasons why this replication strategy is not exact. First of all, there is still a credit risk on the loans, even though it has been reduced to overnight risk. Secondly, there is a risk of default for the OIS counterparties. There is also a risk of default for the counterparty of the FX forward, with an associated replacement risk of the FX position. This risk is particularly big as the whole notional is involved (and not just coupon payments as for single currency swaps). Finally, the overnight rates used in the OISs are based on fixings from banks without commitment to trading. The consequence is that the proceeds of the overnight reinvestment strategies do not exactly cancel out the floating legs of the OISs.

Despite the fact that a strategy of using OISs does not exactly succeed in replicating FX forwards, it can be relatively close. It means that it makes sense to let the FX forward curve inherit features of the OIS curves such as jumps at monetary policy meeting dates and the possible existence of (positive and negative) spikes.

During the credit crisis that started 2007, the breakdown of covered interest rate parity became particularly obvious due to dollar shortage. European banks had problems obtaining unsecured dollar loans while US banks were reluctant to lend dollars. To raise dollars, the European banks took loans in European currencies and used FX swaps or cross currency swaps to convert them into dollars. This created a one-sided pressure on the FX forward (and cross currency swap) markets which led to a theoretical arbitrage that could not be picked up for the reasons described above.

The effects discussed here have an interesting impact on the pricing of cross currency swaps. To understand this thoroughly, we need a more detailed understanding of the contract specifications. Cross currency swaps (and not too short-dated FX forwards) are most often traded collateralized with a choice of collateral from several currencies. The receiver of the collateral has to pay the overnight interest rate for the currency that was posted as collateral. Because of the reasons mentioned above, the posting of collateral in one currency cannot replicate the posting of collateral in a different currency. Thus, the price of a cross currency swap depends on the choice of collateral currency. The consequence is that the price of a cross currency swap should be computed using the cheapest choice of collateral currency. It means that the discounting in a cross currency swap should be done using the OIS curve corresponding to the cheapest collateral currency and not to the native currencies. Observe that this argument is not restricted to cross currency swaps as CSA agreements that allow for collateral from a choice of currencies can be found for various product types, such as single currency swaps.

Consider a cross currency swap between CCY1 and CCY2, and assume that CCY2 is the cheapest collateral. CCY2 cash flows should obviously be discounted with the CCY2 OIS curve. CCY1 cash flows, on the other hand, need to be discounted taking into account that CCY2 is the collateral. The result is a CCY1 discount curve that prices cross currency swaps to par and is different from the CCY1 OIS curve. The consequence is that we for each currency end up with several discount curves, depending on the allowed collateral.

The currency that is cheapest to post as collateral today might not coincide with the cheapest collateral in the future. By using forward curves, it is possible to account for deterministic changes in the cheapest collateral. The complexity in the pricing can be increased one more step by taking into account stochastic changes in the choice of the cheapest currency. The result is the inclusion of a type of option premium in the cross currency swap price.

There is currently a controversy as to whether the pricing actually should include the possibility to replace collateral during the lifetime of a cross currency swap as there have been disputes in the situation when one of the counterparties has asked for such a replacement. The pricing of CSA regulated contracts gets even more complicated as it is common to not only allow currencies as collateral but also other assets, e.g. government bonds.

Bibliography

Reiswich D, Wystup U (2009) FX volatility smile construction. Center for Practical Quantitative Finance, Frankfurt School of Finance & Management, Research Report No. 20

Appendix A
Mathematical Preliminaries

A.1 Measure Theory, Random Variables and Integration

A σ-algebra $\mathcal{F}$ over a set Ω is a non-empty collection of subsets that contains the empty set $\varnothing$ and is closed under the operations of taking complements and countable unions. An example is the Borel σ-algebra defined as the smallest σ-algebra containing the open sets in a topological space. A measure ν is a function $\mathcal{F} \to \mathbb{R}_+$ that satisfies $\nu(\varnothing) = 0$ and is σ-additive: $\nu(\cup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \nu(A_i)$ for $\{A_i \in \mathcal{F}\}$ pairwise disjoint as subsets of Ω. A measure P satisfying $P(\Omega) = 1$ is called a probability measure.

$X : \Omega \to \Omega'$ is said to be measurable (with respect to the σ-algebras $\mathcal{F}$ and $\mathcal{F}'$) if $X^{-1}(A') \in \mathcal{F}$ for every $A' \in \mathcal{F}'$. If the measure on $\mathcal{F}$ is a probability measure then a measurable function is referred to as a random variable. We are only concerned with the case when $\Omega' = \mathbb{R}^n$ and $\mathcal{F}'$ is the corresponding Borel σ-algebra. Unless stated otherwise, we assume that $n = 1$ as the generalization to arbitrary n is often straightforward. An example of measurable functions is given by the simple functions that by definition can be written as $\sum_{i=1}^{n} c_i \mathbb{1}_{A_i}$ where $c_i \in \mathbb{R}$, $A_i \in \mathcal{F}$ and $\mathbb{1}_A$ is the indicator function with the property that $\mathbb{1}_A(\omega)$ is equal to 1 if $\omega \in A$ and 0 otherwise.

The integral of a positive simple function is given by $\sum_{i=1}^{n} c_i P(A_i)$. The integral $\int_\Omega X dP$ of a positive random variable is defined in the Lebesgue sense as the supremum of integrals over simple functions smaller than X. If $\int_\Omega |X| dP < \infty$ then X is said to be P-integrable. The integral is then defined as the difference between the integrals over the positive part and the negative part.

For a random variable X, the probability density function (PDF) $p = p_X : \mathbb{R} \to \mathbb{R}_+$ is defined by $p(x)dx = P(X^{-1}([x, x + dx]))$, see Fig. A.1. The expectation $E[X]$ of X is by definition the integral $\int_\Omega X dP$ which by a change of integration variable can be expressed with the PDF as $\int_\mathbb{R} x p(x) dx$. We are often not interested in the set Ω of events but only in the probabilities for a random variable to attain its various values. This is exactly the information contained by the PDF or the cumulative density function (CDF) $F(x) = \int^x p(x')dx'$.

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2,

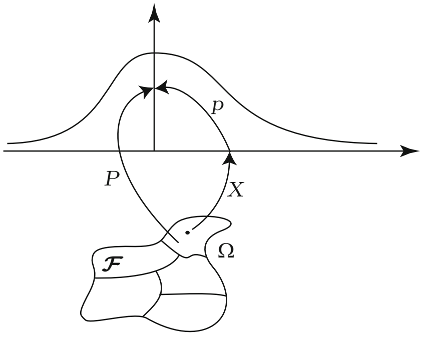

Fig. A.1 Relation between a random variable and its PDF

A.2 The Gaussian Distribution

A normally distributed $\mathcal{N}(\mu,\sigma^2)$ variable X is a random variable with $p(x) = \exp((x-\mu)^2/2\sigma^2)/\sqrt{2\pi}\sigma$. A standard normal random variable has $\mu = 0$ and $\sigma = 1$. The normal distribution is also referred to as the Gaussian distribution. In higher dimensions, it is defined by

$$p(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ is symmetric and positive semi-definite. The characteristic function

$$E\left[e^{ik^TX}\right] = \exp\left(ik^T\mu - \frac{1}{2}k^T\Sigma k\right), \quad k \in \mathbb{R}^n$$

can be computed by diagonalizing Σ. The moments can then be calculated through

$$E\left[X^j\right] = \frac{d^j}{i^j\,dp^j} E\left[e^{ik^TX}\right]\bigg|_{k=0}$$

and we see in particular that the Gaussian distribution is completely determined by its expectation $E[X] = \mu$ and covariance $\mathrm{Covar}(X) = \Sigma$.

$X = \{X_i\}$ being a Gaussian is equivalent with λ^TX being a 1-dimensional Gaussian for an arbitrary vector λ. Indeed, if X is a Gaussian then

$$E\left[e^{ik\lambda^TX}\right] = \exp\left(ik\lambda^T\mu - \frac{1}{2}k\lambda^T\Sigma\lambda k\right), \quad k \in \mathbb{R}$$

which shows that λ^TX is a Gaussian with mean $\lambda^T\mu$ and variance $\lambda^T\Sigma\lambda$. Conversely, if λ^TX is a Gaussian then

$$E\left[e^{ik\lambda^T X}\right] = \exp\left(ikm - \frac{1}{2}k^2 s\right), \quad k \in \mathbb{R}$$

where

$$m = E[\lambda^T X] = \lambda^T \mu$$
$$s = \mathrm{Var}(\lambda^T X) = \lambda^T \Sigma \lambda$$

Inserting this in the above equation with $k = 1$ shows that X is Gaussian.

The covariance is a diagonal matrix if $\{X_i\}$ is a Gaussian with independent components. The converse is also true as

$$E\left[e^{ik^T X}\right] = e^{ik^T\mu - \frac{1}{2}k^T \Sigma k} = \prod_j e^{ik_j\mu_j - \frac{1}{2}\Sigma_{jj}k_j^2} = \prod_j E\left[e^{ik_j X_j}\right]$$

and the characteristic function can be written in product form if and only if the random variable has independent components.

A.3 Copulas

A function $\mathcal{C} : [0,1]^n \to [0,1]$ satisfying:

- $\mathcal{C}(u) = 0$ if $\exists k; u_k = 0$
- $\mathcal{C}(u) = u_j$ if $u_k = 1, \forall k \neq j$
- $\mathcal{C}$ is n-increasing

is called a copula. The n-increasing property means that the weighted sum of the copula evaluated on the vertices of an arbitrary n-dimensional rectangle must be positive, where the weight is equal to 1 if there is an even number of lower points in the rectangle and –1 if there is an odd number. For example, in two dimensions the condition reads:

- $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0, \forall u_2 \geq u_1, v_2 \geq v_1$

Sklar's theorem. *Let $F_X(x_1, x_2, \ldots, x_n) = P(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n)$ be the cumulative density function for a random variable $X = \{X_i\}_{i=1}^n$ and $F_{X_i}(x_i) = P(X_i < x_i)$ the marginal functions. Then there exists a copula $\mathcal{C}$ such that*

$$F_X(x_1, x_2 \ldots, x_n) = \mathcal{C}(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_n}(x_n))$$

The copula is unique if the marginal functions are continuous. Conversely, for $\mathcal{C}$ a copula and $\{F_{X_i}\}_i$ cumulative density functions, $F_X(x_1, x_2 \ldots, x_n)$ defined as above is a multivariate cumulative density functions with marginal distributions $\{F_{X_i}\}_i$.

The theorem can be proven by defining the copula as

$$\mathcal{C}(u_1, u_2, \dots, u_n) = F_X\left(F_{X_1}^{-1}(x_1), F_{X_2}^{-1}(x_2), \dots, F_{X_n}^{-1}(x_n)\right)$$

The statement then follows by identifying the three defining conditions for a copula with corresponding necessary conditions for a multivariate cumulative density function. □

Copulas are commonly used for the problem of constructing a higher-dimensional random variable from a set of marginal distributions. Perhaps the most popular copula for this purpose is the Gaussian copula. This is the copula that relates a Gaussian random variable to its marginal distributions. For example, in two dimensions we use the marginal distributions

$$F_{X_i}(x_i) = N(x_i) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i} e^{-z^2/2} dz$$

and the joint distribution

$$\begin{aligned} &P(X_1 < x_1, X_2 < x_2) \\ &\quad = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \exp\left(-\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)/2\left(1-\rho^2\right)\right) dz_1 dz_2 \end{aligned}$$

to obtain the form

$$\begin{aligned} &\mathcal{C}(u_1, u_2) \\ &\quad = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{N^{-1}(u_1)} \int_{-\infty}^{N^{-1}(u_2)} \exp\left(-\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)/2\left(1-\rho^2\right)\right) dz_1 dz_2 \end{aligned}$$

of the copula.

Many of the popular copulas belong to the class of Archimedean copulas defined by

$$\mathcal{C}(u_1, u_2, \dots, u_n) = \phi^{[-1]}\left(\sum_{i=1}^{n} \phi(u_i)\right)$$

where $\phi : [0, 1] \to [0, \infty]$ is a continuous strictly decreasing convex function with $\phi(1) = 0$. $\phi^{[-1]}(z)$ is the "pseudo-inverse" defined as $\phi^{-1}(z)$ for $z \in [0, \phi(0)]$ and 0 for $z \geq \phi(0)$. An example is the Clayton copula defined by

$$\phi(u) = u^{-\theta} - 1, \quad \theta > 0$$

An important property of the Clayton copula and the Gaussian copula is that they can interpolate between the instances of complete independence and complete dependence of the components. For example, in the 2-dimensional case, $\theta = 0$ and $\rho = 0$ gives independence: $F_X(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$, while $\theta \to \infty$ and $\rho = 1$ gives complete dependency $F_X(x_1, x_2) = \min(F_{X_1}(x_1), F_{X_2}(x_2))$.

A.4 Processes

A set of σ-algebras $\{\mathcal{F}_i\}_{i\in I}$ is called a filtration of $\mathcal{F}$ if every set in $\mathcal{F}_i$ also belongs to $\mathcal{F}_j$ and $\mathcal{F}$, for $j \geq i$, where I is the positive integers or the positive real numbers. $\{X_i\}$ is called a (adapted) process if for every i, X_i is a random variable that is $\mathcal{F}_i$-measurable (for a fixed image σ-algebra $\mathcal{F}'$). We assume that the subsets of null sets (sets with measure zero) in $\mathcal{F}$ are included in $\mathcal{F}_0$ (and therefore in $\mathcal{F}_i$ for all i). When $I = \mathbb{R}_+$ we assume that the filtration is right continuous, $\mathcal{F}_i = \cap_{j>i}\mathcal{F}_j$ for all i. Furthermore, for $I = \mathbb{R}_+$ we say that a process is continuous if the paths $X(\omega) : \mathbb{R}_+ \ni i \mapsto X_i(\omega) \in \mathbb{R}$ are continuous almost surely (a.s.), i.e. for all ω in the complement of a null set. A process is sometimes referred to as a stochastic process to emphasize the dependence on the set of outcomes Ω. A process that only depends on the index i and not on $\omega \in \Omega$ is said to be non-stochastic.

Ex: Let $\Omega = \{\text{hh}, \text{ht}, \text{th}, \text{tt}\}$ be the set of outcomes after tossing a coin twice and $\{\mathcal{F}_i\}_{i=0,1,2}$ the filtration such that $\mathcal{F}_i$ is the σ-algebra of information available after the i:th toss, e.g. $\mathcal{F}_1 = \{\emptyset, \{\text{hh}, \text{ht}\}, \{\text{th}, \text{tt}\}, \{\text{hh}, \text{ht}, \text{th}, \text{tt}\}\}$. If we make a bet on which we win \$1 on heads and lose \$1 on tails, the process $\{X_i\}_{i=0,1,2}$ describes our earnings: $X_0(\text{hh}) = \$0$, $X_0(\text{ht}) = \$0$, $X_0(\text{th}) = \$0$, $X_0(\text{tt}) = \$0$; $X_1(\text{hh}) = \$1$, $X_1(\text{ht}) = \$1$, $X_1(\text{th}) = -\$1$, $X_1(\text{tt}) = -\$1$ and $X_2(\text{hh}) = \$2$, $X_2(\text{ht}) = \$0$, $X_2(\text{th}) = \$0$, $X_2(\text{tt}) = -\$2$.

A.5 Brownian Motion

A standard Brownian motion is a process W_t, $t \geq 0$ satisfying

- For each $s \geq 0, t > 0$, $W_{t+s} - W_s \sim \mathcal{N}(0, t)$
- $\{W_{t_{i+1}} - W_{t_i}\}$ are independent for given $0 \leq t_0 \leq \cdots \leq t_n$
- $W_0 = 0$
- W_t is continuous in t

There are several possible techniques of constructing new Brownian motions from a given one. For example, it is straightforward to show that the following processes satisfy the above conditions and are therefore standard Brownian motions:

- cW_{t/c^2}
- $tW_{1/t}$ if $t > 0$, $W_0 = 0$
- $\{W_s - W_{s-t}\}_{0\leq t\leq s}$ for any fixed $s \geq 0$

The maximum of a standard Brownian motion has the distribution

$$P(\max_{0\leq t\leq T} W_t > m) = \frac{2}{\sqrt{2\pi T}} \int_m^\infty e^{-x^2/2T}\, dx$$

This follows from

$$\begin{aligned}
&P\left(\max_{0\leq t\leq T} W_t > m\right)\\
&\quad = P\left(\max_{0\leq t\leq T} W_t > m \wedge W_T > m\right) + P\left(\max_{0\leq t\leq T} W_t > m \wedge W_T < m\right)\\
&\quad = 2P\left(\max_{0\leq t\leq T} W_t > m \wedge W_T > m\right) = 2P(W_T > m) \qquad \square
\end{aligned}$$

In the second step of the proof, we used the fact that if a Brownian motion hits the level m at $t < T$, there is an equal probability of it ending up above m at T as of it ending up below m at T. Indeed, for every Brownian path ending up above m, there is a mirror path that ends up below m. The mirror path is obtained by reflecting the Brownian path in the point m after the time τ when this point is first hit:

$$\tilde{W}_t = \begin{cases} W_t & t \leq \tau \\ 2m - W_t & t > \tau \end{cases}$$

The method of reflecting Brownian paths is called the reflection principle.

Letting $T \to \infty$ in the above result reveals that a Brownian motion hits every fixed point m with probability 1. This brings us to the question of what happens if we allow the point to move.

The probability of a standard Brownian motion to hit a point that moves linearly in time is given by

$$P(a,b) := P(\exists t \in [0,\infty) | W_t > a + bt) = e^{-2a_+ b_+}$$

By definition, $a_+ = \max(a, 0)$ and similar for b_+. As the statement is obviously true for $a < 0$ or $b < 0$, we let $a, b > 0$. Let $q(x)$ be the probability that the process $W_t - bt$ hits the point a if it is currently located at $x < a$. With $p(t', x', t, x)dx'$ the probability of the process to end up in $[x', x' + dx']$ at t' conditional on that it was in x at t, we obtain

$$q(x) = \int_{-\infty}^{a} p(\epsilon, x', 0, x) q(x') dx' + \xi$$

The term ξ contains information regarding paths that crosses the level a before time ϵ. Obviously, ξ tends quickly to zero when $\epsilon \to 0$. Using the backward Kolmogorov equation (see the main text, Sect. 3.7), we obtain

$$-bq_x + \frac{1}{2}q_{xx} = 0$$

Assuming for a moment that the boundary condition $q(-\infty) = 0$ holds, we obtain together with $q(a) = 1$ the solution

$$q(x) = e^{2b(x-a)}$$

The statement follows by setting $x = 0$.

It remains to show that $q(-\infty) = 0$, or equivalently that the probability for a standard Brownian motion to exceed $a + bt$ tends to zero for $a \to \infty$ and $b > 0$. But this follows from the above expression for the maximum of a standard Brownian motion. Indeed, for small t the probability goes to zero because of the large a while for large t the probability goes to zero because of the linear increase in the level coming from b. □

The process $\sigma W_t + \mu t$ is referred to as a Brownian motion with drift μ and volatility σ. The following statement is useful in the analysis of Brownian motions: *The joint distribution of a Brownian motion and its maximum is given by*

$$\begin{aligned} &P(\sigma W_T + \mu T \in [x, x + dx) \wedge \max_{0 \le t \le T}(\sigma W_t + \mu t) \in [m, m + dm)) \\ &\quad = \frac{2(2m - x)}{\sqrt{2\pi\sigma^6 T^3}} \exp\left(-(x - \mu T)^2/2\sigma^2 T - 2m(m - x)/\sigma^2 T\right) dxdm, \\ &\quad\quad x < m, 0 < m \end{aligned}$$

Using

$$\begin{aligned} &P(\exists t \in [0, T)|W_t > a + bt \wedge W_T \in [x, x + dx)) \\ &\quad = P(\exists t \in [0, T)|tW_{1/t} > a + bt \wedge TW_{1/T} \in [x, x + dx)) \\ &\quad = P(\exists t \in [0, T)|W_{1/t} > a/t + b \wedge W_{1/T} \in [x, x + dx)/T) \\ &\quad = P(\exists t \in [1/T, \infty)|W_t > at + b \wedge W_{1/T} \in [x, x + dx)/T) \\ &\quad = P(W_{1/T} \in [x, x + dx]/T)P(a/T + b - x/T, a) \end{aligned}$$

we obtain

$$\begin{aligned} &P(\exists t \in [0, T)|\sigma W_t + \mu t > m \wedge \sigma W_T + \mu T \in [x, x + dx)) \\ &\quad = P(\exists t \in [0, T)|W_t > m/\sigma - \mu t/\sigma \wedge W_T \in [x, x + dx)/\sigma - \mu T/\sigma) \\ &\quad = P(W_{1/T} \in [x, x + dx)/\sigma T - \mu/\sigma)P(m/\sigma T - \mu/\sigma - x/\sigma T + \mu/\sigma, m/\sigma) \\ &\quad = \frac{dx}{\sqrt{2\pi\sigma^2 T}} \exp\left(-(x - \mu T)^2/2\sigma^2 T - 2m_+(m - x)_+/\sigma^2 T\right) \end{aligned}$$

The statement follows by taking the m-derivative. □

The first hitting time $\tau_m = \min\{t|\sigma W_t + \mu t > m\}$ *of the level* $m > 0$ *for a Brownian motion with drift has the distribution*

$$P\left(\tau_m \in [t, t+dt)\right) = \frac{m}{\sqrt{2\pi\sigma^2 t^3}} e^{-(x-\mu t)^2/2\sigma^2 t} dt$$

Proof:

$$\begin{aligned}
P(\tau_m < t) &= P(\max_{0<t'\leq t}(\sigma W_{t'} + \mu t' > m)) \\
&= \int_{-\infty}^{m} P(\sigma W_t + \mu t \in [x, x+dx) \wedge \max_{0<t'\leq t}(\sigma W_{t'} + \mu t' > m)) \\
&\quad + \int_{m}^{\infty} P(\sigma W_t + \mu t \in [x, x+dx)) \\
&= \frac{1}{\sqrt{2\pi\sigma^2 T}} \int_{-\infty}^{m} e^{-(x-\mu t)^2/2\sigma^2 t - 2m(m-x)/\sigma^2 t} dx \\
&\quad + \frac{1}{\sqrt{2\pi\sigma^2 T}} \int_{m}^{\infty} e^{-(x-\mu t)^2/2\sigma^2 t} dx \\
&= e^{2m\mu/\sigma^2} N\left(\frac{-m-\mu t}{\sigma\sqrt{t}}\right) + N\left(\frac{-m+\mu t}{\sigma\sqrt{t}}\right)
\end{aligned}$$

□

A.6 Total Variation and Bounded Variation

Let π be a partition of $[0, t]$, i.e. $\pi = \{t_i\}_{i=0}^{n}$ with $0 = t_0 < t_1 < \cdots < t_n = t$. For $f : \mathbb{R} \to \mathbb{R}$, consider

$$V_{f,\pi}^{p}(t) = \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|^p$$

The p-variation of f is defined as the function $V_f^p(t) = \lim_{\|\pi\|\to 0} V_{f,\pi}^{p}(t)$, where the norm is $\max_i\{|t_{i+1} - t_i|\}$. Important special cases are the total variation ($p = 1$) and the quadratic variation ($p = 2$). The total variation of a process is defined pointwise on Ω, $V_X^2(t, \omega) = V_{X(\omega)}^2(t)$. A process is said to be of bounded variation if the total variation $V_X = V_X^1$ is finite a.s on every compact time interval.

A.7 Martingales

For an integrable random variable X, the conditional expectation $E_s[X_t] = E[X_t|\mathcal{F}_s]$, $s < t$ is by definition the a.s. unique $\mathcal{F}_s$-measurable variable Y_s that satisfies

$$E\left[Y_s \mathbb{1}_A\right] = E\left[X_t \mathbb{1}_A\right], \quad \forall A \in \mathcal{F}_s$$

The conditional expression fulfills the double expectation theorem:

$$E_s\left[E_{s'}\left[X_t\right]\right] = E_s\left[X_t\right], \quad s < s' < t$$

An integrable process X_t with $E_s[X_t] = X_s\ \forall s, t$ is called a martingale. For a stopping time T, i.e. a function $T\colon \Omega \to [0,\infty]$ with $T^{-1}([0,t)) \in \mathcal{F}_t\ \forall t$, the process X^T defined by $X_t^T = X_{\min(t,T)}$ is called the process X stopped at time T. Thus, based on the information available at t the stopping time determines whether X should continue to be stochastic after t or if it should be constant. If X is a martingale, then so is X^T.

A process is said to be a local martingale if there exists a set $\{T_i\}_{i=0}^{\infty}$ of stopping times satisfying $T_i < T_{i+1}$ a.s. for all i and $T_i \to \infty$ a.s. when $i \to \infty$ such that $\mathbb{1}_{\{T_i>0\}} X_t^{T_i}$ is a martingale for all i. The factor $\mathbb{1}_{\{T_i>0\}}$ has been introduced so that processes with X_0 non-integrable can be included in the definition. We have chosen to only prove the statements in this Appendix as applied to martingales. The proofs for the general case of local martingales are often straightforward but will be omitted in order to not obscure the basic ideas of the proofs.

A continuous local martingale of bounded variation is constant a.s.

Assume first that $|V(t)| \le m$ a.s. for all t. For a given partition π we have

$$\begin{aligned} E\left[(X_t - X_0)^2\right] &= E\left[X_t^2 - X_0^2\right] = E\left[\sum_{i=0}^{n-1}\left(X_{t_{i+1}}^2 - X_{t_i}^2\right)\right] \\ &= E\left[\sum_{i=0}^{n-1}\left(X_{t_{i+1}} - X_{t_i}\right)^2\right] \le E\left[\max_i \left|X_{t_{i+1}} - X_{t_i}\right| V(t)\right] \end{aligned}$$

As the integrand $\max_i \left|X_{t_{i+1}} - X_{t_i}\right| V(t)$ is uniformly bounded and converges to zero as $\|\pi\| \to 0$ a.s, the dominated convergence theorem for integrals proves that $E[(X_t - X_0)^2] = 0$. The integrand is positive so we must have $X_t = X_0$ a.s.

For the general case, note that the above part of the proof holds for $X_t^{T_m}$ with stopping time $T_m = \inf\{t \ge 0 | V(t) > m\}$. Thus, $X_t^{T_m} = X_0$ a.s. The finiteness of $V(t)$ implies that $T_m \to \infty$ a.s. when $m \to \infty$ from which it follows that $X_t = X_0$ a.s. □

A process that can be decomposed into a sum of a local martingale and a bounded-variation process is called a semimartingale. The two parts in the decomposition are referred to as the local martingale part and the compensator part. By adding a constant to one of the parts and subtracting it from the other, we can always assume that the compensator is zero at $t = 0$. The process is said to be a continuous semimartingale if both parts in the semimartingale decomposition are continuous. It then follows from the above statement that the decomposition is unique.

Let X be a continuous local martingale. Then, for all t, $V^2_{X,\pi}(t)$ converges in probability to $\langle X\rangle_t$, where the quadratic variation $\langle X\rangle$ is the unique continuous bounded-variation process starting at 0 such that $X^2 - \langle X\rangle$ is a local martingale

First of all, a set of random variables $\{Y_i\}$ is said to converge to Y in probability if $\lim_{i\to\infty} P(|Y_i - Y| > \epsilon) = 0$. For a given partition π of $[s,t]$ we have

$$\begin{aligned} &E_s\left[X_t^2 - \sum_i (X_{t_{i+1}} - X_{t_i})^2\right] - X_s^2 \\ &\quad = \sum_i E_s\left[X^2_{t_{i+1}} - X^2_{t_i} - (X_{t_{i+1}} - X_{t_i})^2\right] = 0 \end{aligned}$$

Taking the limit $\|\pi\| \to 0$ proves the statement formally. It remains to prove that the integral and the limit can be interchanged and that the limit exists. This part of the proof adds nothing to the understanding of martingales necessary for this book and is therefore omitted. As $V^2_{X,\pi}(t)$ is an increasing function of t, the quadratic variation must be of bounded variation. The uniqueness follows from the uniqueness of the semimartingale decomposition. □

Motivated by the polarization identity

$$\begin{aligned} &(X_{i+1} - X_i)(Y_{i+1} - Y_i) \\ &\quad = \frac{1}{4}\Big(((X_{i+1} + Y_{i+1}) - (X_i + Y_i))^2 - ((X_{i+1} - Y_{i+1}) - (X_i - Y_i))^2\Big) \end{aligned}$$

where $X_i = X(t_i)$ and $Y_i = Y(t_i)$, we define the covariation for two continuous local martingales as

$$\langle X, Y\rangle = \frac{1}{4}\left(\langle X + Y\rangle - \langle X - Y\rangle\right)$$

It is then straightforward to generalize the above statement:

For X, Y continuous local martingales, $XY - \langle X, Y\rangle$ is a martingale

An immediate consequence is that the product of two continuous martingales is a semimartingale. From next statement it follows that a continuous process with bounded variation does not contribute to the covariation:

$\langle \mu, X\rangle = 0$ if μ is a continuous process with bounded variation and X is a continuous semimartingale

Consider

$$\left|\sum_{i=0}^{n-1}\left(\mu_{t_{i+1}} - \mu_{t_i}\right)\left(X_{t_{i+1}} - X_{t_i}\right)\right| \le \max_i \left|X_{t_{i+1}} - X_{t_i}\right| \sum_{i=0}^{n-1}\left|\mu_{t_{i+1}} - \mu_{t_i}\right|$$

The second factor on the right-hand side converges to the total variation of μ when $\|\pi\| \to 0$. As X is continuous, the first factor converges uniformly to 0 on the interval $[0,t]$. The statement therefore follows in the limit $\|\pi\| \to 0$. □

A Brownian motion is obviously a continuous martingale and must therefore have unbounded variation. The quadratic variation is given by:
$\langle W\rangle_t = t$ *for* W *a standard Brownian motion*

As

$$W_t^2 = \langle W\rangle_t + \text{martingale}$$

is the unique semimartingale decomposition of W^2, the statement holds if we can prove that $W_t^2 - t$ is a martingale. This follows from

$$\begin{aligned} E_s\left[W_t^2 - W_s^2\right] &= E_s\left[\sum_{i=0}^{n-1}\left(W_{t_{i+1}}^2 - W_{t_i}^2\right)\right] \\ &= E_s\left[\sum_{i=0}^{n-1}\left(W_{t_{i+1}} - W_{t_i}\right)^2\right] = \sum_{i=0}^{n-1}(t_{i+1} - t_i) = t - s \end{aligned}$$ □

A.8 Integrals with Martingales

A process μ of bounded variation induces a (signed) measure ν on the time-axis $\mathbb{R}_+$ by letting it be pathwise defined on Ω: $\nu_\omega([a,b]) = \mu_\omega(b) - \mu_\omega(a)$. The integral $\int_0^t H d\mu := \int_0^t H d\nu$ of continuous processes H over bounded-variation processes can be defined in the Lebesgue sense. In particular, it is possible to define the integral over $\langle X\rangle$ if X is a continuous martingale.

For X a continuous martingale and H a continuous process the integral $\int_0^t HdX$ is defined by

$$\sum_{i=0}^{n-1} H_{t_i}\left(X_{t_{i+1}} - X_{t_i}\right)$$

in the limit $\|\pi\| \to 0$. The integral is obviously a continuous martingale itself. The expectation of the square of the above expression is equal to

$$\begin{aligned} &E\left[\sum_{ij} H_{t_i} H_{t_j}\left(X_{t_{i+1}} - X_{t_i}\right)\left(X_{t_{j+1}} - X_{t_j}\right)\right] \\ &= E\left[\sum_i H_{t_i}^2 E_{t_i}\left[\left(X_{t_{i+1}} - X_{t_i}\right)^2\right]\right] \\ &= E\left[\sum_i H_{t_i}^2 E_{t_i}\left[X_{t_{i+1}}^2 - X_{t_i}^2\right]\right] \end{aligned}$$

$$= E\left[\sum_i H_{t_i}^2 E_{t_i}\left[\langle X\rangle_{t_{i+1}} - \langle X\rangle_{t_i}\right]\right]$$

$$= E\left[\sum_i H_{t_i}^2 \left(\langle X\rangle_{t_{i+1}} - \langle X\rangle_{t_i}\right)\right]$$

In the limit $\|\pi\| \to 0$ we obtain

$$E\left[\left(\int_0^t H dX\right)^2\right] = E\left[\int_0^t H^2 d\langle X\rangle\right]$$

Note that the integral on the right-hand side is finite a.s. as H is continuous on the bounded interval $[0, t]$. As usual, the results can be extended to continuous local martingales X.

When X is a Brownian motion, the approximating sum for the integral consists of independent normally distributed variables. The integral is therefore also normally distributed. It has zero mean (as it is a martingale) and the calculation above gives the variance. Thus, $\int f dW_t \sim \mathcal{N}\left(0, \int f^2 dt\right)$ if f is non-stochastic.

By separating a continuous semimartingale into its local martingale part and its bounded-variation part $Y = X + \mu$, the integral of continuous processes H over continuous semimartingales is defined by

$$\int H dY = \int H dX + \int H d\mu$$

A.9 Ito's Lemma

Let $X : \Omega \to \mathbb{R}$ be a continuous martingale and f a C^2 real-valued function defined on an open set containing the range of X. Then,

$$f(X_t) = f(X_0) + \int_0^t f_X dX + \frac{1}{2}\int_0^t f_{XX} d\langle X\rangle, \quad \text{a.s.}$$

For a given partition π we have

$$f(X_t) - f(X_0) = \sum_{i=0}^{n-1} \left(f(X_{t_{i+1}}) - f(X_{t_i})\right)$$

$$= \sum_{i=0}^{n-1} f_X(X_{t_i})\left(X_{t_{i+1}} - X_{t_i}\right) + \frac{1}{2}\sum_{i=0}^{n-1} f_{XX}(\xi_i)\left(X_{t_{i+1}} - X_{t_i}\right)^2$$

where $\xi_i \in [X_{t_i}, X_{t_{i+1}}]$. The first sum converges to $\int_0^t f_X dX$ when $\|\pi\| \to 0$. It remains to show that the second sum converges to $\frac{1}{2}\int_0^t f_{XX} d\langle X\rangle$, or equivalently that

$$J = \sum_{i=0}^{n-1} f_{XX}(\xi_i)\left(X_{t_{i+1}} - X_{t_i}\right)^2 - \sum_{i=0}^{n-1} f_{XX}(X_{t_i})\left(X_{t_{i+1}} - X_{t_i}\right)^2 \to 0$$

Clearly,

$$|J| \le C\sum_{i=0}^{n-1}\left(X_{t_{i+1}} - X_{t_i}\right)^2, \quad C = \max_i |f_{XX}(\xi_i) - f_{XX}(X_{t_i})|$$

Assuming $|X| < m$ we see that $C \to 0$ when $\|\pi\| \to 0$ while the sum converges to $\langle X\rangle$ which is finite a.s. For unbounded X the proof follows by considering the stopped processes $X_t^{T_m}$ with stopping times $T_m = \inf\{t \ge 0 | V_t > m\}$ so that $T_m \to \infty$ a.s. □

It is straightforward to prove the multidimensional generalization of Ito's lemma: *Let $X : \Omega \to \mathbb{R}^n$ be a continuous semimartingale and $f(t, X)$ a $C^{1,2}$ real-valued function defined on an open set containing the range of (t, X_t). Then a.s.:*

$$f(t, X_t) = f(0, X_0) + \int_0^t f_u(u, X)du + \sum_i \int_0^t f_{X_i} dX_i + \frac{1}{2}\sum_{ij}\int_0^t f_{X_i X_j} d\langle X_i, X_j\rangle$$

Expressed in terms of infinitesimals:

$$df = f_t dt + \sum_i f_{X_i} dX_i + \frac{1}{2}\sum_{ij} f_{X_i X_j} d\langle X_i, X_j\rangle$$

Ito's lemma can be remembered through the product rule $dX_i dX_j = d\langle X_i, X_j\rangle$ ($(dW_t)^2 = dt$ for a Brownian motion) and all other differential products equal to 0.

When $f(X, Y) = XY$, Ito's lemma becomes the product rule of differentiation:

$$d(XY) = XdY + YdX + d\langle X, Y\rangle$$

A.10 Lévy's Characterization of the Brownian Motion

A continuous local martingale X that satisfies $\langle X\rangle_t = t$ and $X_0 = 0$ is a Brownian motion

Setting $X'_t = X_t - X_0 - \frac{1}{2}\langle X\rangle_t$, Ito's lemma implies that $\exp\left(X'_t\right)$ is a martingale for an arbitrary martingale X. Using $\langle X\rangle_t = t$ and $X_0 = 0$ we see that

$$Y_t = \exp\left(\sigma X_t - \frac{1}{2}\sigma^2 t\right)$$

is a martingale. Rearranging the martingale condition $Y_s = E_s[Y_t]$ gives

$$E_s\left[\exp\left(\sigma\left(X_t - X_s\right)\right)\right] = \exp\left(\frac{1}{2}\left(t-s\right)\sigma^2\right)$$

As this is the generating function for the Gaussian distribution, we conclude that $X_t - X_s$ is $\mathcal{N}(0, t-s)$-distributed and is independent of $\mathcal{F}_s$. □

A.11 Measure Change and Girsanov's Theorem

Let P_t be the restriction of P to $\mathcal{F}_t$, i.e. P_t is $\mathcal{F}_t$-measurable and satisfies $P_t(A) = P(A)$ for all $A \in \mathcal{F}_t$. For a given set Ω with filtration $\{\mathcal{F}_t\}$, two measures P and Q are said to be equivalent if for all t, the restrictions P_t and Q_t have the same null sets. Equivalent measures are related by the Radon-Nikodym derivative which is the process M_t defined by $Q_t(A) = \int_A M_t dP_t$ for all $A \in \mathcal{F}_t$. For $A \in \mathcal{F}_s$ it follows that

$$\begin{aligned} E^Q\left[M_s^{-1} E_s^P\left[M_t X_t\right] \mathbb{1}_A\right] &= E^P\left[E_s^P\left[M_t X_t\right] \mathbb{1}_A\right] \\ &= E^P\left[E_s^P\left[M_t X_t \mathbb{1}_A\right]\right] = E^P\left[M_t X_t \mathbb{1}_A\right] = E^Q\left[X_t \mathbb{1}_A\right] \end{aligned}$$

and from the definition of conditional expectation we conclude that

$$E_s^Q\left[X_t\right] = M_s^{-1} E_s^P\left[M_t X_t\right]$$

for X_t a Q-integrable process. We conclude that X_t is a Q-martingale if and only if $M_t X_t$ is a P-martingale. In particular, setting $X_t = 1$ in the above equation proves that the Radon-Nikodym derivative M_t is a P-martingale.

Let X be a continuous P-semimartingale with compensator μ^P. Then X is a continuous Q-semimartingale with compensator

$$\mu^Q = \mu^P + \langle X, \ln M\rangle$$

The product rule of differentiation implies that

$$\begin{aligned} &\left(X - \mu^P - \langle X, \ln M\rangle\right) M \\ &\quad = \int \left(X - \mu^P - \langle X, \ln M\rangle\right) dM + \int M d\left(X - \mu^P\right) \\ &\quad\quad - \int M d\langle X, \ln M\rangle + \langle X - \mu^P - \langle X, \ln M\rangle, M\rangle \end{aligned}$$

By using an approximating sum of the integral, the third term on the right-hand side can be seen to be equal to $-\langle X, M\rangle$. As continuous processes of bounded variation do not contribute to the covariation, this term cancels with the fourth term. The first two terms on the right-hand side are continuous P-martingales from which we conclude that the left-hand side is a continuous P martingale as well. The first factor on the left-hand side must therefore be a continuous Q-martingale. □

Let W be a Brownian motion in P. Then $W - \langle W, \ln M \rangle$ is a Brownian motion in Q.

The previous statement implies that $W - \langle W, \ln M \rangle$ is a continuous Q-martingale. As

$$(W - \langle W, \ln M \rangle)|_{t=0} = 0$$
$$\langle W - \langle W, \ln M \rangle \rangle_t = \langle W \rangle_t = t$$

the statement follows from Levi's characterization of Brownian motions. □

It is now straightforward to prove Girsanov's theorem:

Let

$$M_t = \exp\left(\int_0^t \theta_s dW_s - \frac{1}{2}\int_0^t \theta_s^2 ds\right)$$

where θ_t is integrable with respect to the Brownian motion W_t in the measure P. Then

$$W_t - \int_0^t \theta_s ds$$

is a Brownian motion in Q.

The statement follows from the identity

$$\left\langle W_t, \int_0^t \theta_s dW_s \right\rangle = \int_0^t \theta_s ds$$ □

A.12 No-Arbitrage Pricing

We use a set $\{X^i\}_{i=1}^n$ of strictly positive continuous semimartingales to represent the set of tradable assets in a financial market. By abuse of notation we also let $\{X^i\}_{i=1}^n$ denote the prices of these assets. We assume that the holding of the assets does not result in any cash flows (e.g. dividend payments) and that the assets can be bought or sold at any time in unlimited quantities. Based on these assumptions we develop a model for asset pricing. This section serves as a bridge between the Appendix and the main text. Some of the definitions and results reviewed here can also be found in Chap. 1 and Sect. 3.8.

A trading strategy ϕ is an $\mathbb{R}^n$-valued process describing our holdings ϕ^i of asset X^i. The value of the corresponding portfolio is given by $V_t = \sum_i \phi_t^i X_t^i$. To simplify notation we suppress the summation and write $V_t = \phi_t X_t$. We restrict ourselves to continuous strategies satisfying

$$\phi_t X_t = \phi_0 X_0 + \int_0^t \phi_u dX_u$$

The infinitesimal version reads $d(\phi_t X_t) = \phi_t dX_t$ which means that the price fluctuations in the portfolio come solely from changes in the asset prices. Such a strategy is said to be self financing, i.e. there is no in- or out-flux of money. Examples include:

- ϕ constant: it is then possible to move ϕ outside the integral and the above relation is trivial
- If X is a standard Brownian motion and the portfolio value satisfies $\partial V/\partial t = -(1/2)\partial^2 V/\partial X^2$ then $\phi = \partial V/\partial X$ is self financing. This follows since

$$V_t = V_0 + \int_0^t dV = V_0 + \int_0^t \frac{\partial V}{\partial X} dX$$

holds because of Ito's lemma.

A self-financing strategy ϕ is said to be an arbitrage strategy if there exists a t for which $V_0 = 0, V_t \geq 0$ a.s. and $P(V_t > 0) > 0$. By restricting ourselves to self-financing strategies without arbitrage, we exclude strategies with risk-free gains.

Instead of valuing the assets in dollar terms, the valuation can be done relative to one of the assets. The asset with respect to which the valuation is done is called the numeraire and by a reordering we can assumed it to be X^0. The values of the assets are then given by $(1, X^1/X^0, \ldots, X^n/X^0)$. The concepts of arbitrage and a self-financing strategy are preserved when using a numeraire. For example, the preservation of the self-financing property follows from

$$\begin{aligned} d\left(\phi X^i (X^0)^{-1}\right) &= \phi X^i d(X^0)^{-1} + (X^0)^{-1} d(\phi X^i) + d\langle \phi X^i, (X^0)^{-1}\rangle \\ &= \phi X^i d(X^0)^{-1} + (X^0)^{-1} \phi d(X^i) + \phi d\langle X^i, (X^0)^{-1}\rangle \\ &= \phi d\left(X^i (X^0)^{-1}\right) \end{aligned}$$

Absence of arbitrage implies that if there is a non-zero probability for X_t^i/X_t^0 to be greater than X_0^i/X_0^0 then there must also be a non-zero probability for it to be smaller than X_0^i/X_0^0. By reweighing the probabilities it is then possible to turn this process into a martingale, i.e. there exists a probability measure Q equivalent to P such that the processes X^i/X^0 are martingales. The strict mathematical proof of the fact that this can be done simultaneously for all i does not provide us with further insights and is therefore be omitted. We only need the reverse statement:

If there exists a measure Q such that $\{X^i/X^0\}$ are local martingales then there do not exist any self-financing strategies with arbitrage.

As

$$V_t/X_t^0 = \phi_t X_t^i/X_t^0 = V_0/X_0^0 + \int_0^t \phi d\left(X^i/X^0\right)$$

is a Q-martingale, we have $E[V_t/X_t^0] = V_0/X_0^0$. If $V_0 = 0$ and $V_t \geq 0$ a.s. then $V_t = 0$ a.s. which proves that ϕ cannot be an arbitrage strategy. □

We now describe how to price contracts under this framework. In financial mathematics, one is often faced with the problem of finding the value V_0 of a contract from the knowledge of its value V_T at a future time. We assume that the contract can be replicated with a self-financing strategy $V = \sum_i \phi^i X^i$ in terms of some basic assets $\{X^i\}$. Whether it is really possible to represent (or approximate) V in such a way is usually clear from the context. We then use one of the assets as a numeraire N_t and assume the existence of a martingale measure, i.e a measure such that $(X^0/N, \ldots X^n/N)$ are martingales. This assumption excludes the existence of arbitrage strategies. As ϕ is self financing, V/N is a martingale and

$$\frac{V_0}{N_0} = E\left[\frac{V_T}{N_T}\right]$$

The numeraire N is often chosen to be a simple tradable such as a zero-coupon bond for which we know both the value at T and today's value. The only unknown in the above equation is V_0 which therefore can be computed. Observe that the expectation is taken under Q and not under the real-world measure.

The pricing equation might seem rather abstract at first sight. For example, it is not at all clear at this point how to find the measure Q. However, we use this pricing model throughout the book and hopefully it will be clear how to implement and use it.

When using the pricing model, we need to choose a numeraire, find a corresponding measure for which the tradables are martingales and then calculate expectations. It is often necessary to do the computations for more than one numeraire and measure, for example, when the pricing is done in one measure and the calibration in another. From the identity

$$V_0 = N_0^P \int V_T/N_T^P \, dP = N_0^Q \int V_T/N_T^Q \left(\frac{N_0^P}{N_0^Q}\frac{N_T^Q}{N_T^P}\right) dP$$

we see that changing numeraire from N^P to N^Q implies a change in measure from dP to $dQ = MdP$, with M_T equal to $N_0^P N_T^Q / N_0^Q N_T^P$.

Index

C. Ekstrand, *Financial Derivatives Modeling*, DOI 10.1007/978-3-642-22155-2,