6

Yield Curve Construction and Risk Management

In a nutshell, the job of an interest rate model is to describe the random movement of a curve of discount bond prices through time, starting from a known initial condition. In reality, however, only a few short-dated discount bonds are directly quoted in the market at any given time, a long stretch from the assumption of many models that an initial curve of discount bond prices is observable for a continuum of maturities out to 20–30 years or more. Fortunately, a number of liquid securities depend in relatively straightforward fashion on discount bonds, opening up the possibility of uncovering discount bond prices from prices of such securities. Still, as only a finite set of securities are quoted in the market, constructing a continuous curve of discount bond prices will inevitably require us to complement market observations with an interpolation rule, based perhaps on direct assumptions about functional form or perhaps on a regularity norm to be optimized on. A somewhat specialized area of research, discount curve construction relies on techniques from a number of fields, including statistics and computer graphics. While we cannot possibly do the subject full justice, discount curve construction is a fundamental step in the modeling exercise, and no book on fixed income models is complete without a discussion of basic techniques.

As mentioned in the Preface to this book, the crisis of 2007–2009 have lead to changes in the foundations of interest rate modeling, not least in the area of yield curve construction and risk management. Pre-crisis, is was often sufficient to construct only a single (Libor) discount curve, but nowadays the task is more complicated as a whole collection of inter-related curves is required. Nevertheless, the traditional techniques used for single-curve construction are by no means obsolete, and their mastery is required before more ambitious curve algorithms can be attempted. Accordingly, we have split this chapter into three parts. In the first, and most significant, part, we introduce notations and cover a number of curve construction techniques, moving from simply bootstrapped C^0 curves through "local spline" C^1 curves to full C^2 smoothing splines with and without tension.

Perturbation locality is discussed, as are methods to control behavior under perturbations. In the second part we discuss the management of interest rate curve risk, covering both basic approaches as well as more advanced methods based on Jacobian techniques. In the last part, we discuss a number of specialized issues and contemporaneous extensions, most notably turn-of-year adjustments and techniques to construct separate discount and forward curves. The need for such a separation has long been recognized (albeit neglected in the literature) as a requirement to avoid arbitrages in markets for foreign exchange forwards and for floating-floating cross-currency swaps. More recently, similar issues have appeared in purely domestic markets where the Libor rate is no longer considered a good proxy for the risk-free discount rate, and where a significant *tenor basis* has developed in floating-floating single-currency swaps. Accordingly, we conclude the chapter with a description of techniques for building a *multi-index curve group*, a self-consistent arbitrage-free collection of discount and forward curves suitable for valuation of different types of swaps and other interest rate derivatives.

6.1 Notations and Problem Definition

6.1.1 Discount Curves

Throughout this chapter, we use the abbreviated notation $P(T) = P(0,T)$ where $P : [0, \mathcal{T}] \to (0, 1]$ is a continuous, monotonically decreasing *discount curve*. $\mathcal{T}$ denotes the maximum maturity considered, typically given as the longest maturity in the set of securities the curve is built from. Let there be N such securities — the *benchmark set* — with observable prices $V_1, \ldots, V_N$. We assume that the time 0 price $V_i = V_i(0)$ of security i can be written as a linear combination of discount bond prices at different maturities,

$$V_i = \sum_{j=1}^{M} c_{i,j} P(t_j), \quad i = 1, \ldots, N, \tag{6.1}$$

where $0 < t_1 < t_2 < \ldots < t_M \leq \mathcal{T}$ is a given finite set of dates, in practice obtained by merging together the cash flow dates of each of the N benchmark securities. Let $T_1, T_2, \ldots, T_N$ denote the final maturities of the N benchmark securities, in which case we necessarily must have

$$c_{i,j} = 0, \quad t_j > T_i.$$

Securities that can be represented by pricing expressions of the form (6.1) obviously include coupon and discount bonds, but also FRAs and fixed-floating interest rate swaps. For instance, consider a newly issued unit-notional fixed-floating swap, paying a coupon of $c\tau$ at times τ, 2τ, 3τ,..., $n\tau$. If no spread is paid on the floating rate, the time 0 total swap value to the fixed payer is

$$V_{\text{swap}} = 1 - P(t_n) - \sum_{j=1}^{n} c\tau P(j\tau),$$

as already discussed in Chapter 5. We can rewrite this as

$$1 - V_{\text{swap}} = P(t_n) + \sum_{j=1}^{n} c\tau P(j\tau), \tag{6.2}$$

which is in the form[1] (6.1) once we interpret $V_i = 1 - V_{\text{swap}}$. In practice, swaps used for discount curve construction are nearly always newly issued and par-valued, in the sense that the coupon c is set to make $V_{\text{swap}} = 0$. To give another example, consider an FRA on the $[T, T+\tau]$ Libor rate, for which formula (5.2) in Chapter 5 gives, at $t = 0$,

$$V_{\text{FRA}} = \tau \left(L(0, T, T+\tau) - k\right) P(T+\tau), \tag{6.3}$$

where k is the quoted FRA rate. From the definition of $L(0, T, T+\tau)$ this is just

$$V_{\text{FRA}} = P(T) - P(T+\tau) - k\tau P(T+\tau) = P(T) - (1+k\tau)\,, P(T+\tau)$$

which is in the form (6.1). As for swaps, FRAs used for curve construction are newly issued and typically have k set such that $V_{\text{FRA}} = 0$.

The choice of the securities to be included in the benchmark set depends on the market under consideration. For instance, to construct a Treasury bond curve, it is natural to choose a set of Treasury bonds and T-Bills. On the other hand, if we are interested in constructing a discount curve applicable for bonds issued by a particular firm, we would naturally use bonds and loans issued by the firm in question. For our purposes, the most important discount curve is the *Libor curve*, constructed out of market quotes for Libor deposits, swaps and Eurodollar futures. In the construction of this curve, most firms would use a few certificates of deposit for the first 3 months of the curve, followed by a strip of Eurodollar futures[2] (with maturities staggered 3 months apart) out to 3 or 4 years. Par swaps are then used for the rest of the curve, with typical maturities being 5, 7, 10, 12, 15, 20, 25, and 30 years.

---

[1] For swaps where payment schedules do not coincide perfectly with the accrual periods of the Libor rates, the expression (6.2) is only an approximation, albeit a very good one. In practice we can construct the yield curve assuming that (6.2) holds, and then perform a small post-processing clean-up iteration, along the lines of the algorithm in Section 6.5.2.4.

[2] We note that Eurodollar futures contracts do not allow for a pricing expression of the form (6.1), so a pre-processing step is normally employed to convert the futures rate quote to a forward rate (FRA) quote. See Proposition 4.5.3 or Chapter 16 for more on this.

6.1.2 Matrix Formulation

Define the M-dimensional discount bond vector[3]

$$\mathbf{P} = (P(t_1), \ldots, P(t_M))^\top ,$$

and let $\mathbf{V} = (V_1, \ldots, V_N)^\top$ be the vector of observable security prices. Also let $\mathbf{c} = \{c_{i,j}\}$ be an $(N \times M)$-dimensional matrix containing all the cash flows produced by the chosen set of securities. The matrix $\mathbf{c}$ would typically be quite sparse, with many rows containing only a few non-zero entries. A typical, albeit simplified, form of the matrix $\mathbf{c}$ might be ($\times$ marks a non-zero element)

$$\mathbf{c} = \begin{pmatrix}
\times & & & & & & & & & & & & & & \\
 & \times & & & & & & & & & & & & & \\
 & & \times & \times & & & & & & & & & & & \\
 & & & \times & \times & & & & & & & & & & \\
 & & & & \times & \times & & & & & & & & & \\
 & & & & & \times & \times & & & & & & & & \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times & & & & & \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & & & & \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times
\end{pmatrix},$$

corresponding to two certificates of deposit (first two rows); four FRAs or Eurodollar futures (next four rows); and three swaps (last three rows).

In a consistent, friction-free market without arbitrage opportunities, the fundamental relation

$$\mathbf{V} = \mathbf{cP} \tag{6.4}$$

must be satisfied, giving us a starting point to determine $\mathbf{P}$.

6.1.3 Construction Principles and Yield Curves

In practice, we normally have more cash flow dates than benchmark security prices, i.e. $M > N$, in which case (6.4) is insufficient to uniquely determine $\mathbf{P}$. The problem of curve construction essentially boils down to supplementing (6.4) with enough additional assumptions to allow us to extract $\mathbf{P}$ and to determine $P(T)$ for values of T not in the cash flow timing set $\{t_j\}_{j=1}^M$.

As it is normally easier to devise an interpolation scheme on a curve that is reasonably flat (rather than exponentially decaying), it is common to perform the curve fitting exercise on *discount yields*, rather than directly on bond prices[4]. Specifically, we introduce a continuous yield function $y : [0, T] \to \mathbb{R}_+$ given by

[3]For extra clarity, throughout this chapter we use boldface type for vectors and matrices.

[4]See e.g. Shea [1984] for a discussion of the pitfalls associated with curve interpolators that work directly on the discount function $P(T)$.

$$e^{-y(T)T} = P(T) \quad \Rightarrow \quad y(T) = -T^{-1} \ln P(T), \tag{6.5}$$

such that in (6.4)

$$\mathbf{P} = \left(e^{-y(t_1)t_1}, \ldots, e^{-y(t_M)t_M}\right)^\top.$$

The mapping $T \mapsto y(T)$ is known as the *yield curve*; it is related to the discount curve by the simple transformation (6.5). Of related interest is also the *instantaneous forward curve* $f(T)$, given by

$$P(T) = e^{-\int_0^T f(u)du}. \tag{6.6}$$

Notice that

$$f(T) = y(T) + \frac{dy(T)}{dT}T. \tag{6.7}$$

For alternative transformations, and a discussion of their relative merits, see Andersen [2005]. Unless explicitly stated, in the remainder of this chapter we shall work with yields, i.e. we treat $y(T)$ as the fundamental curve to be estimated.

Whatever space we elect to work in, we have at least three options for solving (6.4).

1. We can introduce new and unspanned securities such that $N = M$ and (6.4) allows for exactly one solution.
2. We can use a parameterization of the yield curve with precisely N parameters, using the N equations in (6.4) to recover these parameters.
3. We can search the space of all solutions to (6.4) and choose the one that is "optimal" according to a given criterion.

Let us provide some comments to these three ideas. First, in option 1 above, introduction of new securities might not truly be possible — such securities may simply not exist — but sometimes interpolation rules applied to the given benchmark set may allow us to provide reasonable values for an additional set of "fictitious" securities. Although it can occasionally be useful in pre-processing to pad an overly sparse benchmark set, this idea will often require some quite ad-hoc decisions about the specifics of the fictitious securities, and excessive use may ultimately lead to odd-looking curves and suboptimal hedge reports. When an interpolation rule is to be used, it is typically better to apply it directly on more fundamental quantities such as zero-coupon yields or forward rates, thereby maintaining a higher degree of control over the resulting discount curve.

In option 2 above, parametric functional forms (e.g. Nelson and Siegel [1987]) are sometimes used, but it is far more common to work with a spline representation with N user-selected knots (typically at the maturity dates of the benchmark securities), with the level of the yield curve at these knots constituting the N unknowns to be solved for. We discuss the details of

this approach in Section 6.2, using a number of different spline types. Some required elements of basic spline theory can be found in Appendix 6.A of this chapter.

Option 3 constitutes the most sophisticated approach and can often be stated in completely non-parametric terms, with the yield curve emerging naturally as the solution to an optimization problem. If carefully stated, this approach can easily be modified to handle the situation where the system of equations (6.4) is (near-) singular, in the sense that either no solutions exist or all solutions are irregular and non-smooth[5]. Technically, we handle this by working with *smoothing splines*, in the process replacing (6.4) with a penalized least-squares optimization problem. We discuss elements of this idea in Section 6.3 below.

6.2 Yield Curve Fitting with N-Knot Splines

In this section we discuss a number of well-known yield curve algorithms based on polynomial and exponential (tension) splines of various degrees of differentiability. Throughout, we assume that we can select and arrange our benchmark set of securities to guarantee that the maturities of the benchmark securities satisfy

$$T_i > T_{i-1}, \quad i = 2, 3, \ldots, N, \tag{6.8}$$

where the inequality is strict. Equation (6.8) constitutes a "spanning" condition and allows us to select the N maturities as distinct knots in our yield curve splines.

6.2.1 C^0 Yield Curves: Bootstrapping

If continuity of the yield curve is all that we require, we can work with a common iterative procedure known as *bootstrapping*. The basic idea is encapsulated in the following iteration:

1. Let $P(t_j)$ be known for $t_j \leq T_{i-1}$, such that prices for benchmark securities $1, \ldots, i-1$ are matched.
2. Make a guess for $P(T_i)$.
3. Use an interpolation rule to fill in $P(t_j)$, $T_{i-1} < t_j < T_i$.
4. Compute V_i from the now-known values of $P(t_j)$, $t_j \leq T_i$.
5. If V_i equals the value observed in the market, stop. Otherwise return to Step 2.
6. If $i < N$, set $i = i + 1$ and repeat.

[5]Intuitively, this situation can arise if, say, two or more securities in the benchmark set have near-identical cash flows, yet have significantly different present values.

The updating of guesses when iterating over Steps 2 through 5 can be handled by any standard one-dimensional root-search algorithm (e.g., the Newton-Raphson or secant methods).

There are strong limitations on what kind of interpolation rule can be applied in Step 3. For instance, one might consider using a representation in terms of instantaneous forwards $f(T)$ (see (6.6)), with the assumption that $f(T)$ is a continuous piecewise linear function on the maturity grid $\{T_i\}_{i=1}^N$. While based on seemingly natural assumptions, this interpolation rule can, however, be shown to be numerically unstable and prone to oscillations. Some stable, and standard, choices for interpolation rules are covered in the next two sections; common for both is that the resulting yield curve is continuous, but non-differentiable. This, in turn, implies that the instantaneous forward curve is discontinuous (see (6.7)).

6.2.1.1 Piecewise Linear Yields

The most common discount curve bootstrap algorithm assumes that the continuously compounded yield $y(T)$ in (6.5) is a continuous piecewise linear function on $\{T_i\}_{i=1}^N$. Formally, the interpolation rule in Step 3 of the algorithm in Section 6.2.1 writes $P(T) = e^{-y(T)T}$, where

$$y(T) = y(T_i)\frac{T_{i+1} - T}{T_{i+1} - T_i} + y(T_{i+1})\frac{T - T_i}{T_{i+1} - T_i}, \quad T \in [T_i, T_{i+1}]. \tag{6.9}$$

To initiate the iterative bootstrap algorithm, we note that the interpolation rule (6.9) may require us to provide an equation for $y(t)$, $t < T_1$. There are a number of ways to do this; one common choice is to simply set $y(t) = y(T_1)$, $t < T_1$.

To give a feel for the types of yield curves produced by linear yield bootstrapping, let us consider a simple example with a benchmark set of $N = 10$ swaps, with maturities and quoted par swap rates as given in Table 6.1[6].

The swaps are assumed to pay on a semi-annual basis,

$$t_j = j \cdot 0.5, \quad j = 1, 2, \ldots, 50.$$

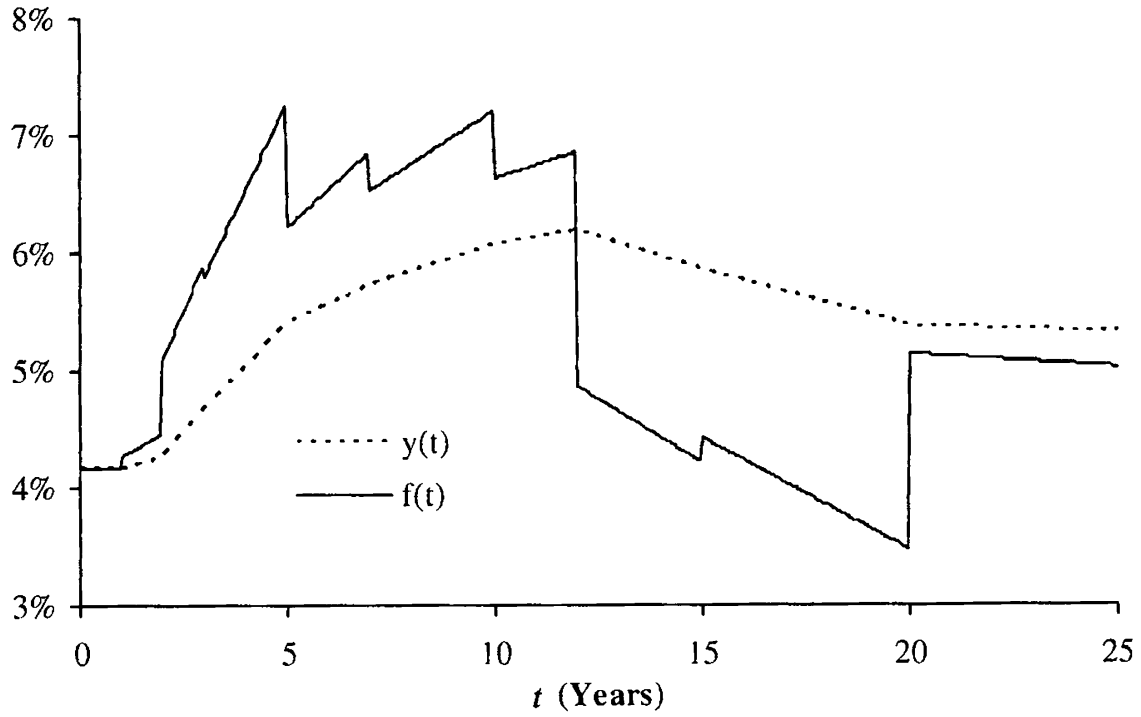
Setting $y(t) = y(1)$, $t < 1$, and then running the bootstrap procedure on the swap price expression (6.2) results in the yield shown in Figure 6.1. The same figure also shows the continuously compounded forward curve, as computed by equation (6.7). The discontinuous "saw-tooth" shape of the forward curve is characteristic for bootstrapped yield curves with piecewise linear yield.

[6]In actual markets, swap yields are most often increasing functions of the swap maturity, rather than humped as in Table 6.1. The data in Table 6.1 was picked to stress the curve construction algorithms, in order to emphasize their strengths and weaknesses.

Maturity (Years)	Swap Par Rate
1	4.20%
2	4.30%
3	4.70%
5	5.40%
7	5.70%
10	6.00%
12	6.10%
15	5.90%
20	5.60%
25	5.55%

Table 6.1. Swap Benchmark Set for Numerical Tests

Fig. 6.1. Yield and Forward Curve



Notes: Yield curve is constructed by bootstrapping, assuming piecewise linear yields. Swap data is in Table 6.1.

6.2.1.2 Piecewise Flat Forward Rates

Assume now that the instantaneous forward curve is piecewise flat, switching to a new level at each point in $\{T_i\}$, i.e.

$$f(T) = f(T_i), \quad T \in [T_i, T_{i+1}), \tag{6.10}$$

with $T_0 \triangleq 0$. This corresponds to an interpolation rule where $\ln P(T)$ is linear in T, or

$$P(T) = P(T_i)e^{-f(T_i)(T-T_i)}, \quad T \in [T_i, T_{i+1}),$$

where a bootstrap algorithm can be used to establish the values of the N unknown constants $f(T_0)$, $f(T_1)$,..., $f(T_{N-1})$. From the equation

$$y(T)T = \int_0^T f(u)\,du,$$

we see that the assumption of piecewise flat forwards gives, for $T \in [T_i, T_{i+1})$,

$$y(T) = \frac{y(T_i)T_i + f(T_i)(T - T_i)}{T} = f(T_i) + \frac{(y(T_i) - f(T_i))\,T_i}{T},$$

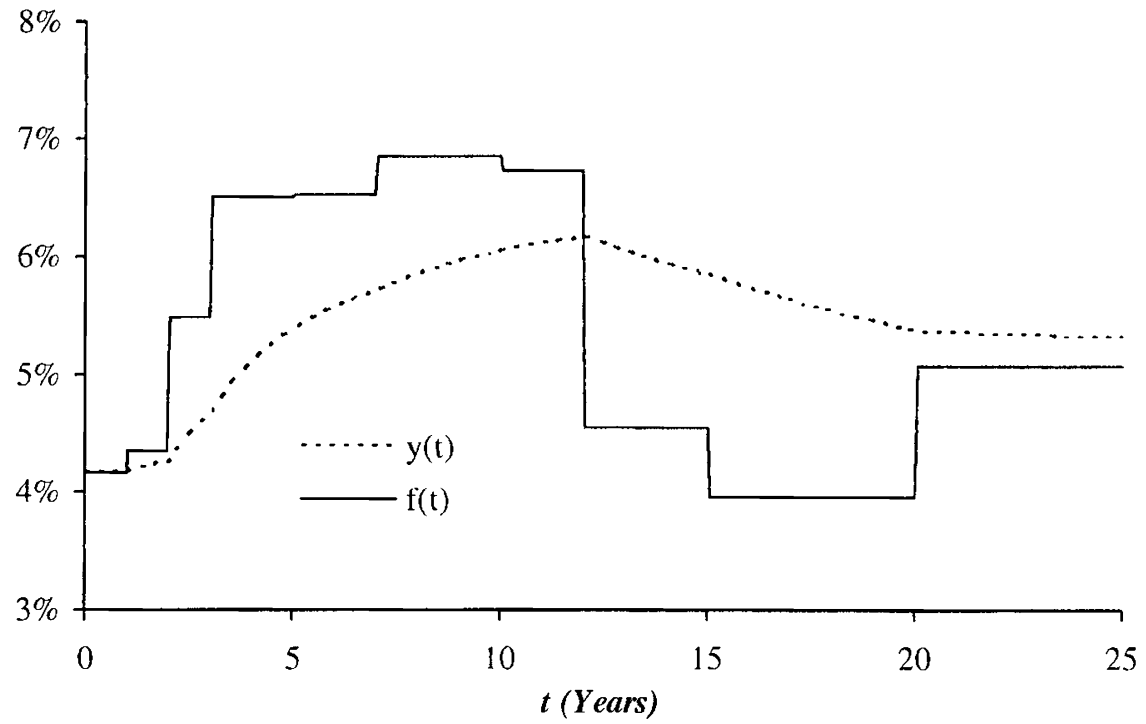
or

$$y(T) = \frac{1}{T}\left(T_i y(T_i)\frac{T_{i+1} - T}{T_{i+1} - T_i} + T_{i+1} y\,(T_{i+1})\,\frac{T - T_i}{T_{i+1} - T_i}\right).$$

The yield curve will remain continuous.

Figure 6.2 below shows the results of applying (6.10) to the swap data in Table 6.1. Notice the non-linear behavior of yields between maturity dates and the staircase shape of the forward curve.

Fig. 6.2. Yield and Forward Curve



Notes: Yield curve is constructed by bootstrapping, assuming piecewise flat forward rates. Swap data is in Table 6.1.

6.2.2 C^1 Yield Curves: Hermite Splines

As we have seen, simply bootstrapped curves generally result in a discontinuous forward curve. From an empirical/economic perspective, such discontinuities are often unrealistic, and may also result in distortions of derivative prices[7] and technical difficulties in dynamic yield curve models. In this section, we consider a simple scheme to extend the bootstrapping technique to produce a once-differentiable yield curve and a continuous forward curve. Our scheme relies on *Hermite cubic splines*, where we write

$$y(T) = a_{3,i}(T - T_i)^3 + a_{2,i}(T - T_i)^2 + a_{1,i}(T - T_i) + a_{0,i}, \quad T \in [T_i, T_{i+1}], \tag{6.11}$$

for a series of constants $a_{3,i}$, $a_{2,i}$, $a_{1,i}$, $a_{0,i}$ to be determined from given values of $y(T_i)$, $y(T_{i+1})$, $y'(T_i)$, and $y'(T_{i+1})$. Appendix 6.A.1 contains a review of Hermite spline theory.

A particularly popular choice among Hermite splines is the *Catmull-Rom spline*, where derivatives $y'(T_i)$, $i = 1, \ldots, N$, are constructed by finite differences, relieving the user from directly specifying them. As shown in Appendix 6.A.1, for the Catmull-Rom spline we can organize (6.11) in a vector/matrix form as

$$y(T) = \mathbf{D}_i(T)^\top \mathbf{A}_i \begin{pmatrix} y_{i-1} \\ y_i \\ y_{i+1} \\ y_{i+2} \end{pmatrix}, \quad T \in [T_i, T_{i+1}], \quad i = 1, \ldots, N-1, \tag{6.12}$$

where, adapting as necessary the notation from the previous section,

$$\mathbf{D}_i(T) = \begin{pmatrix} d_i^3 \\ d_i^2 \\ d_i \\ 1 \end{pmatrix}, \quad d_i = \frac{T - T_i}{h_i}, \quad y_i = y(T_i), \quad h_i = T_{i+1} - T_i,$$

and the matrix $\mathbf{A}_i$ is as given in (6.54)–(6.56) in Appendix 6.A.1. While nominally (6.12) involves the values y_{N+1} and y_0, the matrices $\mathbf{A}_{N-1}$ and $\mathbf{A}_1$ are such that these values are irrelevant.

The Catmull-Rom spline prescription (6.12) completely specifies the yield curve on the interval $[T_1, T_N]$, given the N constants $y_1, \ldots, y_N$. To extend the yield curve to cover the interval $[0, T_1)$, we need to supply additional extrapolation assumptions. As in bootstrapping, possible choices for this additional equation is $y_0 = y(0) = y_1$, or perhaps the slope condition

$$\frac{y_1 - y_0}{h_0} = \frac{y_2 - y_1}{h_1}. \tag{6.13}$$

[7] For instance, as deal maturity crosses a point of discontinuity on the forward curve, the price of an FRA or a caplet on a short-tenor rate will jump.

Away from the boundaries, we notice that the price of security i depends only on $y_1, \ldots, y_{i+1}$, as the pricing equations take the diagonal form

$$\begin{aligned} V_1 &= F_1(y_1, y_2, y_3), \\ V_2 &= F_2(y_1, y_2, y_3), \\ V_3 &= F_3(y_1, y_2, y_3, y_4) \\ &\vdots \\ V_{N-1} &= F_{N-1}(y_1, \ldots, y_N), \\ V_N &= F_N(y_1, \ldots, y_N), \end{aligned}$$

for non-linear functions F_i. Here F_i is typically only mildly sensitive to y_{i+1}, so the system of equations is nearly, but not quite, in bootstrap form. This makes solving for the y_i's an easy fare for a standard non-linear root-search algorithm (see Press et al. [1992] for several algorithms). We can also consider an iteration on a series of bootstrap procedures. To describe this idea, let $y_i^{(k)}$ be the value for y_i found in the k-th iteration, and consider then the following algorithm:

1. Let $y_j^{(k)}$, $j = 1, \ldots, i-1$, and $y_{i+1}^{(k-1)}$ all be known.
2. Make a guess for $y_i^{(k)}$.
3. Compute $V_i = F_i(y_1^{(k)}, \ldots, y_i^{(k)}, y_{i+1}^{(k-1)})$.
4. If V_i equals the market value stop. Otherwise return to Step 2.
5. If $i < N$, set $i = i + 1$ and repeat.

We emphasize that the iteration over Steps 2–4 is still only one-dimensional, as in the bootstrapping algorithm of Section 6.2.1. Upon completion, the algorithm above yields $y_1^{(k)}, \ldots, y_N^{(k)}$. Iterating over k, we repeat the algorithm until the differences between the yields found at the k-th and $(k+1)$-th iteration are sufficiently small, say when

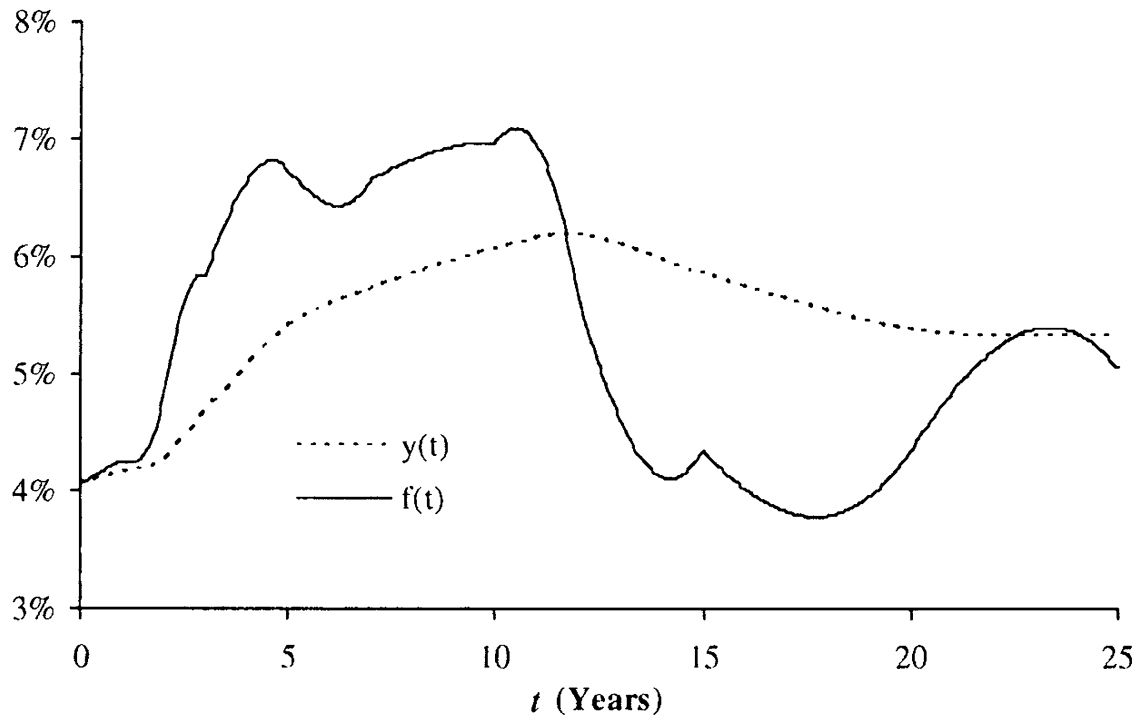
$$N^{-1} \sum_{i=1}^{N} \left(y_i^{(k+1)} - y_i^{(k)} \right)^2 < \varepsilon^2,$$

where ε is a given tolerance. To initialize the iteration over k, we need a starting guess $y_1^{(0)}, \ldots, y_N^{(0)}$; a good choice is the yield curve constructed by regular bootstrapping.

In Figure 6.3, we show the results of applying the algorithm above (using the boundary choice (6.13)) to the numerical example of Sections 6.2.1.1 and 6.2.1.2. We see that, as desired, the yield curve is smooth and the instantaneous forward curve is continuous. As the yield curve by construction is only once differentiable, equation (6.7) shows that the forward curve is not differentiable; this is obvious from the figure.

We can easily extend the procedure above beyond Catmull-Rom splines to more complicated C^1 cubic splines in the Hermite class, using results

Fig. 6.3. Yield and Forward Curve



Notes: Yield curve is assumed to be a Catmull-Rom cubic spline. Swap data is in Table 6.1.

from Appendix 6.A. For instance, it is relatively straightforward to add *tension* to the Catmull-Rom spline. We cover twice-differentiable tension splines later in this chapter.

6.2.3 C^2 Yield Curves: Twice Differentiable Cubic Splines

While the spline method introduced in the previous section often produces acceptable yield curves, the method is heuristic in nature and ultimately does not produce a smooth forward curve. To improve on the latter, one alternative is to remain in the realm of cubic splines, but now insist that the curve is twice differentiable everywhere on $[T_1, T_N]$. We then write (see Appendix 6.A.2)

$$y(T) = \frac{(T_{i+1}-T)^3}{6h_i}y_i'' + \frac{(T-T_i)^3}{6h_i}y_{i+1}'' + (T_{i+1}-T)\left(\frac{y_i}{h_i} - \frac{h_i}{6}y_i''\right) + (T-T_i)\left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6}y_{i+1}''\right), \quad T \in [T_i, T_{i+1}], \quad (6.14)$$

where $y_i'' = d^2y(T_i)/dT^2$, $y_i = y(T_i)$, and $h_i = T_{i+1} - T_i$. The appendix demonstrates that continuity of the second derivative across the $\{T_i\}$ knots requires that the y_i'' and y_i are connected through a tri-diagonal linear system of equations, see equation (6.62). To state the expressions explicitly in matrix format, let $\mathbf{y}'' = (y_2'', y_3'', \ldots, y_{N-2}'', y_{N-1}'')^\top$ and

$\mathbf{y} = (y_2, y_3, \ldots, y_{N-2}, y_{N-1})^\top$ such that

$$\mathbf{B}\mathbf{y}'' = \mathbf{C}\mathbf{y} + \mathbf{M}\left(y_1, y_N, y_1'', y_N''\right), \tag{6.15}$$

where the matrices $\mathbf{B}$ and $\mathbf{C}$ are both $(N-2) \times (N-2)$ tri-diagonal, with elements given by

$$B_{i,i} = \frac{h_i + h_{i+1}}{3}, \quad B_{i,i+1} = \frac{h_{i+1}}{6}, \quad B_{i,i-1} = \frac{h_i}{6},$$

and

$$C_{i,i} = -\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right), \quad C_{i,i+1} = \frac{1}{h_{i+1}}, \quad C_{i,i-1} = \frac{1}{h_i}.$$

The $(N-2)$-dimensional vector $\mathbf{M}(y_1, y_N, y_1'', y_N'')$ captures boundary terms at T_1 and T_N. The most important — and, as discussed later, in a sense *best* — boundary specification is that of the *natural spline*, where we set $y_1'' = y_N'' = 0$. In this case, we have

$$\mathbf{M}(y_1, y_N, y_1'', y_N'') = \mathbf{M}(y_1, y_N) = \left(\frac{y_1}{h_1}, 0, 0, \ldots, 0, 0, \frac{y_N}{h_{N-1}}\right)^\top.$$

Notice that application of a natural boundary condition at time T_1 allows us to recover yields inside the time period $[0, T_1]$ by linear interpolation, using the gradient $y'(T_1)$ at time T_1 (which can easily be found by differentiating (6.14)).

We notice that (6.14) combined with (6.15) allows us to turn any guess of $y_1, y_2, \ldots, y_N$ into a guess for the vector $\mathbf{P}$ in (6.4). Specifically, we perform the following steps:

1. Compute the right-hand side of (6.15).
2. Use a standard tri-diagonal LU solver (see Press et al. [1992]) to invert (6.15) and recover $\mathbf{y}''$.
3. Apply (6.14) to determine[8] all values of $y(t_j)$, $j = 1, \ldots, M$, extrapolating as necessary when $t_j < T_1$.
4. Use (6.5) to establish $\mathbf{P}$.

The computational effort of Steps 1 through 4 are $O(N)$, $O(N-2)$, $O(M)$, and $O(M)$, respectively.

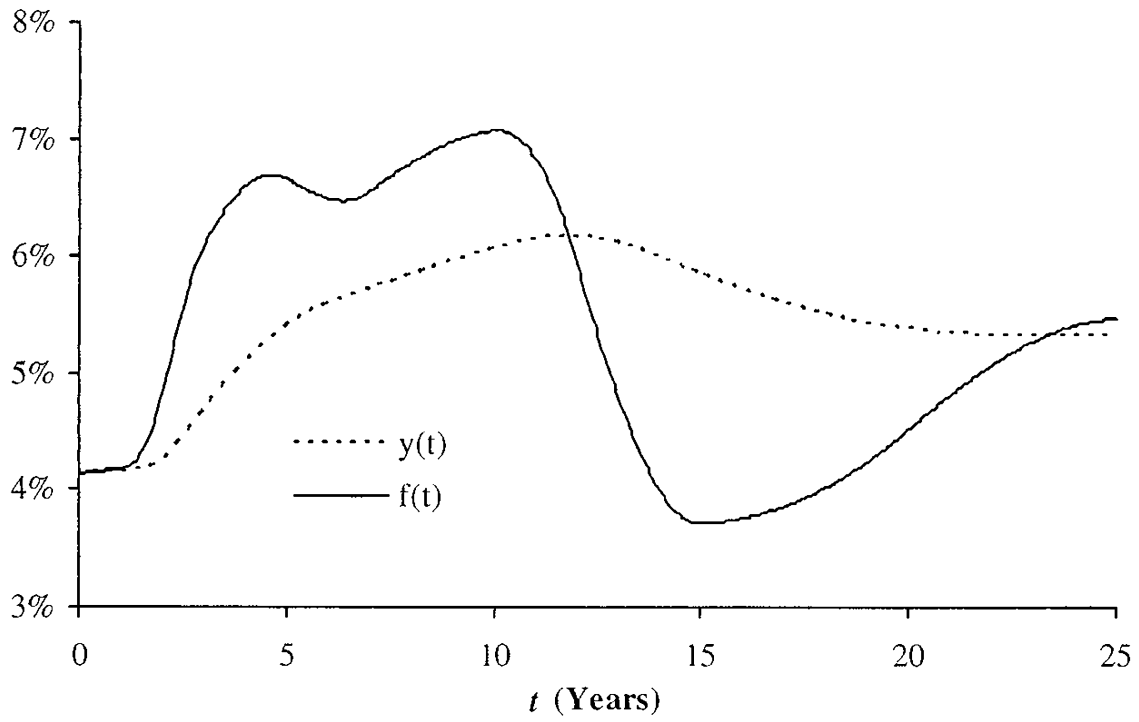
To solve for the correct values of $y_1, y_2, \ldots, y_N$, we iterate on Steps 1–4 using a non-linear root-search algorithm, terminating when (6.4) is satisfied to within acceptable tolerances. The fitting problem is typically good-natured, and virtually all standard root-search packages (see Press et al. [1992]) can tackle it successfully. Tanggaard [1997], for instance, uses

[8] For computational reasons, the terms multiplying the various y and y'' in (6.14) should be pre-cached, to avoid wasting effort when we ultimately perform an iteration.

a simple Gauss-Newton scheme with good results. Whatever root-search algorithm is selected, a good first guess can always be found by simple bootstrapping.

In Figure 6.4, we show the results of applying the algorithm above to a natural cubic spline representation of the yield curve example used in earlier sections. The yield curve is smooth and, unlike the Hermite spline case in Figure 6.3, the instantaneous forward curve is now differentiable, as desired.

Fig. 6.4. Yield and Forward Curve



Notes: Yield curve is assumed to be a C^2 natural cubic spline. Swap data is in Table 6.1.

While the C^2 cubic spline discussed here has attractive smoothness, it is not necessarily an ideal representation of the yield curve. As discussed in Andersen [2005] and Hagan and West [2004], among others, twice differentiable cubic spline yield curves are often subject to oscillatory behavior, spurious inflection points, poor extrapolatory behavior, and non-local behavior when prices in the benchmark set are perturbed. We shall return to the concept of non-local perturbation effects in Section 6.4 below, but for now just note that perturbation of a single benchmark price can cause a slow-decaying "ringing" effect on the C^2 cubic yield curve, with the effect of the perturbation of the benchmark instrument price spilling into the entire yield curve. This behavior is not surprising, given that the spline is constructed through a full $(N-2) \times (N-2)$ matrix system, where interpolation behavior on the interval $[T_i, T_{i+1}]$ depends on *all* values y_j, $j = 1, \ldots, N$. In contrast, the simple linear-yield bootstrapping method in Section 6.2.1 interpolation on

the interval $[T_i, T_{i+1}]$ involves only the two points y_i and y_{i+1}, and the Hermite spline approach involves only the four points $y_{i-1}, y_i, y_{i+1}, y_{i+2}$.

6.2.4 C^2 Yield Curves: Twice Differentiable Tension Splines

C^1 Hermite cubic splines are less prone to non-local behavior than C^2 cubic splines, but accomplish this in a somewhat ad-hoc fashion by giving up one degree of differentiability. Rather than taking such a draconian step, one wonders whether there may be a way to retain the C^2 feature of the cubic spline in Section 6.2.3, yet still allow control of curve locality and "stiffness". As it turns out, an attractive remedy to the shortcomings of the pure C^2 cubic spline is to insert some tension in the spline, that is, to apply a tensile force to the end-points of the spline. Appendix 6.A.3 lists the necessary details of this approach, using the classical *exponential tension spline* construction[9] in Schweikert [1966]. When applied to the yield-curve setting, the construction involves a modification of the cubic equation (6.14) for $y(T)$, $T \in [T_i, T_{i+1}]$, to

$$\begin{aligned} y(T) = &\left(\frac{\sinh\left(\sigma\left(T_{i+1} - T\right)\right)}{\sinh\left(\sigma h_i\right)} - \frac{T_{i+1} - T}{h_i} \right) \frac{y_i''}{\sigma^2} \\ &+ \left(\frac{\sinh\left(\sigma\left(T - T_i\right)\right)}{\sinh\left(\sigma h_i\right)} - \frac{T - T_i}{h_i} \right) \frac{y_{i+1}''}{\sigma^2} \\ &+ y_i \frac{T_{i+1} - T}{h_i} + y_{i+1} \frac{T - T_i}{h_i}, \end{aligned} \tag{6.16}$$

where $\sigma \geq 0$ is the *tension factor*, and where we recall the definition $h_i = T_{i+1} - T_i$.

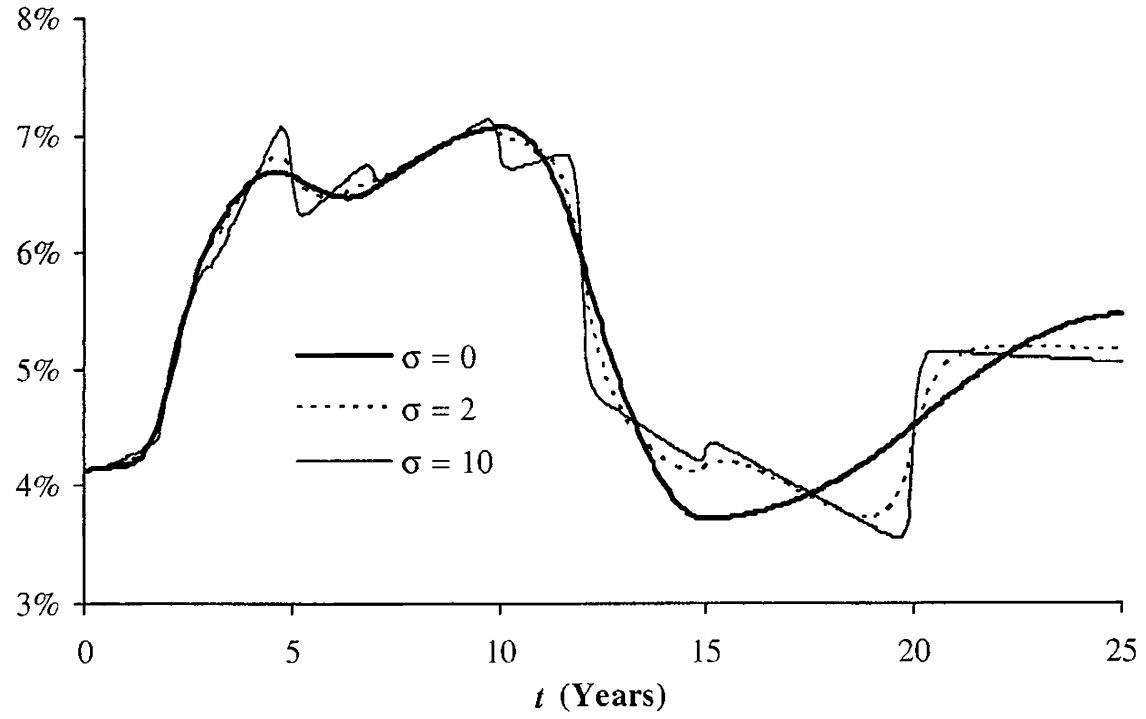
Appendix 6.A.3 discusses a number of properties of tension splines, the most important perhaps being the fact that setting $\sigma = 0$ will recover the ordinary C^2 cubic spline, whereas letting $\sigma \to \infty$ will make the tension spline uniformly approach a linear spline (i.e. the spline we used in Section 6.2.1.1). Loosely, we can thus think of a tension spline as a twice differentiable hybrid between a cubic spline and a linear spline. Equally loosely: as we increase σ, spurious inflections and ringing in the cubic spline are gradually "stretched" out of the curve, accompanied by rising (absolute values of) second derivatives at the knot points. More details on tension splines can be found in Andersen [2005], who also discusses application of computationally efficient local spline bases and the usage of a T-dependent tension factor to gain further control of the curve.

We observe that (6.16) is structurally similar to (6.14), and allows for a matrix representation of the same form as (6.62), albeit with suitably

[9] The exponential tension spline is not the only class of twice differentiable tension splines, but is probably the most common. Other classes are discussed in Kvasov [2000] and Andersen [2005].

modified definitions of the vector $\mathbf{M}$ and the matrices $\mathbf{B}$ and $\mathbf{C}$; we leave these modifications as an exercise to the reader. Suffice to say that once a value of σ has been decided upon, the numerical search for the unknown levels y_i, $i = 1, \ldots, N$, can proceed along the same principles as in Section 6.2.3 above. Figure 6.5 below shows an example; notice how increasing the tension parameter gradually moves us from cubic spline behavior to bootstrap behavior.

Fig. 6.5. Forward Curve



Notes: The yield curve is constructed as a C^2 natural tension spline, with tension parameters as given in the graph (only the forward curve $f(t)$ is shown). Swap data is in Table 6.1.

Remark 6.2.1. If the tension spline is applied not to yields, but to the logarithm of discount factors $\ln P(t)$, the limit of $\sigma \to \infty$ will produce a piecewise flat forward curve, as in Figure 6.2.

The reader may at this point wonder whether there are any firm rules as to what σ should be. We have no definitive answers to this question, and we do not try to determine σ automatically (although such routines do exist, see Renka [1987]). Instead, we normally treat σ as an "extra knob" that allow users to balance curve smoothness, shape preservation, and perturbation locality to their particular tastes. Inevitably some element of experimentation is required here.

6.3 Non-Parametric Optimal Yield Curve Fitting

The techniques we have outlined so far generally suffice for constructing a discount curve from a "clean" set of non-duplicate benchmark securities, including the carefully selected set of liquid staggered-maturity deposits, futures, and swaps most banks assemble for the purpose of constructing a Libor yield curve. In some settings, however, the benchmark set may be significantly less well-structured, involving illiquid securities with little order in their cash flow timing and considerable noise in their prices. This situation may, say, arise when one attempts to construct a yield curve from corporate bonds. While construction of a Libor curve is the most important task for the purposes of this book, we nevertheless wish to say a few words about techniques suitable for less cooperative benchmark security sets. These techniques can also be applied to Libor curve construction, of course, and are particularly relevant for applications where we are willing to sacrifice some precision in the fit to benchmark prices in return for a smoother yield curve.

6.3.1 Norm Specification and Optimization

When the input benchmark set is noisy, a direct solution of (6.4) may be erratic or may not exist. To overcome this, and to reflect that noise in the input data may make us content to solve (6.4) only to within certain error bounds, we now proceed to replace this equation with a problem of minimization of a penalized least-squares norm. Specifically, define the space $\mathcal{A} = C^2([t_1, t_M])$ of all functions $[t_1, t_M] \to \mathbb{R}$ that are twice differentiable with continuous second derivative, and introduce the M-dimensional discount vector

$$\mathbf{P}(y) = \left(e^{-y(t_1)t_1}, \ldots, e^{-y(t_M)t_M}\right)^\top.$$

Also, let $\mathbf{W}$ be a diagonal $N \times N$ weighting matrix. Then, as our best estimate $\widehat{y}$ of the yield curve we will here use

$$\widehat{y} = \operatorname*{argmin}_{y \in \mathcal{A}} \mathcal{I}(y), \tag{6.17}$$

with

$$\mathcal{I}(y) \triangleq \frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(y)\right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(y)\right) + \lambda \left(\int_{t_1}^{t_M} \left[y''(t)^2 + \sigma^2 y'(t)^2\right] dt \right), \tag{6.18}$$

where λ and σ^2 are positive constants. The norm $\mathcal{I}(y)$ consists of three separate terms:

- A least-squares penalty term

$$\frac{1}{N}\left(\mathbf{V}-\mathbf{cP}(y)\right)^{\top}\mathbf{W}^{2}\left(\mathbf{V}-\mathbf{cP}(y)\right)$$
$$=\frac{1}{N}\sum_{i=1}^{N}W_{i}^{2}\left(V_{i}-\sum_{j=1}^{M}c_{i,j}e^{-y(t_{j})t_{j}}\right)^{2},$$

  where W_i is the i-th diagonal element of $\mathbf{W}$. This term is an outright precision-of-fit norm and measures the degree to which the constructed discount curve can replicate input security prices. The weights W_i can be used to assign different importance to the various securities in the benchmark set, and/or to translate the precision of the fit from raw dollar amounts into more intuitive quantities, such as security-specific quoted yields[10]. Clearly, if (6.4) can be satisfied, then the least-squares penalty term will attain its minimum (of zero) for all yield curves that satisfy (6.4).
- A weighted smoothness term $\lambda\int_{t_1}^{t_M} y''(t)^2\,dt$, penalizing high second-order derivatives of y to avoid kinks and discontinuities.
- A weighted curve-length term $\lambda\sigma^2\int_{t_1}^{t_M} y'(t)^2\,dt$, penalizing oscillations and excess convexity/concavity.

Our choice of calibration norm is, we believe, an attractive one, but other choices obviously are available as well. For instance, in Adams and van Deventer [1994] the norm contains no curve-length term and the smoothing norm is expressed on the forward curve, rather than on the yield curve. Due to the lack of the curve-length penalty term, the resulting curve will tend to behave like the C^2 cubic spline in Section 6.2.3; see Hagan and West [2004] for some numerical tests.

The following result is shown by variational methods in Andersen [2005]:

Proposition 6.3.1. *The curve $\widehat{y}$ that satisfies (6.17) is a natural exponential tension spline with tension factor σ and knots at* all *cash flow dates $t_1, t_2, \ldots, t_M$.*

Proposition 6.3.1 establishes that the curve we are looking for is a tension spline with tension factor σ, but does not in itself allow us to identify the optimal spline directly, beyond the fact that i) it is a natural spline with boundary conditions $y''(t_1) = y''(t_M) = 0$; and ii) it has knots at all t_j, $j = 1, \ldots, M$. Identification of the correct spline involves solving for unknown

[10]Most fixed-income securities are quoted through some type of yield, e.g. $V_i = g_i(r_i)$ where r_i is the quoted yield and g_i is a function that encapsulates the quoting convention. The quantity $D_i = -dg_i/dr_i$ is known as the *duration* of V_i. Setting $W_i = 1/D_i$ in the least-squares norm will turn price deviations into yield deviations.

levels[11] $y(t_1), y(t_2), \ldots, y(t_M)$ to optimize directly (6.18). In this exercise, the following lemma is useful.

Lemma 6.3.2. *For a natural tension spline interpolating the values* $y(t_1)$, $y(t_2)$, ..., $y(t_M)$, *we have*

$$y'(t_j) = \left(-\sigma \frac{\cosh(\sigma(t_{j+1} - t_j))}{\sinh(\sigma(t_{j+1} - t_j))} + \frac{1}{t_{j+1} - t_j}\right) \frac{y''(t_j)}{\sigma^2}$$
$$+ \left(\frac{\sigma}{\sinh(\sigma(t_{j+1} - t_j))} - \frac{1}{t_{j+1} - t_j}\right) \frac{y''(t_{j+1})}{\sigma^2} + \frac{y(t_{j+1})}{t_{j+1} - t_j} - \frac{y(t_j)}{t_{j+1} - t_j},$$

and

$$\lambda \left(\int_{t_1}^{t_M} \left[y''(t)^2 + \sigma^2 y'(t)^2 \right] dt \right) = -\lambda \sum_{j=1}^{M-1} d_j \left(y(t_{j+1}) - y(t_j) \right), \quad (6.19)$$

where $y''(t_1) = y''(t_M) = 0$, *and*

$$d_j \triangleq \frac{y''(t_{j+1}) - \sigma^2 y(t_{j+1})}{t_{j+1} - t_j} - \frac{y''(t_j) - \sigma^2 y(t_j)}{t_{j+1} - t_j}. \quad (6.20)$$

Proof. The result for $y'(t_j)$ follows from direct differentiation of the basic equations for a tension spline (see (6.16) above, applied to the knot grid $\{t_j\}$). To show (6.19), consider the interval $[t_j, t_{j+1}]$ and the integral

$$\int_{t_j}^{t_{j+1}} \left(y''(t)^2 + \sigma^2 y'(t)^2 \right) dt = \int_{t_j}^{t_{j+1}} \left(y''(t) \cdot y''(t) + \sigma^2 y'(t) \cdot y'(t) \right) dt.$$

Integration by parts yields

$$\int_{t_j}^{t_{j+1}} \left(y''(t)^2 + \sigma^2 y'(t)^2 \right) dt$$
$$= \left[y''(t) y'(t) \right]_{t_j}^{t_{j+1}} - \int_{t_j}^{t_{j+1}} \left(y^{(3)}(t) - \sigma^2 y'(t) \right) y'(t)\, dt$$
$$= y''(t_{j+1}) y'(t_{j+1}) - y''(t_j) y'(t_j) - d_j \left(y(t_{j+1}) - y(t_j) \right), \quad (6.21)$$

where d_j is given in (6.20). Here, we have used that, by definition, hyperbolic tension splines have $y^{(3)}(t) - \sigma^2 y'(t)$ piecewise constant and equal to d_j on each interval $[t_j, t_{j+1}]$ (see equation (6.63) in Appendix 6.A). The result (6.20) follows by addition of the terms (6.21) and using the condition $y''(t_1) = y''(t_M) = 0$. □

[11] In Andersen [2005], the search for yield levels has been replaced by the more contemporary idea of searching for weights in a local basis representation of the spline.

Lemma 6.3.2 shows us that we can compute the value of the integral penalty term in (6.18) directly from knowledge of yield levels $y(t_1), \ldots, y(t_M)$ and second derivatives $y''(t_2), \ldots, y''(t_{M-1})$. For each guess for the M unknown levels $y(t_1), y(t_2), \ldots, y(t_M)$ we can proceed as follows.

1. Compute the least-squares penalty term $\frac{1}{N}(\mathbf{V} - \mathbf{cP}(y))^\top \mathbf{W}^2(\mathbf{V} - \mathbf{cP}(y))$ directly from the definition of $\mathbf{P}(y)$.
2. Use the results in Section 6.2.4 to solve for $y''(t_2), \ldots, y''(t_{M-1})$ by solving a tri-diagonal set of equations.
3. Use Lemma 6.3.2 to compute $\lambda(\int_{t_1}^{t_M} [y''(t)^2 + \sigma^2 y'(t)^2]\, dt)$, thereby completing the computation of the norm $\mathcal{I}(y)$.

Embedding Steps 1–3 above in a multi-variate numerical optimizer ultimately allows us to determine the optimal solution $\widehat{y}$. A good generic routine for this optimization step would be the Levenberg-Marquardt algorithm; see Press et al. [1992]. The optimization problem at hand is generally good-natured, and one can also use a simpler Gauss-Newton method, as discussed in Andersen [2005]. If possible, it is often useful to use a simpler method (e.g. bootstrapping) to establish a good guess for the yield curve levels $y(t_1), y(t_2), \ldots, y(t_M)$. A proper implementation of the algorithm should typically construct a yield curve in less than one-tenth of a second on a standard PC.

Remark 6.3.3. If we let $\sigma = 0$, the solution to the optimization problem becomes a *cubic smoothing spline*; see Tanggaard [1997] for more details on this case.

Remark 6.3.4. If we let $\lambda \downarrow 0$, the resulting spline will often end up hitting all benchmark prices exactly, i.e. will satisfy (6.4) in the limit. The resulting spline is then the optimal *interpolating* curve, in the sense that of all twice differentiable yield curves that match the benchmark prices, the spline is the minimizer of the regularity term $\int_{t_1}^{t_M} [y''(t)^2 + \sigma^2 y'(t)^2]\, dt$. If, for $\lambda \downarrow 0$, we do not satisfy (6.4), then the resulting spline can be considered a *least-squares regression spline* solution.

6.3.2 Choice of λ

So far, we have assumed that the parameter λ has been specified exogenously by the user. In practice, however, a good magnitude of λ may sometimes be hard to ascertain by inspection, and a procedure to estimate λ directly from the data is often useful. One possibility is to use a cross-validation approach, either outright or through the more efficient Generalized Cross-Validation (GCV) criterion by Craven and Wahba [1979]. Some results along these lines can be found in Tanggaard [1997] and Andersen [2005], but are outside the scope of our treatment here. A more pragmatic approach is to replace the optimization problem (6.17) with the constrained optimization problem

$$\widehat{y} = \underset{y \in \mathcal{A}}{\operatorname{argmin}} \int_{t_1}^{t_M} [y''(t)^2 + \sigma^2 y'(t)^2]\, dt, \tag{6.22}$$

$$\text{subject to } \frac{1}{N} (\mathbf{V} - \mathbf{cP}(y))^\top \mathbf{W}^2 (\mathbf{V} - \mathbf{cP}(y)) = \gamma^2, \tag{6.23}$$

where γ is an exogenously specified constant. Note that γ is just the allowed weighted root-mean-square (RMS) error in the fit to benchmark securities, an intuitive quantity that most users should have no problem specifying directly based on, say, observed bid-offer spreads. The Lagrangian for the above problem becomes

$$\widehat{y} = \underset{y \in \mathcal{A}}{\operatorname{argmin}} \left(\int_{t_1}^{t_M} [y''(t)^2 + \sigma^2 y'(t)^2]\, dt + \rho \left[\frac{1}{N} (\mathbf{V} - \mathbf{cP}(y))^\top \mathbf{W}^2 (\mathbf{V} - \mathbf{cP}(y)) - \gamma^2 \right] \right), \tag{6.24}$$

where the Lagrange multiplier ρ must be determined such that the constraint (6.23) is satisfied at the optimum of (6.24). Apart from a constant scale, (6.24) is identical to (6.17), so we solve the constrained optimization problem (6.22)–(6.23) through the following iteration over λ:

1. Given a guess for λ, find the optimum value of $y(t_1), y(t_2), \ldots, y(t_M)$, as a solution of (6.17).
2. Compute $\mathcal{S} = \frac{1}{N}(\mathbf{V} - \mathbf{cP}(y))^\top \mathbf{W}^2 (\mathbf{V} - \mathbf{cP}(y))$.
3. If $\mathcal{S} = \gamma^2$, stop; otherwise update λ and go to Step 1.

In Step 1, we can proceed as discussed in Section 6.3.1 above. In general, the precision norm $\mathcal{S} = \mathcal{S}(\lambda)$ will be a declining function in λ and, provided that a root to $\mathcal{S}(\lambda) = \gamma^2$ exists[12], the updating in Step 3 can be done by any standard root search algorithm.

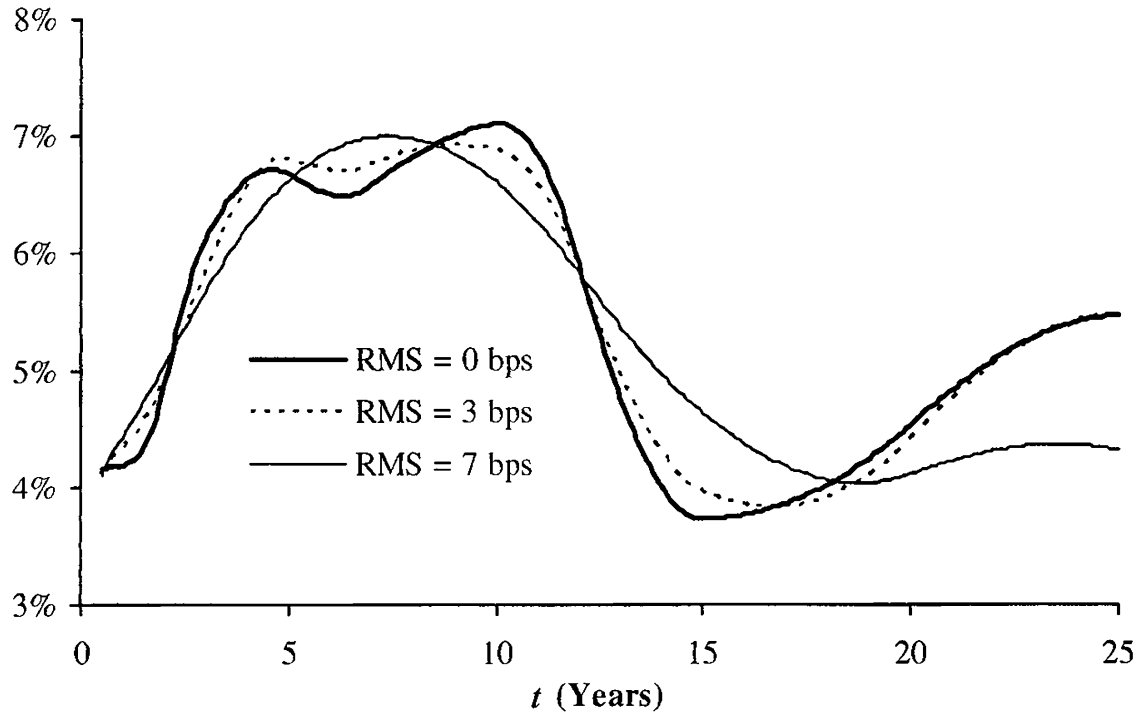
6.3.3 Example

To illustrate the effect of λ, we now apply the algorithm in Section 6.3.2 to the test data in Table 6.1 above. In doing so, we use the matrix $\mathbf{W}$ to normalize (see footnote 10) all price errors to yield-to-maturity errors, allowing us to consider γ in (6.23) as the root-mean-square (RMS) yield error. Setting $\sigma = 0$, the forward curves for various choices of γ are shown in Figure 6.6. As one would expect, the higher we allow γ to be, the smoother the forward (and yield) curves become.

For our test case, the zero-RMS optimal (M-knot) forward curve in Figure 6.6 is virtually identical to the N-knot cubic spline solution in Figure 6.4. In general, the N-knot interpolating curve can be interpreted as a

[12] There may be instances where $\mathcal{S}(0) > \gamma^2$. If the desired precision is unattainable, we can either increase γ^2 or perhaps prune the benchmark security set.

Fig. 6.6. Forward Curve



Notes: The yield curve is constructed as an optimal C^2 natural tension spline, with an RMS yield error constraint as listed in the graph (only the forward curve $f(t)$ is shown). The tension parameter is set to $\sigma = 0$ for all curves. Swap data is in Table 6.1.

constrained solution to (6.17) with $\lambda = 0$, with the constraint requiring that knots be placed only at benchmark maturities $\{T_i\}_{i=1}^N$, rather than at all cash flow dates $\{t_j\}_{j=1}^M$. The effect of enforcing this additional constraint is often rather small, at least for the purposes of constructing a Libor curve.

6.4 Managing Yield Curve Risk

Consider a portfolio of securities with value V_0, where V_0 is a function of the yield curve $y(t)$. The securities in V_0 would typically not be in the benchmark set and could contain, say, interest rate options, seasoned swaps, and so forth. As the yield curve is a function of the benchmark set values $\mathbf{V} = (V_1, \ldots, V_N)^\top$, we may write

$$V_0 = V_0\left(V_1, \ldots, V_N; \theta\right),$$

where the vector θ contains model parameters (e.g. volatilities) and where the function $V_0(\cdot)$ is determined both from the valuation model of the security in question, and from the curve construction algorithm employed. Clearly, then

$$dV_0 = \sum_{i=1}^{N} \frac{\partial V_0}{\partial V_i} dV_i + \sum_i \frac{\partial V_0}{\partial \theta_i} d\theta_i,$$

or, for non-infinitesimal moves,

$$\Delta V_0 \approx \sum_{i=1}^{N} \frac{\partial V_0}{\partial V_i} \Delta V_i + \sum_i \frac{\partial V_0}{\partial \theta_i} \Delta \theta_i. \tag{6.25}$$

For the purpose of managing first-order risk exposure to moves in the yield curve, (6.25) suggests that the collection of derivatives $\partial V_0 / \partial V_i$, $i = 1, \ldots, N$ — often called *(bucketed) interest rate deltas* — forms a natural metric for portfolio risk. In particular, if all these derivatives are zero, our portfolio would, to first order, be immunized against any move in the yield curve that is consistent with the chosen curve construction algorithm. On the other hand, if some or all of the derivatives are non-zero, we could manage our risk by setting up a hedge portfolio of benchmark securities, with notional $-\partial V_0 / \partial V_i$ on the i-th security. We emphasize that the resulting hedge would typically *not* be model-consistent: most interest models assume that yield curve risk originates from only a few stochastic yield curve factors that tend to move the curve smoothly[13], in a predominantly parallel fashion. Theoretically, a bucket-by-bucket immunization against all terms ΔV_i may then be considered an overkill — we typically hedge against far too many risk factors (N) — but is nevertheless standard industry practice and has proven to be robust. Notice that bucket hedging along these lines would, for instance, correctly reject the notion that we could perfectly hedge a 20 year swap with a 1 month FRA, something that a one-factor interest rate model (see Chapter 4 and Chapter 10) would happily accept. We pick up this subject again in Chapter 22.

6.4.1 Par-Point Approach

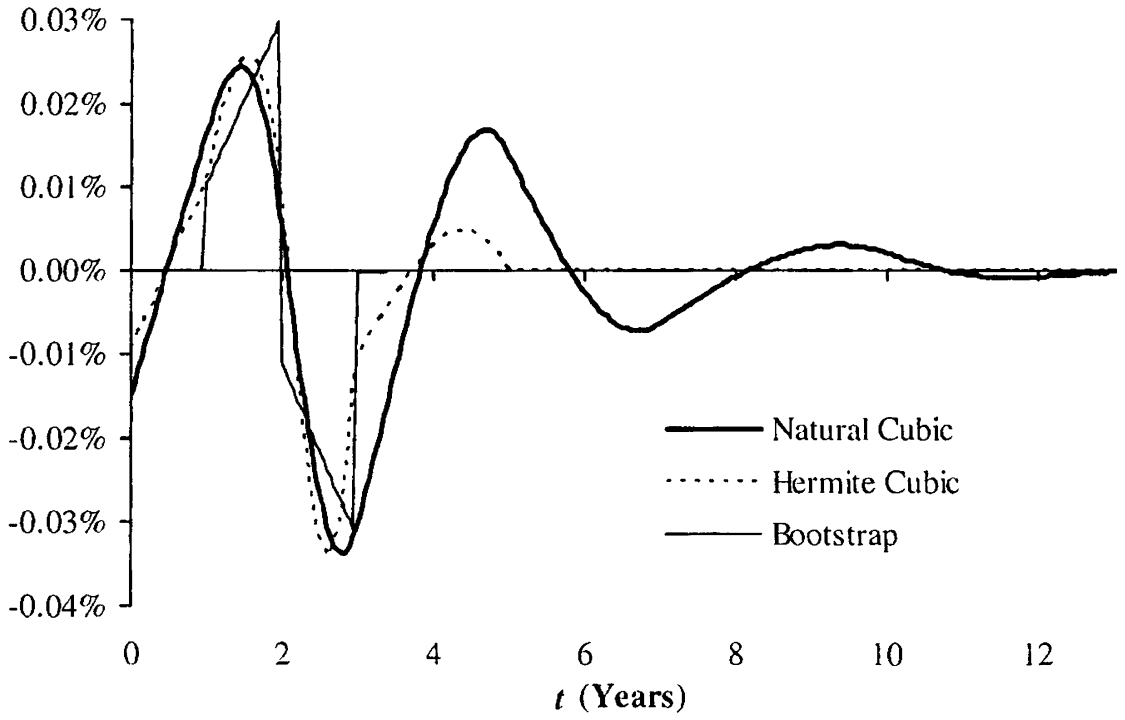
The simplest approach to computation of the delta $\partial V_0 / \partial V_i$ involves a manual bump[14] to V_i, followed by a reconstruction of the yield curve, and a subsequent repricing of the portfolio V_0. This procedure is sometimes known as the *par-point approach*, and resulting derivatives *par-point deltas*. For the approach to work properly, it is important that the yield curve construction algorithm is fast and produces clean, local perturbations of the yield curve when benchmark prices are shifted. For instance, perturbing a short-dated FRA price should not cause noticeable movements in long-term yields, lest we reach the erroneous conclusion (again) that we can perfectly hedge a 20 year swap with a 1 month FRA. As we have discussed earlier, Hermite splines and bootstrapped yield curves both exhibit good perturbation locality, but cubic C^2 splines often do not. To illustrate this, Figure 6.7 considers the

[13] See the principal components analysis in Chapter 14 for more on this.

[14] In practice, rather than bumping the price V_i outright, one may instead bump the yield of the i-th benchmark security (typically by 1 basis point). See also footnote 10.

effect on the forward curves in Figures 6.1, 6.3, and 6.4 from a 1 basis point up-move in the par yield of the 2 year swap in Table 6.1. As we can see, the move causes a noisy, ringing perturbation in the C^2 cubic spline solution, spreading into short- and long-dated parts of the forward curve.

Fig. 6.7. Forward Curve Move



Notes: Change in instantaneous forward curve, from a 1 basis point shift in the 2 year swap yield in Table 6.1. The curve construction methods tested are: bootstrapping with piecewise linear yields ("Bootstrap"), Hermite C^1 cubic spline ("Hermite"), and C^2 natural cubic spline ("Natural Cubic"). Swap data is in Table 6.1.

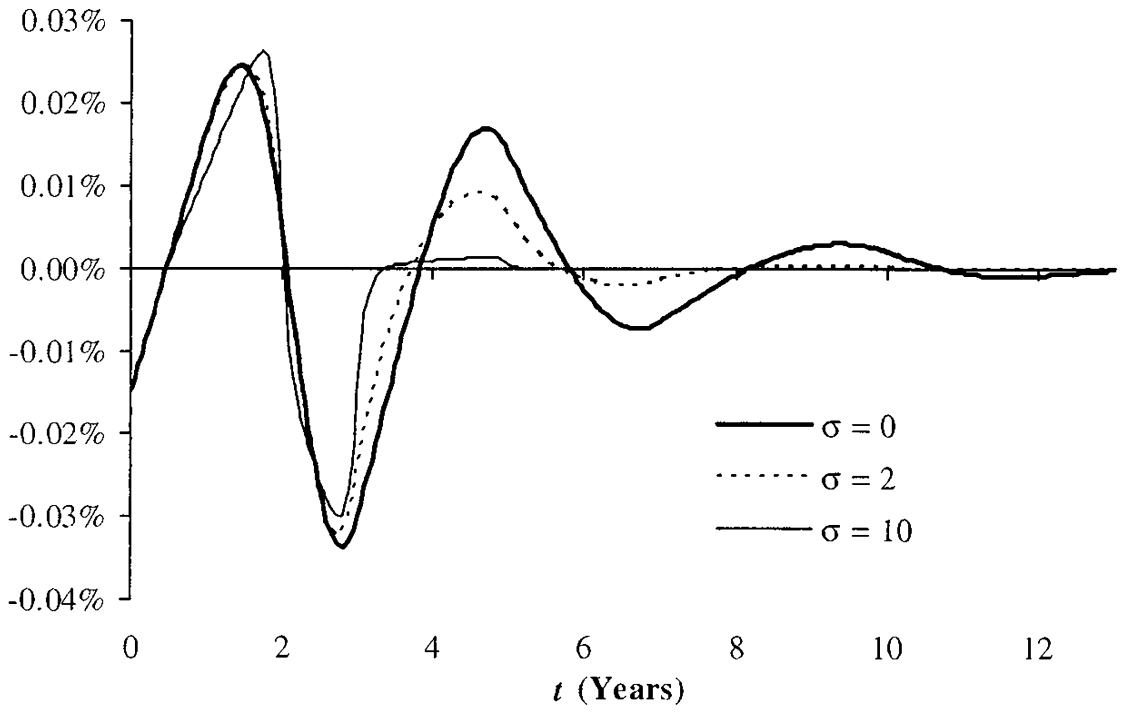
In Figure 6.8, we have followed the recommendations of Section 6.2.4 and added tension to the C^2 spline, causing a dampening of the perturbation noise. Clearly, the usage of a tension factor can have a beneficial impact on risk reports produced by the par-point approach.

6.4.2 Forward Rate Approach

As an alternative to direct perturbation of benchmark security prices, we can consider applying perturbations directly to the discount curve, thereby mostly avoiding the introduction of artifacts specific to the curve construction algorithm. In practice, this technique typically focuses on the forward curve[15] $f(t)$, to which we apply certain functional shifts $\mu_k(t)$, $k = 1, \ldots, K$. Writing

[15] Perturbations may also be performed on discretely, rather than continuously, compounded forward rates.

Fig. 6.8. Forward Curve Move



Notes: Change in instantaneous forward curve, from a 1 basis point shift in the 2 year swap yield in Table 6.1. The yield curve was constructed as a tension spline, with tension factors as given in the graph. Swap data is in Table 6.1.

(loosely) $V_0 = V_0(f)$ to highlight the dependence of V_0 on the forward curve, we then compute functional (Gâteaux) derivatives[16] for V_0:

$$\partial_k V_0 = \left. \frac{dV_0\left(f(t) + \varepsilon \mu_k(t)\right)}{d\varepsilon} \right|_{\varepsilon=0}, \quad k = 1, \ldots, K. \tag{6.26}$$

Standard choices for $\mu_k(t)$ are

$$\text{Piecewise Triangular:} \quad \mu_k(t) = \frac{t - t_{k-1}}{t_k - t_{k-1}} 1_{\{t \in [t_{k-1}, t_k)\}} + \frac{t_{k+1} - t}{t_{k+1} - t_k} 1_{\{t \in [t_k, t_{k+1})\}}, \tag{6.27}$$

$$\text{Piecewise Flat:} \quad \mu_k(t) = 1_{\{t \in [t_k, t_{k+1})\}}, \tag{6.28}$$

where $\{t_k\}$ is a user-specified discretization grid. The resulting sensitivities are often called *forward rate deltas.*

It is common practice to use $\{t_k\}$ grids spaced three months apart, with dates on Eurodollar futures maturities. The number of deltas K is thus typically a rather large number, and the K derivatives $\partial_k V_0$ give a detailed picture of where the portfolio risk is concentrated on the forward curve. As forward rate contracts and Eurodollar futures cease to be liquid beyond 4

[16] For a proper definition of the Gâteaux derivative, see Gâteaux [1913].

or 5 year maturities, the forward rate deltas do not directly suggest hedging instruments for the medium and long end of the yield curve exposure; however it is not difficult to translate forward rate deltas into a hedging portfolio (see the next section). The choice of par point versus forward rate deltas is largely a matter of personal preference, and it is not uncommon for traders to use both at the same time.

6.4.3 From Risks to Hedging: The Jacobian Approach

A collection of forward rate shifts $\mu_k(t)$, $k = 1, \ldots, K$, defines a certain view on the (first-order) risk on the portfolio $V_0(f)$ via the functional derivatives (6.26). In order to be useful, this risk view ultimately needs to be translated into a portfolio of hedging instruments that offsets the risks of V_0. While fixed income traders normally are quite adept at mentally translating forward rate risk into actual hedge transactions, some linear algebra can help out in this exercise, as we now show.

Suppose a set of L hedging instruments is available, with values $\mathbf{H} = (H_1, \ldots, H_L)^\top$. This set may or may not coincide with the benchmark set used for curve construction; for example, one may want to exclude some benchmark securities from the hedging set due to poor liquidity, or one may want to add instruments to the benchmark set to fine-tune hedging. Using (6.26), we denote the sensitivities of hedging instruments to the shifts $\mu_k(t)$ by $\partial_k H_l$, $l = 1, \ldots, L$, $k = 1, \ldots, K$. If the l-th hedging instrument is included in the hedging portfolio with notional weight p_l, and $\mathbf{p} = (p_1, \ldots, p_L)^\top$, then the sensitivity of the hedge portfolio value

$$H_0(\mathbf{p}) = \mathbf{p}^\top \mathbf{H}$$

to the k-th perturbation is given by

$$\partial_k H_0(\mathbf{p}) = \mathbf{p}^\top \partial_k \mathbf{H},$$

where we have denoted

$$\partial_k \mathbf{H} = (\partial_k H_1, \ldots, \partial_k H_L)^\top.$$

In most cases[17] we would like to choose the weights $\mathbf{p}$ in such a way that $\partial_k H_0(\mathbf{p})$ offsets as much of $\partial_k V_0$ as possible, for all $k = 1, \ldots, K$. Let W_k be the relative importance of offsetting the k-th derivative, and U_l a relative "reluctance" to using the l-th hedging instrument (a function of the bid-ask spread, for example). Then, the optimal hedging weights $\widehat{\mathbf{p}}$ can be defined by the condition

[17] Sometimes traders deliberately wish to keep some risk on their books, as a way to speculate on interest rate movements. A non-zero target risk profile is easily accommodated by a change of the optimization target in (6.29).

$$\widehat{\mathbf{p}} = \underset{\mathbf{p}}{\operatorname{argmin}} \left(\sum_{k=1}^{K} W_k^2 \left(\partial_k H_0 \left(\mathbf{p} \right) - \partial_k V_0 \right)^2 + \sum_{l=1}^{L} U_l^2 p_l^2 \right). \tag{6.29}$$

Define the matrix $\partial \mathbf{H}$ to have columns $\partial_1 \mathbf{H}, \ldots, \partial_K \mathbf{H}$, the vector $\partial \mathbf{V}_0$ by

$$\partial \mathbf{V}_0 = \left(\partial_1 V_0, \ldots, \partial_K V_0 \right)^\top,$$

the matrix $\mathbf{W}$ to be diagonal with W_k's on the diagonal, and the matrix $\mathbf{U}$ to be diagonal with U_l's on the diagonal. With this notation (6.29) can be recast as a least-squares problem,

$$\left(\partial \mathbf{H}^\top \mathbf{p} - \partial \mathbf{V}_0 \right)^\top \mathbf{W}^2 \left(\partial \mathbf{H}^\top \mathbf{p} - \partial \mathbf{V}_0 \right) + \mathbf{p}^\top \mathbf{U}^2 \mathbf{p} \to \min. \tag{6.30}$$

The problem (6.30) can be solved by standard methods; a formal solution is given by the linear system

$$\left(\partial \mathbf{H}\, \mathbf{W}^2\, \partial \mathbf{H}^\top + \mathbf{U}^2 \right) \widehat{\mathbf{p}} = \partial \mathbf{H}\, \mathbf{W}^2\, \partial \mathbf{V}_0. \tag{6.31}$$

We note that the addition of the $\mathbf{U}$ term to the optimization problem (6.31) is sometimes called *Tikhonov regularization*, a technique that we shall return to in Chapter 18.

When solving (6.30), one should carefully consider the relative dimensions of the matrices involved. First, if there are fewer hedging instruments than shifts to be immunized ($L < K$), then, in general, not all risks can be offset. In this case, the weights $\mathbf{W}$ gain in importance as they allow the user to focus hedging on risk buckets deemed more important, at the expense of other, less critical ones. Also, when $L < K$ the weights $\mathbf{U}$ are less important, and in most cases can safely be set to zero. Second, if there are more hedging instruments than risk buckets to immunize against ($L > K$), then there are typically multiple hedging portfolios that perfectly offset all risks. In this case, the weights $\mathbf{W}$ can normally be ignored (all set to 1), but the weight matrix $\mathbf{U}$ becomes more critical as it allows one to choose which of the possible hedging portfolios one "likes" best (e.g., the least costly). Finally, if $L = K$, then normally there exists exactly one portfolio that hedges all risks. Both $\mathbf{W}$ and $\mathbf{U}$ are then often of little consequence, although one might still want to specify non-zero weights $\mathbf{U}$ to avoid oscillatory or unstable solutions if the linear equations are ill-posed. We note that in the simple case of $L = K$, $\mathbf{W} = 1$, $\mathbf{U} = \mathbf{0}$ and $\partial \mathbf{H}$ invertible, the solution to the optimization problem is given by

$$\mathbf{p} = \left(\partial \mathbf{H}^\top \right)^{-1} \partial \mathbf{V}_0. \tag{6.32}$$

The method of constructing a hedge portfolio from derivatives to arbitrary shocks of the forward curve via the optimization problem (6.30) is known as the *Jacobian method* for interest rate deltas; the name originates from the fact that the matrix $\partial \mathbf{H}$ is the Jacobian matrix of a hedge set with respect to the forward curve shocks. Combined with the forward rate deltas,

the Jacobian method helps aggregate fine-grained risks into various sets of hedges. The approach has considerable generality as the risk basis functions μ_k and the hedge portfolio can be chosen freely by the user — note, for instance, that even the par-point approach can be seen as a special type of the Jacobian method where we effectively choose the hedging set to coincide perfectly with the benchmark set and where the μ_k's are set to be the shifts of the forward curve that correspond to the bumps of benchmark securities. In this special case, the Jacobian $\partial \mathbf{H}$ is then a unit matrix and (from (6.32)) the original par-point deltas are recovered.

The Jacobian method serves to decouple risk calculations from curve construction. This, potentially, allows for combining smooth curves with localized risk, a feat that is difficult to achieve by other methods. The Jacobian is also useful in applications where curves need to be rebuild over and over, to address the fact that Libor and Treasury benchmark security prices (or yields) change very quickly, often quicker than a sophisticated curve construction algorithm can rebuild the curves. With the aid of a Jacobian, changes in benchmark prices can be quickly translated into changes of the forward curve via a matrix multiplication. A full curve rebuild needs only be triggered when the benchmark prices have moved sufficiently far from their initial values.

6.4.4 Cumulative Shifts and other Common Tricks

As evident in Figure 6.7 (the bootstrap case), a shift to a single swap rate (while keeping other swap rates fixed) typically results in a strong "see-saw" impact on the forward rate curve. Let us attempt to gain some intuition about the magnitude of the forward rate shock. For a back-of-the-envelope calculation, we can assume that a swap rate is a linear combination of forward Libor rates (see (4.11)),

$$S_n \approx \sum_{i=1}^{n} w_{i,n} L_i,$$

where S_n denotes a swap rate for a swap covering n periods (for simplicity assume that each period is 1 year), L_i denotes a forward Libor rate for i-th period (from year $i-1$ to year i), and $w_{i,n} \approx 1/n$. Inverting this relationship yields

$$L_n \approx n \left(S_n - \frac{n-1}{n} S_{n-1} \right). \tag{6.33}$$

As part of a par-point report, assume now that S_n is shifted by the amount δ, but S_{n-1} and S_{n+1} remain unchanged. According to (6.33) L_n will then shift by approximately $n\delta$, and L_{n+1} by $-n\delta$. For instance, if a 30 year swap yield is shifted by 1 basis point, while 29 year and 31 year are kept unchanged, then evidently the forward Libor rate L_{30} will move by 30

basis points, and the rate L_{31} will move by -30 basis points. If the portfolio whose deltas we are computing happens to contain, say, a spread option on the difference between L_{30} and L_{31}, the underlying rate of this option would be shifted by 60 basis points (!). And clearly, a shift of 60 basis points (or 30 basis points, for that matter) is not small, and may be inappropriate for calculating a first-order derivative. We emphasize that what appears to be a benign 1 basis point rate shift translates into a much larger forward curve move that can potentially affect underlying instruments in unexpected ways.

The example above highlights the importance of applying shifts to the forward curve that are consistent with real moves of interest rates. Obviously, it is highly unlikely that a forward curve would move in such a way that a 30 year swap rate has changed but the 29 and 31 year rates have not.

One tweak to the standard par-point approach that goes some way towards the goal of realistic curve shifts is the so-called *cumulative par-point approach* (also known as a *waterfall par-point approach*). The idea is simple: the shift to the i-th benchmark security is retained while calculating the derivative to the $(i+1)$-th (and subsequent) securities. In other words, the two curves for the $(i+1)$-th derivative are constructed from the prices

$$(V_1 + \Delta V_1, \ldots, V_i + \Delta V_i, V_{i+1}, V_{i+2}, \ldots, V_N)$$

(base) and

$$(V_1 + \Delta V_1, \ldots, V_i + \Delta V_i, V_{i+1} + \Delta V_{i+1}, V_{i+2}, , \ldots, V_N)$$

(perturbed). The standard deltas are then computed as differences of two consecutive (cumulative) derivatives. While the resulting deltas should coincide with the standard par-point deltas in the limit of $\Delta V \to 0$, they differ for non-vanishing perturbations.

The forward curve shifts implied by the cumulative par-point method are less extreme than those of the ordinary par-point method, making the cumulative par-point method quite attractive in practice. Another practical advantage of the method is the fact that the sum of deltas computed by the method is always (by definition) exactly equal to the *parallel delta*, i.e. the delta that is obtained by shifting all benchmark yields by the same amount at the same time. Because of the second-order effects, the same is only true for the standard par-point method in the limit of vanishing shifts, not for the non-infinitesimal perturbations used in practice.

The cumulative par-point approach is easy to mimic (and even improve) in the Jacobian framework of Section 6.4.3. Clearly, from (6.33), the i-th cumulative shift roughly corresponds to a piecewise flat move of the forward curve between the maturities of $(i-1)$-th and i-th benchmark. Hence, we can define

$$\mu_i(t) = 1_{\{t \in [T_{i-1}, T_i)\}}, \quad i = 1, \ldots, N, \tag{6.34}$$

with $T_0 \triangleq 0$. Note that this specification involves benchmark maturities $\{T_i\}$, in contrast to (6.28) which is typically set on a 3 month grid; in

particular, (6.34) involves as many shocks as there are benchmark securities. Application of the Jacobian method to (6.34) yields an attractive variation of the cumulative par-point method where all forward curve shocks are similarly scaled, in contrast to the basic cumulative par-point where the size of forward curve shocks grows linearly with maturity, as implied by (6.33).

We should note that to improve accuracy, one may compute deltas as the average of deltas computed using first positive shocks, then negative shocks. This idea applies to par-point, forward-rate, Jacobian, cumulative-par-point, or any other delta calculation method. For the simple par-point method, this boils down to using two-sided finite difference approximations versus one-sided for approximating derivatives, a standard trick. For other methods the relationship is not as straightforward but the end result is the same: improved accuracy and stability of deltas. Using averaged deltas is typically particularly useful for security prices that depend on yields in a strongly non-linear fashion.

Finally, let us mention another popular trick. We have spent a good part of Section 6.2.4 describing ways to build smooth yield curves that exhibit good locality under perturbations. A more simplistic approach to tackle the same problem is based on building two different curves. One — smooth — is then used for pricing and the other — bootstrapped and with good locality — used for risk computations. While certainly helpful in a pinch, this approach tends to suffer from poor P&L predict, in the sense that changes in valuations of a portfolio between two dates are not well explained by first-order sensitivities (because values and sensitivities are calculated using different curves). We spend more time on P&L predict in Chapter 22.

6.5 Various Topics in Discount Curve Construction

6.5.1 Curve Overlays and Turn-of-Year Effects

Many of the curve construction algorithms so far have been designed around the implicit idea that the forward curve should ideally be *smooth*. While this is, indeed, generally a sound principle, exceptions do exist. For instance, it may be reasonable to expect instantaneous forwards to jump on or around meetings of monetary authorities, such as the Federal Reserve in the US. In addition, other "special" situations may exist that might warrant introduction of discontinuities into the forward curve. A well-known example is the turn-of-year (TOY) effect where short-dated loan premiums spike for loans between the last business day of the year and the first business day of the following calendar year.

One common way of incorporating TOY-type effects is to exogenously specify an *overlay curve* $\varepsilon_f(t)$ on the instantaneous forward curve. Specifically, the forward curve $f(t) = f(0,t)$ is written as

$$f(t) = \varepsilon_f(t) + f^*(t), \tag{6.35}$$

where $\varepsilon_f(t)$ is user-specified — and most likely contains discontinuities around special event dates — and $f^*(t)$ is unknown. The yield curve algorithm is then subsequently applied to the construction of $f^*(t)$. That is, rather than solving $\mathbf{cP} = \mathbf{V}$ (see equation (6.4)), we instead write

$$P(T) = e^{-\int_0^T \varepsilon_f(t)dt} e^{-\int_0^T f^*(t)dt} \triangleq P_\varepsilon(T)P^*(T) \tag{6.36}$$

and solve

$$\mathbf{c}_\varepsilon \mathbf{P}^* = \mathbf{V}, \tag{6.37}$$

where $\mathbf{P}^* = (P^*(t_1), \ldots, P^*(t_M))^\top$, and $\mathbf{c}_\varepsilon$ is a modified $N \times M$ coupon matrix, with elements

$$(\mathbf{c}_\varepsilon)_{i,j} = c_{i,j} P_\varepsilon(t_j).$$

Construction of $\mathbf{c}_\varepsilon$ can be done as a pre-processing step, after which any of the algorithms discussed earlier in this chapter can be applied to attack (6.37). Once the curve $P^*(t)$ (or, equivalently, the yield curve $y^*(t) = -t^{-1} \ln P^*(t)$) has been constructed, any subsequent use of the curve for cash flow discounting requires, according to (6.36), a multiplicative adjustment of time t discount factors by the quantity $P_\varepsilon(t)$.

6.5.2 Cross-Currency Curve Construction

In this section we consider the issues involved in constructing yield curves simultaneously in multiple currencies. As it turns out, the market for foreign exchange (FX) forwards and cross-currency basis swaps imposes certain arbitrage constraints that must be considered in the curve construction exercise.

6.5.2.1 Basic Problem

To provide some motivation, consider a US dollar (USD) based firm receiving \$1 for certain at some future time T. Assuming that we have available a risk-free discount curve $P(\cdot)$ for USD-denominated cash flows, we compute the value of this security simply as $P_\$(T)$. Suppose now that the firm enters into a (costless) FX forward where it commits to pay \$1 at time T against receipt of a Japanese yen (JPY) amount of ¥$Y(T)$; $Y(T)$ thereby represents the time 0 forward JPY/USD exchange rate for delivery at time T. By transacting the FX forward, the firm has effectively turned the receipt of \$1 into receipt of ¥$Y(T)$, the USD PV at time 0 of which is

$$X(0)P_¥(T)Y(T),$$

where $P_¥(T)$ is a JPY discount factor and $X(0)$ is the time 0 foreign exchange rate in \$/¥ terms. To avoid an arbitrage, we evidently need

$$P_{\$}(T) = X(0)P_{¥}(T)Y(T) \quad \Rightarrow \quad P_{¥}(T) = \frac{P_{\$}(T)}{Y(T)X(0)}. \tag{6.38}$$

Suppose, say, that we have blindly estimated discount curves $P(\cdot)$ and $P_{¥}(\cdot)$ from the market for USD- and JPY-denominated interest rate swaps, respectively, without paying any attention to FX markets. The discount curves $P_{\$}(\cdot)$ and $P_{¥}(\cdot)$ estimated in this fashion will very likely *not* satisfy (6.38), implying the existence of cross-currency arbitrages. The degree to which (6.38) is typically violated is often small, but any such violation can be highly problematic for a firm engaging in trading of significant amounts of both USD- and JPY-denominated assets.

6.5.2.2 Separation of Discount and Forward Rate Curves

It may appear that there is no way out of this conundrum: after all, our curve construction algorithms imply a unique discount curve out of given swap prices and have few, in any, means of incorporating additional requirements such as (6.38). However, built into our assumptions about how to price a swap (see (6.2)) was an implicit assumption that Libor itself is the proper discount rate for flows based on Libor fixings. As Libor rates represent lending rates between banks, they contain a certain amount of credit risk[18] and it is ex-ante unclear that they are suitable proxies for a "risk-free" rate (or, at least, are suitable for discounting of swap cash flows). More details about this can be found in Collin-Dufresne and Goldstein [2001] and Duffie and Huang [1996]. For our purposes, it suffices to introduce the notion that when computing a swap value we may need two curves: i) the Libor "pseudo-discount" curve $P^{(L)}(t) = P^{(L)}(0,t)$, used to project the Libor-based floating cash flows on the floating leg of the swap; and ii) a real discount curve $P(t) = P(0,t)$, used to discount all cash flows. For, say, a regular JPY-based fixed-floating swap paying a coupon c on a schedule $\{t_i\}_{i=1}^n$, the swap valuation equation thus becomes

$$V_{¥}(0) = \underbrace{\sum_{i=0}^{n-1} c\tau_i P_{¥}(0,t_{i+1})}_{\text{Fixed Leg}} - \underbrace{\sum_{i=0}^{n-1} L_{¥}(0,t_i,t_{i+1})\tau_i P_{¥}(0,t_{i+1})}_{\text{Floating Leg}}, \tag{6.39}$$

where $\tau_i = t_{i+1} - t_i$, $t_0 = 0$, and where we compute $L_{¥}(t,t_i,t_{+1})$ as (compare to (4.2))

[18] Reflecting the average bank credit rating, it is common to think of the Libor curve as a proxy for a AA-rated funding curve. In reality, however, this is not quite accurate, as banks with deteriorating credit are eliminated from the consortium of banks polled when determining the Libor rates. As such, the medium- and long-term forwards of the Libor curve contain *less* credit risk adjustment than similar forwards for a curve used to discount obligations to a single AA-rated firm. For more on this, see Collin-Dufresne and Goldstein [2001].

$$L_{¥}(t, t_i, t_{i+1}) = \frac{1}{\tau_i}\left(P_{¥}^{(L)}(t, t_i)/P_{¥}^{(L)}(t, t_{i+1}) - 1\right).$$

A similar construction can be done for any USD swap, by means of introducing curves $P_{\$}^{(L)}(t)$ and $P_{\$}(t)$. Technically speaking, the Libor forwards $L_{¥}(t, t_i, t_{i+1})$ in (6.39) represent expectations in the t_{i+1}-forward measure — i.e. the martingale measure associated with the numeraire price $P_{¥}(t, t_{i+1})$ (*not* $P_{¥}^{(L)}(t, t_{i+1})$) — of quoted *spot* Libor rates,

$$L_{¥}(t, t_i, t_{i+1}) \triangleq \mathrm{E}_t^{t_{i+1},¥}\left(L_{¥}(t_i, t_i, t_{i+1})\right). \tag{6.40}$$

In this view, the quoted Libor rate is effectively reduced to an observable index that may have little, if any, relationship to a true discount rate. For this reason, the time 0 pseudo-discount curves $P_{\$}^{(L)}(t)$ and $P_{¥}^{(L)}(t)$ are often referred to as *index* curves.

It should be clear that the introduction of the pseudo-discount curves $P_{\$}^{(L)}(t)$ and $P_{¥}^{(L)}(t)$ equips us with enough degrees of freedom to fit both USD-denominated swaps, JPY-denominated swaps, and the market for FX forward contracts. In fact, we have *too many* degrees of freedom: four curves, but only three separate markets to calibrate to. One way of handling this issue is to impose additional assumptions about the relationship between the curves $P(t)$ and $P^{(L)}(t)$ in one chosen currency. Before the 2007–2009 crisis, the following assumption was common.

Assumption 6.5.1. *In USD, the Libor pseudo-discount curve coincides with the real discount curve, i.e.* $P_{\$}^{(L)}(t, T) = P_{\$}(t, T)$ *for all* t *and* T, $T \geq t$.

Assumption 6.5.1 amounts to a convention where the liquidity and credit basis of non-USD Libor rates should all be measured relative to a neutral "bed-rock" established by USD Libor rates. Embedded into Assumption 6.5.1 also is the notion that most firms world-wide can fund themselves by borrowing in USD at levels close to USD-Libor; in the past this was often not a bad assumption. As we discuss in Section 6.5.3 below, post-crisis a non-trivial basis between index and discounting curves has emerged in the US. For simplicity of exposition we proceed in this section with Assumption 6.5.1, but the index-discounting basis in the US could be easily incorporated into the algorithm. The problem of accounting for this basis in *single-currency* curve construction in postponed until Section 6.5.3.

It is common to measure the difference between $P^{(L)}(t)$ and $P(t)$ in yield space, writing

$$P^{(L)}(t) = P(t)e^{-s(t)t},$$

where $s(t)$ is a yield spread often known as the *cross-currency (CRX) yield spread*. By Assumption 6.5.1, $s(t)$ is zero for USD, but will rarely be zero for any other currency. As indicated earlier, $s(t)$ will generally be quite small, often in the magnitude of a few basis points or less. Occasionally, however,

the CRX yield spread may blow out, particularly if banks in a particular country are perceived as having below-average credit quality. For instance, in the late 1990's, the CRX yield spread reached somewhere around -40 basis points in JPY as Japanese banks were perceived as being in economic trouble. During that period of time, foreign banks could generally fund themselves in USD at USD Libor, but in JPY at rates significantly below JPY Libor (due to their superior credit relative to Japanese banks). Had FX forward rates traded without any large CRX basis, foreign banks could have borrowed in JPY and used the FX forward markets to turn their obligations into USD-denominated ones at a borrowing cost below USD Libor, which would have indicated the existence of an inconsistency and an arbitrage. Conversely, in early 2008 the CRX basis spread became significantly positive (up to $+60$ basis points) as the hedging demands of long-dated FX books increased rapidly on the back of significant strengthening of the Yen versus the US Dollar. During the 2007–2009 crisis, many other currencies (including EUR) have experienced similar dramatic moves in the CRX basis spreads against USD.

6.5.2.3 Cross-Currency Basis Swaps

The discussion so far has assumed the existence of a liquid market for FX forwards, as means to observe and tie down the CRX basis between rates in two separate currencies. In reality, the interbank FX forward market is rarely liquid beyond maturities of one year, a far cry from the 30+ year horizons to which we often want to build yield curves. Rather than relying on FX forwards, instead we can turn to the market for *floating-floating cross-currency (CRX) basis swaps*. Briefly speaking, CRX basis swaps are contracts where floating Libor payments in one currency are exchanged for floating Libor payments in another currency, plus or minus a spread. The swaps involve an exchange of notionals at trade inception and at maturity; the ratio between the two notionals is normally set to equal the spot FX exchange rate prevailing at trade inception. CRX basis swaps are closely related to FX forward contracts — indeed a one-period CRX basis swap is identical to an FX forward contract.

As was the case with FX forward contracts, failing to fit to the market for CRX basis swaps can lead to arbitrageable inconsistencies. For instance, consider the pricing of a stream of fixed USD cash flows. One way to determine the JPY price of these cash flows would be through simple discounting by the USD discount curve, followed by a conversion to JPY at the spot exchange rate. Alternatively, the following zero-cost scheme could be implemented to turn the stream of USD cash flows into fixed JPY cash flows:

1. Swap the fixed cash flows to streams of USD Libor plus some spread x, in a regular USD interest rate swap.

2. Swap USD Libor + x against JPY Libor + e + x in a CRX basis swap, e being a market-quoted CRX basis swap spread.
3. Swap JPY Libor + e + x against a stream a fixed JPY coupons, in a regular JPY interest rate swap.

The USD value of the cash flows in Step 3 can be determined through discounting with the JPY discount curve, and subsequent conversion to USD at the spot USD/JPY exchange rate. If the JPY discount curve is inconsistent with the basis-swap market, the value computed this way may not equal the value computed by discounting the original USD cash flows at the USD discount curve. Since the swap transactions 1–3 above are costless, this discrepancy will indicate an arbitrage.

We can use the pricing formalism developed in Section 6.5.2.2 to provide an explicit expression for the value of a CRX basis swap. For concreteness, we again turn to the USD/JPY market and consider a CRX basis swap, where a USD-based corporation receives USD Libor flat in exchange for payments of JPY Libor plus a fixed spread, $e_{¥}$. With payment dates $\{t_i\}_{i=1}^n$, the USD price $V_{\text{basisswap},\$}$ of the basis swap is (assuming a \$1 notional)

$$V_{\text{basisswap},\$}(0) \tag{6.41}$$

$$\begin{aligned}
&= \sum_{i=0}^{n-1} L_{\$}(0,t_i,t_{i+1})\tau_i P_{\$}(0,t_{i+1}) + P_{\$}(0,t_n) \\
&\quad - X(0)\left(\sum_{i=0}^{n-1}(L_{¥}(0,t_i,t_{i+1}) + e_{¥})\tau_i P_{¥}(0,t_{i+1}) + P_{¥}(0,t_n)\right) \\
&= 1 - X(0) \\
&\quad \times \left(\sum_{i=0}^{n-1}\left(\frac{P_{¥}^{(L)}(0,t_i)}{P_{¥}^{(L)}(0,t_{i+1})} - 1 + e_{¥}\tau_i\right) P_{¥}(0,t_{i+1}) + P_{¥}(0,t_n)\right),
\end{aligned} \tag{6.42}$$

where we have used the fact that $P_{\$}$ and $P_{\$}^{(L)}$ are identical (by Assumption 6.5.1), in order to reduce the time 0 price of the USD floating leg to \$1. The market quotes par values $e_{¥}^{par}$ — that is, the value of $e_{¥}$ that will make $V_{\text{basisswap},\$}(0) = 0$ — in a wide range of maturities extending out to 30 years or more.

6.5.2.4 Modified Curve Construction Algorithm

By Assumption 6.5.1, construction of the USD discount and Libor curves can be accomplished through direct application of the routines in Sections 6.2 or 6.3 on benchmark securities consisting of deposits, FRAs, and swaps. For non-USD currencies, however, matters are more complicated as we must now simultaneously estimate both curves $P(t)$ and $P^{(L)}(t)$, $t > 0$,

in a manner ensuring that i) Libor benchmark securities are correctly priced; and ii) par-valued cross-currency swaps against USD are correctly priced. In performing this exercise, we apply (6.42) and adjust valuation expressions for the benchmark securities according to the principles of the swap-pricing[19] equation (6.39). We can make the curve construction problem quite complicated if we insist on $P(t)$ and $P^{(L)}(t)$ both being smooth functions of t; instead, here we will show a simpler idea that applies earlier algorithms in this chapter in iterative fashion.

Working as before with JPY as the foreign currency, we start by assuming that we have somehow managed to construct the correct Libor curve $P_{¥}^{(L)}(t)$. Were we — erroneously — to pretend that $P_{¥}^{(L)}(t)$ were a proper discount curve, we would get, for our N benchmark securities, a vector of values $\mathbf{V}^{(L)}$ that would generally *not* equal the correct JPY market prices $\mathbf{V}$:

$$\mathbf{c}\mathbf{P}_{¥}^{(L)} = \mathbf{V}^{(L)}, \quad \mathbf{V}^{(L)} \neq \mathbf{V}. \tag{6.43}$$

As the $P_{¥}^{(L)}(t)$ discount curve will be used to project forward rates, the yields and forward rates implied by $P_{¥}^{(L)}(t)$ should ideally be smooth. The smoothness requirements of the discount curve $P_{¥}(t)$, however, are significantly lower, as we shall never need to (in effect) differentiate this curve to produce forward Libor rates. Assuming that we have CRX basis swaps maturing on the benchmark set maturity dates $\{T_i\}_{i=1}^{N}$, it is thus, for instance, not unreasonable to write

$$P_{¥}(t) = P_{¥}(T_i)\frac{P_{¥}^{(L)}(t)e^{-\varepsilon_i \cdot (t-T_i)}}{P_{¥}^{(L)}(T_i)}, \quad t \in [T_i, T_{i+1}), \tag{6.44}$$

which assumes that the instantaneous forward rates generated by $P_{¥}(t)$ are given by those computed from $P_{¥}^{(L)}(t)$ plus a piecewise flat function:

$$f_{¥}(t) = f_{¥}^{(L)}(t) + \varepsilon(t), \quad \varepsilon(t) = \sum_{i=0}^{N-1} \varepsilon_i 1_{\{t \in [T_i, T_{i+1})\}}, \tag{6.45}$$

where $T_0 = 0$ as before.

In a cross-currency setting, the expression (6.4) no longer holds and must be replaced with something like

$$\mathbf{f}\left(\mathbf{P}_{¥}^{(L)}, \mathbf{P}_{¥}\right) = \mathbf{V} \tag{6.46}$$

for a non-linear vector-valued function $\mathbf{f}$. Indeed, according to (6.39), many of the coupon payments ($\mathbf{c}$) become a non-linear function of points on $\mathbf{P}_{¥}^{(L)}$ and cannot be considered constants. To salvage the methodologies discussed

[19] Pricing of short-term deposits only involves the discount curve P, whereas FRAs can be treated as one-period swaps. We leave details to the reader.

in Sections 6.2 or 6.3, we avoid working with (6.46) directly, and instead use an iteration based on equations (6.42), (6.43), and (6.45). The iteration attempts to estimate the unknown quantity $\mathbf{V}^{(L)}$ in (6.43) and works as follows:

1. Let $\mathbf{V}^{(L)}(j)$ be the j-th iteration for $\mathbf{V}^{(L)}$. Use $\mathbf{V}^{(L)}(j)$ along with (6.43) to estimate the curve $P_{¥}^{(L)}(t)$, using any of the curve construction methods discussed in earlier sections of this chapter.
2. Given knowledge of $P_{¥}^{(L)}(t)$, use (6.44)–(6.45) combined with (6.42) to imply the N constants $\varepsilon_0, \varepsilon_1, \ldots \varepsilon_{N-1}$, by calibration to the N par-valued CRX basis swaps maturing at time $T_1, \ldots, T_N$. This calibration exercise can be done by simple bootstrapping, and establishes the current guess for $P_{¥}(t)$.
3. Given estimates for both $P_{¥}^{(L)}(t)$ and $P_{¥}(t)$, compute guessed prices $\mathbf{V}(j)$ of all benchmark securities, i.e. evaluate the left-hand side of (6.46). If $\mathbf{V}(j)$ and $\mathbf{V}$ are within a given tolerance, we are done.
4. Update the guess for $\mathbf{V}^{(L)}$ according to $\mathbf{V}^{(L)}(j+1) = \mathbf{V}^{(L)}(j) - (\mathbf{V}(j) - \mathbf{V})$, and proceed to Step 1.

The iteration is initiated at $j = 0$ with the estimate $\mathbf{V}^{(L)}(0) = \mathbf{V}$ and runs until the termination criterion in Step 3 is satisfied. As the approximation $\mathbf{V}^{(L)} \approx \mathbf{V}$ is normally very accurate, only a few iterations are needed to reach acceptable precision.

In this book we shall mostly ignore the existence of a non-zero CRX basis spread. In construction of a model for the evolution of the Libor curve, the reader should, however, keep in mind that it may be necessary to adjust the curve slightly before using it to discount any cash flows. In a dynamic setting, it is quite common to perform this adjustment by simply assuming that $\varepsilon(t)$ in (6.45) is deterministic. A discussion of how to incorporate both stochastic and deterministic spreads in a dynamic model for interest rate evolution can be found in Section 15.5. For now, we note that using deterministic spreads is generally safe, unless pricing securities with strong convexity in $\varepsilon(t)$ — e.g. an option on a CRX basis swap — in which case a separate stochastic model for $\varepsilon(t)$ may be needed.

6.5.3 Tenor Basis and Multi-Index Curve Group Construction

Section 6.5.2 relied extensively on the notion of separating the discount curves used for Libor projection and for outright discounting. This idea is quite powerful and has applications in other settings, including some where only a single currency is involved. For instance, for swaps that pay a non-Libor index — e.g. the Bond Mark Association (BMA) index in the US — it is natural to introduce a basis spread that measures the difference between forward rates of the non-Libor index curve and the Libor curve itself.

More recently, a similar technique has become important even for curves used for pricing standard Libor-based contracts. We have already mentioned (Section 5.1) that the Fed funds rate, the overnight rate used for balances of bank deposits with the Federal reserve, is often considered the closest proxy to the risk-free rate in the US (with Eonia and Sonia rates, see Section 5.1, fulfilling the same function for Euro and GBP). One argument for this choice is that most inter-dealer transactions are collateralized under the *International Swaps and Derivatives Association* (ISDA) Master Agreement, with the rate paid on collateral being the Fed funds rate (for USD; Eonia and Sonia for Euro and GBP), see Piterbarg [2010] (and also the discussion in Section 5.1). While the spread between the Fed funds rate and 3 month Libor rate used to be very small — in the order of a few basis points — after September 2007 it went up to as much as 275 basis points, and it is now generally accepted now that the Libor rate is no longer a good proxy for a discounting rate on collateralized trades. Uncollateralized derivative contracts are subject to credit risk, and a fully consistent pricing approach needs to incorporate the cost of hedging this risk (the co-called *credit valuation adjustment* or *CVA*). These computations are outside the scope of this book and can get very complex, in part because collateral rules can be complicated and are normally enforced on entire counterparty portfolios and not just on individual trades. See Gregory [2009] for further details.

As we discussed in Section 6.5.2, if we make an assumption on the index-discounting basis in one currency (say, USD), we can translate it into the index-discounting basis in any other currency through the market quotes for forward FX contracts and cross-currency basis swaps. However, to estimate this basis in USD (say), we need to rely on domestic markets only; doing otherwise will introduce a circularity into our arguments. Fortunately, the market in the appropriate instruments, the OIS (overnight index swap, a swap of payments based on a compounded Fed funds rate versus fixed rate, see (5.7)–(5.8)) and the Fed funds/Libor basis swaps (see Section 5.7) — has developed in the US with a range of maturities actively traded. Hence, using techniques that we already discussed, we can construct a pair of curves — a curve for discounting and a curve for projecting 3 month (say) forward Libor rates — in a self-consistent way from the market quotes on deposits, FRAs, swaps with 3 month frequency, and overnight index swaps.

Currently there are no countries where both the OIS market and the cross-currency basis swap (vs. USD) market are very liquid, and we can always use one or the other to find the index-discounting basis. As the markets evolve, there may come a time when there will be two liquid sources of discounting curve information. It turns out that potential conflict between the two can be resolved by carefully analyzing the collateral mechanisms used in the two markets and the implications for yield curve construction. This discussion is outside of scope of the current edition of our book, but the interested reader could consult Fujii et al. [2010] for details.

The challenges of curve construction do not end with building separate discount and forecasting curves to take into account the index-discounting basis. We also need to account for the *tenor basis* between vanilla single-currency swaps trading at different frequencies, e.g. 1 month, 3 months, and 6 months. Before proceeding, let us explain in more detail what we mean by tenor basis.

Suppose we construct Libor and discount curves based on, say, vanilla swaps (and for non-USD currencies also CRX basis swaps) paying 3 month Libor on a quarterly schedule. If the resulting index and discount curves are subsequently used to price a vanilla swap paying 1 month Libor on a monthly schedule, the resulting price is typically different from actual market quotes. In other words, there is a basis between the 3 month and 1 month Libor index curves, a basis arising partly from credit considerations and partly liquidity considerations (banks have a natural desire to have longer-term deposits to better match their loan commitments). Historically this basis has also been low; for example, the difference between 1 month and 3 month Libor rates was in the order of one basis point up until September 2007, but since then has been as wide as 50 basis points.

When various basis levels were small, the small discrepancies between different Libor-tenor swaps were often accounted for by building a unique discount curve for the subset of swaps referencing the Libor rate of a particular tenor; this curve would, in addition to generating the floating leg forward rates, then be used to discount floating and fixed cash flows of swaps of that frequency. In a swap pricing framework, this can create an arbitrage since it implies that fixed flows (from the fixed leg) will be discounted at different discount curves, depending on which Libor tenor the fixed flows happen to be paid against. Moreover, it is not clear how to deal with swaps that involve multiple Libor tenors[20], or how to aggregate risks coming from unrelated, individually constructed curves. Again, when the differences were small, these issues were largely ignored.

More recently, the naive approach above has evolved into the idea of using a *multi-index curve group*, a collection consisting of a single discount curve and multiple index curves, one for each Libor tenor covered by the multi-index curve group. The index curves are used in a tenor-specific manner to project Libor forward rates, and the universal discount curve serves to discount all floating and fixed cash flows, irrespective of tenor. The index curves are built sequentially as spreads off previously-built curves, which provides linkage between index curves and also a convenient risk parameterization. This relatively recent idea is discussed in Traven [2008] in considerable details, from where most of the material of this section is derived. Another good reference here is Fujii et al. [2010].

[20] The most common example of this is a swap with a short front stub, i.e. a swap where at inception the first payment period is shorter than subsequent ones.

To discuss multi-index curve groups in detail, let us introduce a superscript k, $k = 1, \ldots, K$, to distinguish quantities related to different tenors. In particular, let t_i^k be the i-th date in the tenor structure for tenor k; $\tau_i^k = t_{i+1}^k - t_i^k$ the corresponding tenor offset; $L^k(t_i^k, t_i^k, t_{i+1}^k)$ the spot Libor rate of tenor k for the i-th period; and so on. If we denote the expected value of $L^k(t_i^k, t_i^k, t_{i+1}^k)$ under the t_{i+1}^k-forward measure by

$$L^k\left(t, t_i^k, t_{i+1}^k\right) \triangleq \mathrm{E}_t^{t_{i+1}^k}\left(L^k\left(t_i^k, t_i^k, t_{i+1}^k\right)\right)$$

(compare to (6.40)), then the value of a fixed-floating k-Libor-tenor swap with n periods and rate c is given by

$$V^k(0) = \underbrace{\sum_{i=0}^{n-1} c\tau_i^k P(t_{i+1}^k)}_{\text{Fixed Leg}} - \underbrace{\sum_{i=0}^{n-1} L^k(0, t_i^k, t_{i+1}^k)\tau_i^k P(t_{i+1}^k)}_{\text{Floating Leg}}. \tag{6.47}$$

Here $P(t)$ is the universal discounting curve. The time 0 index curve for tenor k, $P^k(t)$, is defined by the condition that forward Libor rates (of tenor k) be defined by the familiar formula

$$L\left(0, t_i^k, t_{i+1}^k\right) = \frac{1}{\tau_i^k}\left(P^k\left(t_i^k\right)/P^k\left(t_{i+1}^k\right) - 1\right). \tag{6.48}$$

A multi-index curve group is defined as a collection $\{P(\cdot), P^1(\cdot), \ldots, P^K(\cdot)\}$ of the universal discounting curve and one index curve per tenor, with swaps priced via (6.47) (and equivalent formulas for other linear instruments) and (6.48).

Let us outline how to calibrate a multi-index curve group to market instruments referencing rates of different tenors. For each market, fixed-floating swaps referencing a Libor rate of one particular tenor L^k are usually the most liquid. Assume that this curve is the first index curve in the group, i.e. $k = 1$. The method from Section 6.5.2 can be used to construct a discounting curve, and a *base* index curve $P^1(\cdot)$ from

- Funding instruments such as deposits, forward FX contracts, OIS, overnight rate basis swaps (i.e. floating-floating swaps of an overnight rate versus L^1, see Section 5.7 and a discussion of more general floating-floating single-currency basis swaps in the next paragraph), cross-currency basis swaps, and the like.
- Vanilla instruments referencing L^1 such as FRAs on L^1 and fixed-floating swaps on L^1 versus a fixed rate.

To construct $P^2(\cdot)$, we assume that prices of *floating-floating single-currency basis swaps* are available in the market. A floating-floating basis swap is a swap of payments linked to a Libor rate of a particular tenor — such as L^1 — versus payments based on a Libor rate of different tenor —

such as L^2. Each leg pays on its own schedule of a corresponding tenor. A fixed spread is typically added to one of the legs to make the swap value at inception equal to zero. If a floating-floating basis swap is not traded or not liquid, it can be synthesized from two fixed-floating swaps referencing L^1 and L^2.

If a floating-floating basis swap is traded at par in the market, the values of both legs should be the same at time 0:

$$\sum_{i=0}^{n^2(T)-1} L^2(0, t_i^2, t_{i+1}^2)\tau_i^2 P(t_{i+1}^2) = \sum_{i=0}^{n^1(T)-1} \left(L^1(0, t_i^1, t_{i+1}^1) + e^{1,2}(T)\right)\tau_i^1 P(t_{i+1}^1). \quad (6.49)$$

Here $n^k(T)$ is the number of periods in the tenor structure to date T for tenor k, and $e^{1,2}(T)$ is the quoted floating-floating basis spread for exchanging L^1 for L^2 to maturity T, quoted on the L^1 leg. It could be positive or negative, depending on perceived desirability of payments linked to L^1 versus L^2.

Similar to (6.44)–(6.45), we represent $P^2(\cdot)$ as a multiplicative spread to $P^1(\cdot)$:

$$P^2(t) = P^1(t)e^{-\int_0^t \eta^{1,2}(s)\,ds}, \quad t \geq 0$$

for a given, usually piecewise-constant, spread function $\eta^{1,2}(\cdot)$. With the discounting curve $P(\cdot)$ already constructed and $L^1(0, t_i^1, t_{i+1}^1)$ known for all i from the already-built index curve $P^1(\cdot)$, it is a simple exercise to obtain the spread function $\eta^{1,2}(\cdot)$ by solving (6.49) for different T's.

Having built $P^2(\cdot)$, the remaining index curves $P^k(\cdot)$, $3 \leq k \leq K$ may be constructed in a similar fashion, always using floating-floating basis swap spreads for L^k versus L^1 basis swaps or, more generally, for whatever L^k versus L^i basis swaps are the most liquid with $i < k$. In particular, each index curve $P^k(\cdot)$ for $k > 1$ is built as a *spread*, or *basis*, curve to one of the previous curves.

In the presence of multiple curves, it is not entirely clear from the outset how to most sensibly define risk sensitivities. Fortunately, with this spread-based method of curve group construction, sensitivities to instruments used in the curve group have clear, and orthogonal, meaning:

- Perturbations to instruments used in building the base index curve, e.g. non-basis swaps and FRAs referencing L^1, define risk sensitivities to the overall levels of interest rates. Clearly, with basis spreads for L^1-versus-L^k floating-floating basis swaps, $k = 2, \ldots, K$, kept constant, shifts to fixed-floating L^1-versus-fixed swap rates will move all index curves together by the same amount in forward rate space. These sensitivities are the direct analog of the standard interest rate deltas in the traditional, single-curve, world, see Section 6.4.

- Perturbations to funding instruments define sensitivities to discounting.
- Perturbations to basis swap spreads for L^k-versus-L^1 floating-floating basis swaps define *basis risk*, i.e. the risk that index curves of different tenors do not move in lock step.

The parameterization allows us to naturally aggregate "similar" risks such as overall rate level risks, discounting risks, basis risks, while keeping different kinds separate for efficient risk management. Had we constructed all index curves in separation from each other (from multiple sets of vanilla fixed-floating swaps, say) such automatic aggregation would not be possible.

6.A Appendix: Spline Theory

6.A.1 Hermite Spline Theory

Consider a given set of data points (x_i, f_i), $i = 1, \ldots, N$, where $x_1 < x_2 < \ldots < x_N$. We wish to apply an interpolation rule such that a continuous function $f(x)$, $x \in [x_1, x_N]$, is created. We require that f be piecewise cubic, be at least once differentiable (C^1), and be a true interpolating function, i.e. $f(x_i) = f_i$ for all $i = 1, \ldots, N$.

In the Hermite spline description, tangents at points x_i, $i = 1, \ldots, N$, are assumed exogenously specified. Let f_i' denote the tangent df/dx at $x = x_i$, $i = 1, \ldots, N$. We write f as a piecewise cubic polynomial

$$f(x) = a_{3,i}(x - x_i)^3 + a_{2,i}(x - x_i)^2 + a_{1,i}(x - x_i) + a_{0,i}, \quad x \in [x_i, x_{i+1}],$$

with unknown coefficients $a_{j,i}$ specific to each interval $[x_i, x_{i+1}]$. Expressing that both f and f' should be continuous across each point x_i allows us, after a little rearrangement, to write the spline specification as simply

$$f(x) = \mathbf{D}_i(x)^\top \mathbf{M} \begin{pmatrix} f_i \\ f_{i+1} \\ f_i' h_i \\ f_{i+1}' h_i \end{pmatrix}, \quad x \in [x_i, x_{i+1}], \tag{6.50}$$

where $h_i \triangleq x_{i+1} - x_i$,

$$\mathbf{D}_i(x) = \begin{pmatrix} \delta_i^3 \\ \delta_i^2 \\ \delta_i \\ 1 \end{pmatrix}, \quad \delta_i \triangleq \frac{x - x_i}{h_i},$$

and $\mathbf{M}$ is the *Hermite matrix*

$$\mathbf{M} = \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

One drawback of the Hermite specification is the need to directly specify tangents df/dx. A number of approaches exist to compute these directly from the given data points or by adding additional control points. For our purposes, we highlight the so-called *Catmull-Rom spline* (Catmull and Rom [1974]), where the derivatives are computed as[21]

$$f'_i = \frac{f_{i+1} - f_{i-1}}{x_{i+1} - x_{i-1}}, \quad i = 2, \ldots, N-1. \tag{6.51}$$

At end points (x_1, f_1) and (x_N, f_N) forward and backward differences are used instead:

$$f'_1 = \frac{f_2 - f_1}{x_2 - x_1}; \quad f'_N = \frac{f_N - f_{N-1}}{x_N - x_{N-1}}. \tag{6.52}$$

Notice that with (6.51), the Hermite representation (6.50) can be rewritten in the derivative-free form

$$f(x) = \mathbf{D}_i(x)^\top \mathbf{A}_i \begin{pmatrix} f_{i-1} \\ f_i \\ f_{i+1} \\ f_{i+2} \end{pmatrix}, \quad x \in [x_i, x_{i+1}], \tag{6.53}$$

where

$$\mathbf{A}_i = \begin{pmatrix} -\alpha_i & 2-\beta_i & -2+\alpha_i & \beta_i \\ 2\alpha_i & \beta_i - 3 & 3-2\alpha_i & -\beta_i \\ -\alpha_i & 0 & \alpha_i & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad i = 2, \ldots, N-2, \tag{6.54}$$

with

$$\alpha_i = \frac{h_i}{h_i + h_{i-1}}, \quad \beta_i = \frac{h_i}{h_{i+1} + h_i}.$$

As indicated, equation (6.53) only holds for $i = 2, \ldots, N-2$, with external boundary conditions needed to establish the curve in the segments $[x_1, x_2]$ and $[x_{N-1}, x_N]$. Applying boundary conditions of the type (6.52), we get

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 1-\beta_1 & -1 & \beta_1 \\ 0 & -1+\beta_1 & 1 & -\beta_1 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \tag{6.55}$$

and

[21] Variations exist which use more elaborate finite difference style derivatives, taking into account the fact that the grid may be non-equidistant; see Chapter 2. Given the semi-heuristic nature of the Catmull-Rom spline, it is doubtful that much is gained by such extensions.

$$\mathbf{A}_{N-1} = \begin{pmatrix} -\alpha_{N-1} & 1 & -1+\alpha_{N-1} & 0 \\ 2\alpha_{N-1} & -2 & 2-2\alpha_{N-1} & 0 \\ -\alpha_{N-1} & 0 & \alpha_{N-1} & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{6.56}$$

While Catmull-Rom splines shall suffice for the yield curve applications we have in mind here, let us note that it is possible to introduce further parameters to control local spline behavior. For completeness, let us quickly list one popular extension. First, we allow for the possibility that the curve is locally only C^0 by having incoming and outgoing tangents be different. That is, we define

$$f'_{i,I} = \lim_{\varepsilon \downarrow 0} \frac{f(x_i) - f(x_i - \varepsilon)}{\varepsilon}; \quad f'_{i,O} = \lim_{\varepsilon \downarrow 0} \frac{f(x_i + \varepsilon) - f(x_i)}{\varepsilon},$$

and rewrite the Hermite equation (6.50) as

$$f(x) = \mathbf{D}_i(x)^\top \mathbf{M} \begin{pmatrix} f_i \\ f_{i+1} \\ f'_{i,O} h_i \\ f'_{i+1,I} h_i \end{pmatrix}, \quad x \in [x_i, x_{i+1}]. \tag{6.57}$$

Only when $f'_{i,I} = f'_{i,O}$ for all i is the curve C^1 everywhere. The *Kochanek-Bartels spline* — also known as the *TCB spline* — is defined through the expressions

$$\begin{aligned} f'_{i,I} = {} & \frac{(1-\sigma_i)(1+c_i)(1-b_i)}{2} \frac{f_{i+1} - f_i}{x_{i+1} - x_i} \\ & + \frac{(1-\sigma_i)(1-c_i)(1+b_i)}{2} \frac{f_i - f_{i-1}}{x_i - x_{i-1}}, \end{aligned} \tag{6.58}$$

$$\begin{aligned} f'_{i,O} = {} & \frac{(1-\sigma_i)(1-c_i)(1-b_i)}{2} \frac{f_{i+1} - f_i}{x_{i+1} - x_i} \\ & + \frac{(1-\sigma_i)(1+c_i)(1+b_i)}{2} \frac{f_i - f_{i-1}}{x_i - x_{i-1}}, \end{aligned} \tag{6.59}$$

for parameters $\sigma_i, c_i, b_i \in [-1, 1]$, $i = 1, \ldots, N$. The parameters σ_i, c_i, and b_i are used to control curve *tension*, *continuity*, and *bias*, respectively; clearly, when $\sigma_i = c_i = b_i = 0$, the Kochanek-Bartels spline reduces to the Catmull-Rom spline. A positive value of σ_i will tend to "tighten" the curve around the point (x_i, f_i), and a negative value will generate "slack". The parameters b_i are measures of over- and undershoot. To see this, set $\sigma_i = c_i = 0$ and note that when $b_i = 0$, left and right one-sided tangents are weighted equally producing a regular Catmull-Rom spline; when b_i is close to -1 (1), however, the outgoing (incoming) tangent dominates the path of the curve through the point (x_i, f_i) producing undershoot (overshoot). The parameters c_i control the degree of differentiability of the resulting spline: if a parameter $c_i \neq 0$,

the resulting spline will develop a corner (the direction of which depends on the sign of c_i) at point (x_i, f_i), losing its differentiability. Kochanek-Bartels splines are used extensively in computer graphics applications.

6.A.2 C^2 Cubic Splines

The cubic splines in Section 6.A.1 are generally not twice differentiable, and their second derivatives will jump across each knot. We wish to remedy this, and now consider a twice differentiable C^2 cubic spline $f(x)$ interpolating a set of data points (x_i, f_i), $i = 1, \ldots, N$. By necessity, such a cubic spline interpolant is piecewise linear in its second derivative

$$f''(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} f_i'' + \frac{x - x_i}{x_{i+1} - x_i} f_{i+1}'', \quad x \in [x_i, x_{i+1}], \tag{6.60}$$

where we use primes to denote differentiation and where $f_i'' \triangleq f''(x_i)$. We emphasize that for a C^2 cubic spline, the second derivative is continuous across knot points: $\lim_{x \downarrow x_i} f''(x) = \lim_{x \uparrow x_i} f''(x) = f''(x_i)$. Integrating (6.60) twice and requiring the curve to pass through data points results in the classical spline equation

$$f(x) = \frac{(x_{i+1} - x)^3}{6h_i} f_i'' + \frac{(x - x_i)^3}{6h_i} f_{i+1}'' + (x_{i+1} - x)\left(\frac{f_i}{h_i} - \frac{h_i}{6} f_i''\right) + (x - x_i)\left(\frac{f_{i+1}}{h_i} - \frac{h_i}{6} f_{i+1}''\right), \quad x \in [x_i, x_{i+1}], \tag{6.61}$$

where $h_i \triangleq x_{i+1} - x_i$. The second derivatives f_i'', $i = 1, \ldots, N$, needed to evaluate (6.61) can be obtained by requiring $f'(x)$ to be continuous across data points. The result is

$$\frac{h_{i-1}}{6} f_{i-1}'' + \frac{h_{i-1} + h_i}{3} f_i'' + \frac{h_i}{6} f_{i+1}'' = \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}}, \quad i = 2, \ldots, N-1. \tag{6.62}$$

The set of equations (6.62) is a tri-diagonal system for f_i'' that can be solved in $O(N-2)$ operations once we have specified boundary conditions[22] for f_1'' and f_N''. A classical boundary condition is $f_1'' = f_N'' = 0$, leading to the so-called *natural cubic spline*.

While C^2 cubic splines have a number of useful features, they have, loosely speaking, a built-in aversion to make tight turns (since this will cause large values of f''). This, in turn, will often produce extraneous inflection points and non-local behavior, in the sense that perturbation of a single f_i will significantly affect the appearance of the curve for x-values far from x_i. Also, monotonicity and convexity properties of the original data-set will typically not be preserved.

[22] Such boundary conditions may be indirect and can, among other choices, take the form of specification of a gradient f' at x_1 or x_N. By differentiation of (6.61), a gradient specification can always be turned into a condition on f'' at x_1 or x_N.

6.A.3 C^2 Exponential Tension Splines

An attractive remedy to the shortcomings of the cubic spline is to insert some *tension* in the cubic spline, that is, to apply a tensile force to the end-points of the spline. Formally, this can be accomplished (see Schweikert [1966]) by replacing the equation (6.60) with, $x \in [x_i, x_{i+1}]$,

$$f''(x) - \sigma^2 f(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}\left(f_i'' - \sigma^2 f_i\right) + \frac{x - x_i}{x_{i+1} - x_i}\left(f_{i+1}'' - \sigma^2 f_{i+1}\right), \tag{6.63}$$

where $\sigma > 0$ is a measure of the tension applied to the cubic spline[23]. Notice that we have replaced the assumption of a piecewise linear second derivative with the assumption that the quantity $f''(x) - \sigma^2 f(x)$ is linear on each sub-interval $[x_i, x_{i+1}]$.

Integrating (6.63) twice and requiring that the curve pass through the given data points, one obtains (after some rearrangements)

$$\begin{aligned} f(x) = &\left(\frac{\sinh\left(\sigma\left(x_{i+1} - x\right)\right)}{\sinh\left(\sigma h_i\right)} - \frac{x_{i+1} - x}{h_i}\right)\frac{f_i''}{\sigma^2} \\ &+ \left(\frac{\sinh\left(\sigma\left(x - x_i\right)\right)}{\sinh\left(\sigma h_i\right)} - \frac{x - x_i}{h_i}\right)\frac{f_{i+1}''}{\sigma^2} \\ &+ f_i\frac{x_{i+1} - x}{h_i} + f_{i+1}\frac{x - x_i}{h_i}, \quad x \in [x_i, x_{i+1}], \end{aligned} \tag{6.64}$$

where $h_i = x_{i+1} - x_i$ as before. Requiring continuity of the first derivative we then get for the f_i'' ,

$$\begin{aligned} &\left(\frac{1}{h_{i-1}} - \frac{\sigma}{\sinh\left(\sigma h_{i-1}\right)}\right)\frac{f_{i-1}''}{\sigma^2} \\ &\quad + \left(\frac{\sigma\cosh\left(\sigma h_{i-1}\right)}{\sinh\left(\sigma h_{i-1}\right)} - \frac{1}{h_{i-1}} + \frac{\sigma\cosh\left(\sigma h_i\right)}{\sinh\left(\sigma h_i\right)} - \frac{1}{h_i}\right)\frac{f_i''}{\sigma^2} \\ &\quad\quad + \left(\frac{1}{h_i} - \frac{\sigma}{\sinh\left(\sigma h_i\right)}\right)\frac{f_{i+1}''}{\sigma^2} = \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}}. \end{aligned}$$

Again, this is a tri-diagonal system of equations that can be solved in $O(N-2)$ operations once we specify f_1'' and f_N''.

From the representation (6.64), it is clear that on all intervals $[x_i, x_{i+1}]$ hyperbolic tension splines can be written as linear combinations of the basis functions 1, x, $e^{-\sigma x}$, $e^{\sigma x}$. The representation (6.64), however, has better behavior for large and small values of σ (see Renka [1987] and Rentrop [1980] for details about proper evaluation of the hyperbolic functions in (6.64) for large and small values of σ).

[23] Extension to non-uniform tension parameter is straightforward and involves replacing σ with σ_i in (6.63), with σ_i then being a measure of the tension applied locally to the curve in the interval $[x_i, x_{i+1}]$.

We notice that when the tension parameter $\sigma = 0$, equations (6.63) and (6.60) are identical, i.e. the tension spline degenerates into a regular cubic spline. On the other hand, when $\sigma \gg 1$ (6.63) reduces to piecewise linear interpolation, as

$$\lim_{\sigma \to \infty} f(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} f_i + \frac{x - x_i}{x_{i+1} - x_i} f_{i+1}, \quad x \in [x_i, x_{i+1}]. \tag{6.65}$$

Evidently, the equation (6.63) defines a twice differentiable curve that is a hybrid between a cubic spline and a piecewise linear spline.

The convergence of the tension spline towards a piecewise linear curve as $\sigma \to \infty$ can be shown to be *uniform*, i.e. (6.65) holds uniformly in $[x_i, x_{i+1}]$ for $i = 1, \ldots, N-1$. Similarly

$$\lim_{\sigma \to \infty} f'(x) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} \quad \text{and} \quad \lim_{\sigma \to \infty} f''(x) = 0$$

uniformly in any closed subinterval of $[x_i, x_{i+1}]$. See Pruess [1976] for details and a proof. The uniform convergence is important as it guarantees that the monotonicity and convexity properties of the underlying discrete data set are preserved, simply by choosing a sufficiently high value of the tension factor. Due to this property, hyperbolic tension splines are said to be *shape-preserving*[24]. As the tension factor increases, the resulting spline will also behave in increasingly local fashion towards input perturbations. In the limit $\sigma \to \infty$ each point $f(x)$ on the spline will only be associated with the two nearest-neighbor knots.

[24] Generalizing, suppose we introduce constraints on function values, first derivatives, or second derivatives. As long as these constraints are satisfied by piecewise linear interpolation, there will exist some value of the tension parameter σ (possibly $\sigma = 0$) which will make the tension spline satisfy the constraints. This observation is key to algorithms for automatic selection of σ from externally specified function constraints. See, for instance, Lynch [1982] and Renka [1987] for details and efficient algorithms for automatic tension selection.

7

Vanilla Models with Local Volatility

We have shown in Section 5.10 that European swaptions (and caplets/floorlets which we equate with one-period swaptions) can be valued as European options on forward swap rates. As a consequence, a full term structure model that specifies the dynamics of the whole yield curve through time is essentially unnecessary for European swaption valuation. Instead, we only need a model for the evolution — in fact, just a terminal distribution — of a single swap rate in isolation. Models of this type shall be denoted *vanilla models*, to distinguish them from full term structure models. Vanilla models can be extended by copula methods to describe joint terminal distributions of more than one rate, as we discuss in Chapter 17. Ultimately, however, their primary purpose in this book is to serve as a foundation for development of more widely applicable full term structure models, that is, models which provide consistent dynamics for *all* points on the yield curve simultaneously. Term structure models are extensively covered later in the book.

In this chapter, we review one-factor diffusive models where our ability to alter the terminal distribution stems from a single source: a swap rate dependent diffusion function. Models of this type are often known as *deterministic volatility function* (DVF), or sometimes *local volatility function* (LVF), models. We first discuss the most common tractable specifications of such models — the CEV, displaced diffusion, and quadratic models — and then move on to efficient numerical or expansion-based methods for European option pricing within the general DVF model class. The listed techniques and results will be frequently referenced in later chapters, especially in the context of model calibration.

7.1 General Framework

7.1.1 Model Dynamics

Let $S(t)$ denote a forward Libor or swap rate, and let $W(t)$ be a one-dimensional Brownian motion under a measure P in which $S(\cdot)$ is a martingale. We assume that $S(t)$ follows the one-dimensional SDE

$$dS(t) = \lambda\varphi\left(S(t)\right) dW(t), \tag{7.1}$$

where λ is a positive constant[1] and $\varphi : \mathbb{R} \to \mathbb{R}$ satisfies regularity conditions, such as those in Theorem 1.6.1. In most applications we would ideally want $S(t)$ to be non-negative, which is easily seen to impose the restriction

$$\varphi(0) = 0. \tag{7.2}$$

In some cases we may consciously decide to violate (7.2) for the sake of model tractability.

When dealing with vanilla models, we primarily work in the measure P, so we typically abbreviate E^{P} to E when there is no possibility of confusion.

7.1.2 Volatility Smile and Implied Density

The role of the function φ is to match the distribution of S to that observed through puts and calls traded in the market. Specifically, let $c(t, S; T, K)$ denote the (non-deflated) time t value of a T-maturity European call option struck at K with $S(t) = S$, i.e.

$$c(t, S(t); T, K) = \mathrm{E}_t\left(\left(S(T) - K\right)^+\right). \tag{7.3}$$

The time t probability density of $S(T)$ can be derived from time t observed values of c, as (proceeding heuristically)

$$\mathrm{P}_t\left(S(T) \in dK\right)/dK = \mathrm{E}_t\left(\delta\left(S(T) - K\right)\right) \tag{7.4}$$

$$= \mathrm{E}_t\left(\frac{\partial^2 c(T, S(T); T, K)}{\partial K^2}\right) = \frac{\partial^2 c\left(t, S(t); T, K\right)}{\partial K^2}, \tag{7.5}$$

where δ is the Dirac delta function. This classical result is due to Breeden and Litzenberger [1978] and allows us to construct the marginal density of $S(T)$ from prices of T-maturity call options for a continuum of strikes K.

In option markets, it is common to express the strike dependency of call (and put) options in terms of the so-called *implied volatilities*. Specifically,

[1] We allow for time dependence later in the chapter, starting in Section 7.6.

for a given option price c at strike K and maturity T, we define the time t implied volatility function $\sigma_{\mathrm{B}}(t, S; T, K)$ as the solution to

$$c(t, S; T, K) = S\Phi(d_+) - K\Phi(d_-), \tag{7.6}$$
$$d_\pm = \frac{\ln(S/K) \pm \frac{1}{2}\sigma_{\mathrm{B}}(t, S; T, K)^2(T-t)}{\sigma_{\mathrm{B}}(t, S; T, K)\sqrt{T-t}}.$$

We recognize the right-hand side of (7.6) as the Black-Scholes-Merton formula for a martingale process, i.e. the Black model (see Remark 1.9.4), with constant volatility $\sigma_{\mathrm{B}}(t, S; T, K)$. The mapping $K \mapsto \sigma_{\mathrm{B}}(t, S; T, K)$ is known as the T-maturity *volatility smile*[2]. In most established fixed income markets, the volatility smile is predominantly downward-sloping[3] in K, although it is not uncommon for σ_{B} to eventually increase in K for sufficiently large values of K.

7.1.3 Choice of φ

Had we allowed φ to depend on time, results by Dupire [1994] and Andersen and Brotherton-Ratcliffe [1998] demonstrate that any arbitrage-free marginal distribution of $S(T)$ can be realized by suitable choice of $\varphi = \varphi(t, S)$, $t \in [0, T]$. Indeed, non-parametric expressions exist to uniquely imply $\varphi(t, K)$ from observations of $\sigma_{\mathrm{B}}(0, S(0); t, K)$ for the double continuum $(t, K) \in [0, T] \times [0, \infty)$. Unless the resulting φ happens to be monotonically increasing or decreasing in S, however, the resulting model will imply non-stationary volatility smile behavior, which is contrary to typical behavior of actual markets. To expand on this issue, consider setting

$$\varphi(S) = a + (S - S(0))^2, \tag{7.7}$$

where $a > 0$. The function $\varphi(S)$ is thus a U-shaped function with a minimum value of a at $S = S(0)$. Using formulas from Section 7.3 below, it can be verified (and is intuitively obvious) that the time 0 volatility smile σ_{B} produced by this parameterization is also U-shaped. Moving forward to time $t > 0$, consider the smile generated at t by (7.7) if $S(t) \gg S(0)$. At a large level of $S(t)$, $\varphi(S)$ will appear to be a strongly increasing function of S, causing (7.7) to produce a volatility smile no longer U-shaped, but instead monotonically increasing at all statistically relevant strikes. Conversely, if $S(t)$ diffuses below $S(0)$ such that $S(t) \ll S(0)$, a monotonically *de*creasing smile will arise at time t.

---

[2] In case the smile is monotonically downward or upward sloping, i.e. not U-shaped, it is often called a *volatility skew*. *Skew* then refers to the slope of the smile.

[3] This is not necessarily true for emerging markets where the volatility smile, when observed, can be significantly upward sloping or convex.

Strong level-dependence of the basic volatility smile shape is often at odds with observable market behavior, and non-monotonic specifications of $\varphi(S)$ — such as (7.7) — should consequently be approached with some care. As a consequence, the basic model (7.1) is most appropriate for markets where the volatility smile is (close to) a monotonic function of K. A classical monotonic choice for φ is the *constant elasticity of variance* (CEV) specification

$$\varphi(S) = S^p, \tag{7.8}$$

for some constant p. As we proceed to show, this specification is analytically tractable.

7.2 CEV Model

7.2.1 Basic Properties

In this section, we examine the CEV specification (7.8) in detail. We start out with the following proposition:

Proposition 7.2.1. *Consider the stochastic differential equation*

$$dS(t) = \lambda S(t)^p \, dW(t), \tag{7.9}$$

where $p > 0$ is constant and $W(t)$ is a one-dimensional Brownian motion. The following holds:

1. *All solutions to (7.9) are non-explosive.*
2. *For $p \geq 1/2$, the SDE (7.9) has a unique solution.*
3. *For $0 < p < 1$, $S = 0$ is an attainable boundary for (7.9); for $p \geq 1$, $S = 0$ is an unattainable boundary for (7.9).*
4. *For $0 < p \leq 1$, $S(t)$ in (7.9) is a martingale; for $p > 1$, $S(t)$ is a strict supermartingale.*

Proof. Property 1 follows from a remark on page 332 and equation (5.5.19) in Karatzas and Shreve [1997], and Property 2 follows from Example 5.2.14 in Karatzas and Shreve [1997]. Property 3 can be proven using the classical Feller boundary classification techniques based on speed/scale measure integral, see Section 5.5 of Karatzas and Shreve [1997]; Andersen and Andreasen [2000b] have the details. Property 4 is proven in Sin [1998]. □More details on boundary characterization for CEV processes can be found in Davydov and Linetsky [2001].

Remark 7.2.2. For $p \geq 1/2$, the solution to (7.9) is unique. Hence, if the solution ever reaches the origin ($S = 0$), it stays there, i.e. is *absorbed.* For $0 < p < 1/2$, however, there are solutions that stay at origin if they reach it, and there are solutions that jump out if it. Hence, to define a unique

solution, a boundary condition at $S = 0$ must be specified for (7.9). In practice, we set $S = 0$ to be an *absorbing* barrier: if $S(t)$ hits 0 for the first time at $t = \tau$, $S(u) = 0$ for all $u \geq \tau$. This condition is not only imposed to be consistent with the case of $p \geq 1/2$, but is also the only boundary condition consistent with the absence of arbitrage.

Remark 7.2.3. While it is common to require the parameter p to be positive, the process is well-defined for negative p, $p < 0$, as well, with the same absorbing boundary condition at $S = 0$ as for the case $0 < p < 1/2$ above. This enlargement of the domain of applicability of the process is occasionally useful in the fixed income markets, although much less so than in equity or FX markets where the smiles can generally be much more downward sloping.

For $p < 1$ and $t > 0$, the time 0 probability that $S(t) = 0$ is non-zero. In fact, it can be shown (see, for example, Cox [1996]) that if τ, the first time $S(\cdot)$ hits 0, is greater than t, then

$$\mathrm{P}_t\left(\tau < T | \tau > t\right) = G\left(|\vartheta|, \frac{X(t)}{2\lambda^2(T-t)}\right), \quad T > t,$$

where

$$\vartheta = \frac{1}{2(p-1)}, \tag{7.10}$$

$$X(t) = \frac{S(t)^{2(1-p)}}{(1-p)^2}, \tag{7.11}$$

and G is the *complementary Gamma function*

$$G(a,x) \triangleq \frac{\Gamma(a,x)}{\Gamma(a)},$$

with the incomplete Gamma function $\Gamma(a,x)$ given by

$$\Gamma(a,x) = \int_x^\infty u^{a-1}e^{-u}\,du, \quad \Gamma(a) = \Gamma(a,0). \tag{7.12}$$

If the absorption probability is substantial, one may want to consider regularizing the process to prevent absorption; see Section 7.2.3 for this.

Due to the result in Proposition 7.2.1, Property 4, we normally prefer to avoid using $p > 1$. As $p > 1$ will produce volatility smiles increasing in K (and thereby different from those in fixed income markets), this restriction on p is often of little practical concern.

The transition density of $S(\cdot)$ in (7.9) is known in closed form and is listed below for reference, along with a short proof that highlights the relationship between CEV processes and squared Bessel processes.

Lemma 7.2.4. *Consider the SDE (7.9) for any $p \neq 1$ (including $p < 0$ and $p > 1$), and let ϑ and $X(t)$ be as in (7.10)–(7.11). Let $q(X(T)|X(t))$ be the conditional* P*-density of $X(T)$ given $X(t) > 0$, $t < T$. If the level $S = 0$ is defined to be an absorbing boundary for (7.9) when $p \leq 1/2$, then*

$$q(X(T)|X(t)) = \frac{1}{2\lambda^2(T-t)} \exp\left(-\frac{X(T)+X(t)}{2\lambda^2(T-t)}\right) \times \left(\frac{X(t)}{X(T)}\right)^{-\vartheta/2} I_{|\vartheta|}\left(\frac{\sqrt{X(T)X(t)}}{\lambda^2(T-t)}\right),$$

where $I_a(x)$ is the modified Bessel function of the first kind of order a*:*

$$I_a(x) = \sum_{j=0}^{\infty} \frac{(x/2)^{a+2j}}{j!\Gamma(a+j+1)}.$$

Proof. According to Ito's lemma, the process $X(t)$ satisfies the SDE

$$dX(t) = \lambda^2 \frac{1-2p}{1-p}\, dt + 2\lambda\sqrt{X(t)}\, dW(t).$$

Define the process $Y(v)$ by $Y(v) = X(v/\lambda^2)$. Applying a time change, it follows that

$$dY(v) = \frac{1-2p}{1-p}\, dv + 2\sqrt{Y(v)}\, d\widetilde{W}(v),$$

where $\widetilde{W}(\cdot)$ is a Brownian motion, up to the absorption time $\inf\{v > 0 : Y(v) = 0\}$. The process for Y can be identified as a so-called *squared Bessel process of index ϑ*. Standard results for this process (see e.g. p. 117 of Borodin and Salminen [1996]) give the result in the lemma. □

Remark 7.2.5. By the usual transformation rules for densities, the density for $S(T)$ conditional on $S(t)$ is

$$q(X(T)|X(t)) \cdot 2S(T)^{2(1-p)-1}/|1-p|.$$

7.2.2 Call Option Pricing

Consider now the valuation of European call options in the CEV model, requiring evaluation of the expectation

$$c_{\mathrm{CEV}}(t, S(t); T, K) \triangleq \mathrm{E}_t\left((S(T)-K)^+\right)$$

for $S(\cdot)$ that follows (7.9). Using the definition (7.11), we can rewrite this as

$$c_{\mathrm{CEV}}(t, S(t); T, K) = \mathrm{E}_t\left(\left(\left[(1-p)^2 X(T)\right]^{-\vartheta} - K\right)^+\right) = \int_0^{\infty} \left(\left[(1-p)^2 X\right]^{-\vartheta} - K\right)^+ q(x|X(t))\, dx,$$

where we have assumed $p \neq 1$ and the density $q(x|X(t))$ is given in Lemma 7.2.4. A straightforward, but tedious, integration exercise (see e.g. Schroder [1989] or Andersen and Andreasen [2000b]) yields the following result:

Proposition 7.2.6. *Consider the CEV model (7.9). Let $\chi^2_\nu(\gamma)$ be a non-central chi-square distributed variable with ν degrees of freedom and non-centrality parameter γ, and let $\Upsilon(x,\nu,\gamma) = \mathrm{P}(\chi^2_\nu(\gamma) \leq x)$ be the cumulative distribution function for $\chi^2_\nu(\gamma)$. Also define*

$$a = \frac{K^{2(1-p)}}{(1-p)^2\lambda^2(T-t)}, \quad b = |p-1|^{-1}, \quad c = \frac{S^{2(1-p)}}{(1-p)^2\lambda^2(T-t)}.$$

Then, for $0 < p < 1$ and an absorbing boundary at $S = 0$ we have, for $K > 0$,

$$c_{\mathrm{CEV}}(t,S;T,K) = S\left(1 - \Upsilon(a,b+2,c)\right) - K\Upsilon(c,b,a). \tag{7.13}$$

Remark 7.2.7. The result above in fact holds for all $p < 1$, including negative p. A complimentary result holds for $p > 1$,

$$c_{\mathrm{CEV}}(t,S;T,K) = S\left(1 - \Upsilon(c,b,a)\right) - K\Upsilon(a,b+2,c). \tag{7.14}$$

Remark 7.2.8. The special case $p = 1$ leads to the Black pricing formula with volatility λ, see (1.43) and Remark 1.9.4), so that

$$c_{\mathrm{B}}(t,S;T,K;\lambda) = S\Phi(d_+) - K\Phi(d_-), \tag{7.15}$$

where

$$d_\pm = \frac{\ln(S/K) \pm \lambda^2(T-t)/2}{\lambda\sqrt{T-t}},$$

and $\Phi(\cdot)$ is the standard Gaussian CDF.

Remark 7.2.9. For the case $p = 0$, if we remove the assumption of an absorbing barrier at the origin, $S(t)$ is a Gaussian process. In this case, it is straightforward to compute that the option pricing formula, sometimes called the *Normal*, *Gaussian* or *Bachelier* pricing formula with (Normal) volatility[4] λ, becomes

$$c_{\mathrm{N}}(t,S;T,K;\lambda) = (S-K)\,\Phi(d) + \lambda\sqrt{T-t}\phi(d), \quad d = \frac{S-K}{\lambda\sqrt{T-t}}, \tag{7.16}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard Gaussian CDF and PDF, respectively.

Further details about the non-central chi-square distribution can be found in Chapter 3. A number of efficient numerical algorithms exist to compute $\Upsilon(x,\nu,\gamma)$; see Johnson et al. [1995] for a survey. A standard algorithm can be found in Ding [1992]. Figure 7.1 on page 298 gives some examples of volatility skews produced by the CEV model.

[4] Also known as Gaussian volatility; when applied to interest rates, Gaussian volatilities are often called *basis-point*, or *bp*, volatilities.

7.2.3 Regularization

As discussed earlier, the CEV process implies a positive probability of absorption at $S = 0$ (for $p < 1$). This phenomenon is not necessarily a problem for pricing of simple European call options, but is obviously not desirable from an empirical standpoint[5], and might also create some difficulties in pricing of more exotic structures. To avoid absorption, we can specify a regularized version of the CEV model by letting,

$$\varphi(x) = x \min\left(\varepsilon^{p-1}, x^{p-1}\right), \quad \varepsilon > 0, p < 1. \tag{7.17}$$

Roughly speaking, when $S(t)$ crosses the level ε, the resulting process becomes (locally) a geometric Brownian motion with finite volatility ε^{p-1}. With $\varphi(x)$ now Lipschitz continuous, it is straightforward to verify that the process for $S(t)$ can no longer reach the origin. On the other hand, the specification (7.17) will not allow for closed-form call option pricing but will, in principle at least, require the usage of numerical methods such as the finite difference method (see Section 7.4). On the other hand, for small to moderate values of ε, we would expect the CEV pricing formulas from Proposition 7.2.6 to hold as a good approximation. Andersen and Andreasen [2000b] verify numerically that this holds quite robustly, for strikes not too far from the spot value of S. More formally, we have the following result:

Proposition 7.2.10. *For $p < 1$ and $\varepsilon > 0$, let*

$$\begin{aligned} dx(t) &= \lambda x(t)^p \, dW(t), \\ dy(t) &= \lambda y(t) \min\left(\varepsilon^{p-1}, y(t)^{p-1}\right) \, dW(t), \end{aligned}$$

where $x(0) = y(0) > 0$ and $W(t)$ is a one-dimensional Brownian motion in measure P. *For $p < 1/2$, 0 is assumed to be an absorbing boundary for x. For some $T > t$ and some constant K, we then have*

$$\lim_{\varepsilon \downarrow 0} \left| \mathrm{P}\left(x(T) < h\right) - \mathrm{P}\left(y(T) < h\right) \right| = 0,$$

$$\lim_{\varepsilon \downarrow 0} \left| \mathrm{E}\left((x(T) - K)^+\right) - \mathrm{E}\left((y(T) - K)^+\right) \right| = 0.$$

The result is intuitive, but the proof is somewhat technical, and we skip it. Details can be found in Andersen and Andreasen [2000b].

[5] As the measure P is equivalent to the real-life (statistical) measure, a non-zero probability of absorption under P implies a non-zero probability of absorption under the real-life measure.

7.2.4 Displaced Diffusion Models

An easy extension of the CEV model that is sometimes useful involves adding a displacement constant to the CEV specification. Specifically, we write

$$\varphi(x) = (\alpha + x)^p \tag{7.18}$$

for some constant α. In the process (7.1), (7.18), let us set $Z(t) = \alpha + S(t)$. By Ito's lemma, $Z(t)$ then satisfies the CEV SDE

$$dZ(t) = \lambda Z(t)^p \, dW(t).$$

With $Z(t)$ having an absorbing boundary at 0, $S(t)$ then must have an absorbing boundary at $-\alpha$. Call option pricing with (7.18) is straightforward:

Proposition 7.2.11. *Let*

$$c_{\text{DCEV}}(t, S(t); T, K, \alpha) = \mathrm{E}_t\left((S(T) - K)^+\right)$$

be the call option price associated with the displaced CEV process (7.1), (7.18). Then

$$c_{\text{DCEV}}(t, S; T, K, \alpha) = c_{\text{CEV}}(t, S + \alpha; T, K + \alpha), \quad S, K > -\alpha, \tag{7.19}$$

where the right-hand side is given by Proposition 7.2.6.

Proof. The result follows directly from the observation that

$$\mathrm{E}_t\left((S(T) - K)^+\right) = \mathrm{E}_t\left((Z(T) - (K + \alpha))^+\right),$$

where $Z(t) = \alpha + S(t)$ follows a regular CEV process. □

Introduction of the displacement constant α allows for a (somewhat) richer family of volatility smiles than those of the pure CEV specification. In practice, however, the main use of displacement constants is for the special case of the *displaced log-normal*, or *shifted log-normal*, process where $p = 1$. The call option price formula for this case is listed below, for later reference.

Proposition 7.2.12. *Consider the displaced log-normal process*

$$dS(t) = \lambda(\beta + \zeta S(t)) \, dW(t), \tag{7.20}$$

where $W(t)$ is a one-dimensional Brownian motion in measure P, *and $\zeta, \lambda \neq 0$. Assuming $S(t), K > -\beta/\zeta$, we have*

$$\begin{aligned} c_{\text{DLN}}(t, S(t); T, K) &\triangleq \mathrm{E}_t\left((S(T) - K)^+\right) \\ &= \left(S(t) + \frac{\beta}{\zeta}\right)\Phi(d_+) - \left(K + \frac{\beta}{\zeta}\right)\Phi(d_-), \\ d_\pm &= \frac{\ln\left(\frac{S(t)+\beta/\zeta}{K+\beta/\zeta}\right) \pm \frac{1}{2}\zeta^2\lambda^2(T-t)}{\zeta\lambda\sqrt{T-t}}. \end{aligned}$$

Proof. The result follows directly from the Black-Scholes equation (see Section 1.9) and (7.19), after setting $\alpha = \beta/\zeta$ and writing $\lambda(\beta + \zeta S(t)) = \lambda\zeta(\alpha + S(t))$. □

Remark 7.2.13. It is often convenient to rewrite the displaced log-normal process in a slightly different form

$$dS(t) = \sigma\left(bS(t) + (1-b)\,L\right)\,dW(t). \tag{7.21}$$

The parameter L is often set to near, or exactly at, the initial value $S(0)$. In this parameterization, σ is expressed in the units of relative volatility, just like in the Black model, because $bS(0) + (1-b)L \approx S(0)$. In particular, σ always has the same scale for all values of b. Moreover, the effects of σ and b are almost "orthogonal", in the sense that the parameter σ changes the overall level of the implied volatility smile but not its slope, whereas b only changes the slope (skew) of the implied volatility smile but not its overall level (i.e. not the at-the-money implied volatility). We use the parameterization (7.21) extensively in later chapters.

Remark 7.2.14. Consider the general local volatility model (7.1). Expanding the local volatility function $\varphi(\cdot)$ around at-the-money to the first order, we obtain

$$dS(t) \approx \lambda\left(\varphi\left(S(0)\right) + \varphi'\left(S(0)\right)\left(S(t) - S(0)\right)\right)\,dW(t),$$

which we identify as being of the form (7.21) with

$$\sigma = \lambda\frac{\varphi\left(S(0)\right)}{S(0)}, \quad b = \varphi'\left(S(0)\right)\frac{S(0)}{\varphi\left(S(0)\right)}, \quad L = S(0).$$

Hence, a first-order approximation to any local volatility process is of displaced log-normal type. In view of this, displaced log-normal processes are extensively used in various types of approximations and asymptotic expansions.

The previous remark can be applied to the CEV process:

$$\sigma = \lambda S(0)^{p-1}, \quad b = p, \quad L = S(0).$$

The approximation of the CEV process with (7.21) turns out to be particularly close, and we later use it to increase the tractability of certain stochastic volatility models. We also use it as a justification to freely switch from one type of process to the other. It is worth noting, however, that (7.20) has certain drawbacks relative to a pure CEV process. First, the process for $S(\cdot)$ can become negative if β (as is usual) is positive. Second, in stochastic volatility applications the asymptotic linear growth of $\varphi(x)$ in x can sometimes lead to technical problems and unbounded second moments of $S(\cdot)$. We shall return to this issue shortly, in Chapter 8.

7.3 Quadratic Volatility Model

In practice, volatility smiles in fixed income markets are not always perfectly monotonic in strike; indeed, as mentioned earlier, for sufficiently high strikes it is not uncommon for the smile to reverse direction and start increasing in strike. This type of behavior is inconsistent with a pure CEV model, but can, to some extent, be captured by the displaced CEV specification $\varphi(x) = (\alpha + x)^p$. Often, however, this model is hard to fit to actual data. A more powerful approach involves overlaying the CEV process with stochastic volatility, something that we turn to in Chapter 8. If we here wish to stay within the realm of DVF processes, one way to generate arbitrarily convex smiles is to use a *quadratic volatility model*, where

$$\varphi(x) = \alpha + \beta x + \gamma x^2, \tag{7.22}$$

for constants α, β, γ. We develop some aspects of this model here, but remind the reader of the caveats discussed in Section 7.1.3; in particular, for the model to be realistic, γ should probably be small.

Before commencing with derivations, let us note that the behavior of a DVF model (7.1) equipped with volatility function (7.22) will depend strongly on the root configuration in the quadratic polynomial $\alpha + \beta x + \gamma x^2$. For instance, if φ has two real roots l, u, $l < u$, straddling the initial value $S(0)$, it is clear that $S(t)$ will itself be bound to this range, i.e. $S(t) \in [l, u]$. Specifically, whenever $S(t)$ gets close to either l or u, $\varphi(x)$ will approach zero and the diffusion for $S(t)$ will gradually slow down. As such range-bound dynamics are rather unrealistic for interest rate applications[6], we do not consider it in the following.

7.3.1 Case 1: Two Real Roots to the Left of $S(0)$

We first consider the case where $\alpha + \beta x + \gamma x^2$ has two real roots l and u, $l < u$, both lying to the left of $S(0)$. Without loss of generality, we may then consider the normalized process

$$dS(t) = \frac{(S(t) - u)(S(t) - l)}{u - l}\, dW(t), \quad S(0) > u > l. \tag{7.23}$$

We start by listing a few lemmas.

Lemma 7.3.1. *The range for $S(t)$ in (7.23) is $S(t) \in (u, \infty)$. In particular, the process for $S(t)$ does not explode in measure* P.

[6]For an application of the range-bound quadratic model to FX markets (where currency controls may potentially create upper and lower bounds), see Ingersol [1997].

Proof. That $S(t)$ cannot go below u is obvious; further, Feller's boundary criteria (e.g. Karlin and Taylor [1981], Chapter 15.6) establishes that u is not accessible when $S(0) > u$. As $S(t)$ is described by a time-homogeneous one-dimensional SDE, it cannot explode (Karatzas and Shreve [1997], p. 332). □

While the process for $S(t)$ is non-explosive, the super-linear growth[7] of $\varphi(x)$ causes some interesting technical problems, In particular, we have the following result, proved in Andersen [2010].

Lemma 7.3.2. *The process (7.23) is a strict supermartingale in measure* P.

As the process for S is not a martingale, the usual pricing results require some modifications. For the purpose of pricing puts and calls, we need use the following.

Lemma 7.3.3. *Suppose that $S(t)$ satisfies (7.23) in some measure* P *and assume that put-call parity holds. Then the prices at time 0 for the put (p) and call (c) are*

$$p\left(0, S(0); T, K\right) = \mathrm{E}\left(\left(K - S(T)\right)^{+}\right),$$

$$c\left(0, S(0); T, K\right) = p\left(0, S(0); T, K\right) + S(0) - K > \mathrm{E}\left(\left(S(T) - K\right)^{+}\right).$$

Proof. (Sketch only). In the absence of arbitrage, the put price is a local martingale in measure P. As a bounded local martingale is a martingale and the put payout is bounded between 0 and $K - u$, it follows then that the put price in fact must be a true P-martingale. The expression for $p(0, S(0); T, K)$ follows. Applying put-call parity (see Chapter 1) yields the result for $c(0, S(0); T, K)$, where the inequality follows from Lemma 7.3.2. □

We emphasize the non-standard result $c(0, S(0); T, K) > \mathrm{E}(c(T, S(T); T, K))$ which is a consequence of the supermartingale property of $S(t)$. The inequality holds for arbitrarily large strikes; indeed, rather counter-intuitively, $\lim_{K\to\infty} c(0, S(0); T, K) = S(0) - \mathrm{E}(S(T)) > 0$. We should also note that our assumption of put-call parity being valid is critical here, as it allows us to produce unique prices of both puts and calls. As described in Heston et al. [2007] and Andersen [2010], it is, however, possible to work with other assumptions without violating no-arbitrage.

With 7.3.3 we are now ready to tackle the derivation of an option pricing formula. We will be using the shorthand

$$p(t) \triangleq p(t, S(t); T, K),$$

and so forth. First, notice the useful relationship

[7] A similar issue is present in CEV processes with $p > 1$, as noted earlier.

$$S - K = \frac{(S-u)(K-l) - (K-u)(S-l)}{u-l}, \tag{7.24}$$

which allows us to write

$$\begin{aligned} p(T) &= \frac{1}{u-l}\left((K-u)(S(T)-l) - (S(T)-u)(K-l)\right)^+ \\ &= \frac{(K-u)(S(T)-l)}{u-l} 1_{\{(K-u)(S(T)-l)-(S(T)-u)(K-l)>0\}} \\ &\quad - \frac{(S(T)-u)(K-l)}{u-l} 1_{\{(K-u)(S(T)-l)-(S(T)-u)(K-l)>0\}} \\ &\triangleq p_1(T) - p_2(T). \end{aligned} \tag{7.25}$$

The payouts p_1 and p_2 have identical structure, so it suffices to focus our attention on pricing one of them, e.g. p_1.

From Lemma 7.3.3, we have $p_1(0) = \mathrm{E}(p_1(T))$, which we rewrite as

$$p_1(0) = \frac{K-u}{u-l}\mathrm{E}\left((S(T)-l)\, 1_{\{(S(T)-u)/(S(T)-l)<(K-u)/(K-l)\}}\right). \tag{7.26}$$

At this point our first instinct would be to perform a measure shift that eliminates that factor $S(T) - l$ in the expectation, i.e. we would like to introduce a new measure $\widetilde{\mathrm{P}}$ such that

$$\widetilde{\mathrm{P}}(B) = \frac{1}{S(0)-l}\mathrm{E}((S(T)-l)B),$$

for any $\mathcal{F}_T$-measurable event B. We recall, however, that $S(t)$ (and therefore $S(t)-l$) is not a martingale in P, so such a measure shift cannot be performed outright. Let us nevertheless try. Proceeding mechanically as if $S(t)$ *were* a martingale, we would get, for the process $Y(t) \triangleq (S(t)-u)/(S(t)-l)$,

$$dY(t) \stackrel{?}{=} Y(t)\, d\widetilde{W}(t), \quad Y(0) = \frac{S(0)-u}{S(0)-l} < 1, \tag{7.27}$$

where $\widetilde{W}$ is a Brownian motion in $\widetilde{\mathrm{P}}$. Clearly, however, there are technical problems here: the range for $Y(t)$ in (7.27) is $[0,\infty)$, whereas we know that in measure P we have $Y(t) \in (0,1)$ (since $S(t) \in (u,\infty)$); the two measures therefore cannot be equivalent. For option pricing purposes, it turns out that the correct way to handle the technical conflict involves inserting an *absorbing boundary* at $Y = 1$ in (7.27).

Proposition 7.3.4. *Let*

$$dY(t) = Y(t)\, d\widetilde{W}(t), \quad Y(0) = \frac{S(0)-u}{S(0)-l} < 1,$$

be geometric Brownian motion in $\widetilde{\mathrm{P}}$. *Define* $\tau = \inf\{t > 0 : Y(t) = 1\}$, *and let* $K > u$. *Then* $p_1(0)$ *in (7.26) is given by*

$$p_1(0) = \frac{(K-u)(S(0)-l)}{u-l}\mathrm{E}^{\widetilde{\mathrm{P}}}\left(1_{\{Y(T)<(K-u)/(K-l)\}}1_{\{\tau>T\}}\right). \tag{7.28}$$

Stated explicitly,

$$p_1(0) = K_1\Phi\left(\frac{-\ln(X_1/K_1)+T/2}{\sqrt{T}}\right) - X_2\Phi\left(\frac{\ln(X_2/K_2)+T/2}{\sqrt{T}}\right), \tag{7.29}$$

with Φ being the Gaussian cumulative distribution function, and

$$K_1 = \frac{(K-u)(S(0)-l)}{u-l}, \quad X_1 = \frac{(S(0)-u)(K-l)}{u-l},$$
$$K_2 = \frac{(K-l)(S(0)-l)}{u-l}, \quad X_2 = \frac{(S(0)-u)(K-u)}{u-l}.$$

Proof. The result (7.28) is proven in Andersen [2010]. The result (7.29) follows by direct calculations, similar to those leading to the Black-Scholes-Merton formula. □

Following similar steps leads to an expression for $p_2(0)$, which in turn leads to the following result for $p(0) = p_1(0) - p_2(0)$.

Proposition 7.3.5. *Let K_i, X_i, $i = 1, 2$, be given as in Proposition 7.3.4. Assuming $K > u$, the put price $p(0)$ for the model (7.23) has the explicit representation*

$$p(0, S(0); T, K) = K_1\Phi\left(-d_-^{(1)}\right) - X_2\Phi\left(d_+^{(2)}\right) - X_1\Phi\left(-d_+^{(1)}\right) + K_2\Phi\left(d_-^{(2)}\right),$$

$$d_\pm^{(i)} = \frac{\ln(X_i/K_i) \pm T/2}{\sqrt{T}}, \quad i = 1, 2.$$

An application of put-call parity then immediately gives the call price:

Corollary 7.3.6. *Assuming put-call parity holds, the call price for the model (7.23) is*

$$c(0, S(0); T, K) = S(0) - K + p(0, S(0); T, K),$$

with $p(0, S(0); T, K)$ given in Proposition 7.3.5.

We recall that Proposition 7.3.5 applies to (7.23), rather than our original process which, at the root configuration in question, is

$$dS(t) = \lambda\gamma(S(t)-u)(S(t)-l)\,dW(t) = q\frac{(S(t)-u)(S(t)-l)}{u-l}\,dW(t), \tag{7.30}$$

where $q = \lambda\gamma(u-l)$. The constant in front of the quadratic polynomial is easily handled by time-scaling: to price options in (7.30) we simply set the put price equal to $p(0, S(0); q^2T, K)$, where $p(0, S(0); \cdot, K)$ is given by the formula in Proposition 7.3.5.

7.3.2 Case 2: One Real Root to the Left of $S(0)$

If we let the single root to $\alpha + \beta x + \gamma x^2$ be denoted u, $u < S(0)$, it suffices to consider the normalized process

$$dS(t) = (S(t) - u)^2 \, dW(t). \tag{7.31}$$

But this process is a special case, with power equal to 2, of the displaced CEV model in Section 7.2.4, and the option pricing formulas from that section then apply directly. As these formulas are rather complicated in their dependence on the non-central chi-square distribution, it is worthwhile noticing that simple expressions exist for the special case of power equal to 2. The result is listed below.

Proposition 7.3.7. *For the process (7.31), the put option price is*

$$\begin{aligned} p(0, S(0); T, K) = (S(0) - u)\,(K - u)\,\sqrt{T} \\ \times \left\{ d_+ \Phi(d_+) + \phi\,(d_+) - d_- \Phi(d_-) - \phi\,(d_-) \right\}, \end{aligned}$$

where $\phi(x)$ is the Gaussian density, and

$$d_\pm = \frac{\pm \frac{1}{S(0)-u} - \frac{1}{K-u}}{\sqrt{T}}.$$

Proof. We observe that for the process

$$dS(t) = (S(t) - u)\,(S(t) - l)\,dW(t), \quad l < u < S(0),$$

the put price can be computed from the result in Proposition 7.3.5, after a time-change, from T to $T(u - l)^2$; see the comments at the end of Section 7.3.1. Taking the limit of the put price as $l \uparrow u$ then establishes the result. □

The call option price can, as before, be found by put-call parity. To establish put and call option prices for the original diffusion

$$dS(t) = \lambda \left(\alpha + \beta S(t) + \gamma S(t)^2 \right) dW(t) = \lambda \gamma \, (S(t) - u)^2 \, dW(t),$$

we simply change T to $\lambda^2 \gamma^2 T$ in Proposition 7.3.7.

7.3.3 Extensions and Other Root Configurations

The results listed in Sections 7.3.1 and 7.3.2 have given a flavor of how to deal with quadratic volatility process, and shall suffice for the purposes of this book. Other root configurations are treated in detail in Andersen [2010], including the case where $\varphi(x)$ has no roots (in which case the put and call option price formulas are infinite sine-series). Andersen [2010] also discusses the case where an absorbing barrier has been inserted at the origin to prevent $S(t)$ from going negative.

7.4 Finite Difference Solutions for General φ

For general specifications of φ, closed-form solutions for European options will not exist. In such cases, we may instead rely on the finite difference methods discussed in Chapter 2. Consider again the evaluation of

$$c(t, S(t); T, K) = \mathrm{E}_t\left((S(T) - K)^+\right),$$

with $S(t)$ following (7.1). With suitable regularity conditions on φ, the Feynman-Kac theorem of Section 1.8 shows that $c(t, S) = c(t, S; T, K)$ (with T, K fixed) satisfies the PDE

$$\frac{\partial c(t,S)}{\partial t} + \frac{1}{2}\lambda^2 \varphi(S)^2 \frac{\partial^2 c(t,S)}{\partial S^2} = 0, \tag{7.32}$$

subject to the terminal condition

$$c(T, S) = (S - K)^+. \tag{7.33}$$

This PDE can be solved numerically using, say, the Crank-Nicholson finite difference grid method in Chapter 2. A direct discretization of (7.32) is normally sufficient, but we note that it may occasionally be possible to take advantage of special forms of φ and introduce transformations of S to improve the properties of the finite difference scheme. For example, as we have already seen in Chapter 2, when $\varphi(S) = S$, it is customary (and appropriate) to introduce $y(S) = \ln S$ and discretize in y. More generally, for sufficiently regular φ, the transformation

$$y(S) = \int \frac{dS}{\varphi(S)} \tag{7.34}$$

(see (2.81)–(2.82)) might offer numerical advantages over a direct discretization provided, of course, that the inverse in (7.34) exists. The following semi-heuristic argument explains the rationale. With the transform (7.34), the SDE for $y(t) = y(S(t))$ is (ignoring the drift)

$$dy(t) = O(dt) + \lambda\, dW(t).$$

The diffusion coefficient in the process for y is independent of the state of S, suggesting that a differential operator expressed in terms of y may have better numerical properties than the one expressed in terms of S. Even if y is not used for discretization, the transformation (7.34) suggests the discretization grid in the S-domain. In particular, $\{S_n\}_{n=0}^{m+1}$ can be defined by the condition that $y_n = y(S_n)$, $n = 0, \ldots, m+1$, are equidistant over $[y(S_0), y(S_{m+1})]$. For $n = 0, \ldots, m+1$ this gives ($y^{-1}(\cdot)$ is the inverse transform of (7.34))

$$y_n = y(S_0) + \frac{n}{m+1}\left(y(S_{m+1}) - y(S_0)\right),$$

$$S_n = y^{-1}(y_n) = y^{-1}\left(y(S_0) + \frac{n}{m+1}\left(y(S_{m+1}) - y(S_0)\right)\right).$$

7.4.1 Multiple λ and T

In applications, we often need to compute the values of $c(t, S; T, K)$ for many different values of T and/or λ. This need arises, for instance, in a standard model calibration exercise where we use a root-search algorithm to determine the value of λ that will make the computed call prices at different maturities T equal to values observed in the market. In such cases, we note that one should *not* simply solve (7.32) over and over (at great computational expense), but instead rely on the following observation:

Proposition 7.4.1. *Let $g(\tau, x)$ solve the following PDE*

$$-\frac{\partial g(\tau,x)}{\partial \tau} + \frac{1}{2}\varphi(x)^2 \frac{\partial^2 g(\tau,x)}{\partial x^2} = 0, \tag{7.35}$$

with initial *condition*

$$g(0,x) = (x-K)^+ . \tag{7.36}$$

Let $c(t,S)$ solve the backward PDE (7.32)–(7.33) for a given value of λ. Then

$$c(t,S;T,K) = g\left(\lambda^2(T-t), S\right). \tag{7.37}$$

Proof. Follows directly from a variable transformation $\tau(t) = \lambda^2(T-t)$ in (7.32)–(7.33), taking advantage of the time-homogeneity of φ. □

Using finite difference techniques to solve the PDE (7.35), we can construct the function g on a (τ, S)-grid; once this grid is stored in memory, (7.37) is used to recover $c(t, S; T, K)$ for arbitrary choices of S, λ and T by simple lookup or interpolation. We emphasize that this approach involves the numerical solution of only a *single* PDE. Also note that PDE is solved forward in time from a known *initial* condition, rather than backwards from a *terminal* condition.

7.4.2 Forward Equation for Call Options

While the function g from (7.35) is conveniently independent of T and λ, it does depend on K through the initial condition (7.36). In some applications, we may wish to use different strikes for different values of T, in which case the approach in Section 7.4.1 requires us to numerically solve as many finite difference grids as there are different values of K. We can improve on this by replacing the backward equation (7.32) with the forward equation of Dupire [1994]. In this approach, calendar time t and the initial value of S are considered fixed, whereas maturity T and strike K are variable. In view of this, we define $c(T,K) = c(t, S; T, K)$ for fixed t, S. We need the following proposition:

Proposition 7.4.2. *Define the function* $c(T,K) \triangleq c(t,S;T,K)$ *where* t, S *are fixed and* $c(t,S;T,K)$ *is defined by (7.3) for the model (7.1). Then* $c(T,K)$ *satisfies the forward PDE*

$$-\frac{\partial c(T,K)}{\partial T} + \frac{1}{2}\lambda^2\varphi(K)^2\frac{\partial^2 c(T,K)}{\partial K^2} = 0, \tag{7.38}$$

for $T > t$, *subject to the time* t *initial condition*

$$c(t,K) = (S-K)^+ .$$

Proof. In Dupire [1994], the result is proven by combining the Fokker-Planck equation (see Section 1.8) with the result (7.5), followed by a series of integrations. A more intuitive line of attack proceeds as follows. Consider the function $H(t) = (S(t)-K)^+$. While $H(t)$ clearly does not satisfy the smoothness requirements of Ito's lemma, the Tanaka extension nevertheless justifies the following result, obtained by formally applying Ito's lemma to H:

$$dH(t) = 1_{\{S(t)>K\}}\lambda\varphi\left(S(t)\right)\,dW(t) + \frac{1}{2}\delta\left(S(t)-K\right)\lambda^2\varphi\left(S(t)\right)^2\,dt. \tag{7.39}$$

That is,

$$\begin{aligned} H(T) &= H(t) + \int_t^T 1_{\{S(u)>K\}}\lambda\varphi(S(u))\,dW(u) \\ &\quad + \frac{1}{2}\int_t^T \delta\left(S(u)-K\right)\lambda^2\varphi\left(S(u)\right)^2\,du \\ &= H(t) + M(T) + \frac{1}{2}\int_t^T \delta\left(S(u)-K\right)\lambda^2\varphi(K)^2\,du, \end{aligned}$$

where δ is the Dirac delta function and $M(t)$ is a continuous martingale with $M(t) = 0$. From (7.3), we have that

$$\begin{aligned} c\left(t,S(t);T,K\right) &= \mathrm{E}_t\left(H(T)\right) \\ &= H(t) + \frac{1}{2}\int_t^T \mathrm{E}_t\left(\delta\left(S(u)-K\right)\right)\lambda^2\varphi(K)^2\,du \\ &= H(t) + \frac{1}{2}\lambda^2\varphi(K)^2\int_t^T \frac{\partial^2 c\left(t,S(t);u,K\right)}{\partial K^2}\,du, \end{aligned}$$

where we have used the martingale property of M as well as the result (7.5). Differentiating this equation with respect to T gives the result in Proposition 7.4.2. □

As mentioned in Chapter 1, the term $\int_t^T \delta(S(u)-K)\,du$ in the expression for $H(T)$ is known as the *local time* of $S(\cdot)$ at the level K. Local time and the Tanaka extension are deep subjects (see Karatzas and Shreve [1997] for a

formal discussion) and have many interesting applications in finance, see for instance Andersen et al. [2002], Andersen and Andreasen [2000a], Andersen and Buffum [2003], Henderson and Hobson [2000], Carr and Jarrow [1990], Carr and Wu [2003], among many others.

We emphasize that while the backward equation (7.32) holds for European derivative securities on S in general, the forward equation (7.38) is unique to calls and puts, as only put and call payouts allow for the basic result (7.5).

Equipped with Proposition 7.4.2, the following result immediately follows from the proof of Proposition 7.4.1. Notice the difference in the initial conditions (7.36) and (7.40).

Proposition 7.4.3. *Let $h(\tau, x)$ solve the following PDE*

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

with initial condition

$$h(0,x) = (S-x)^+ . \tag{7.40}$$

Then

$$c(t,S;T,K) = h\left(\lambda^2(T-t), K\right).$$

As long as the initial value of $S(t)$ is kept constant, the result in Proposition 7.4.3 allows us use a single finite difference grid to price call options with multiple maturities, strikes, and λ's. We note, however, that in many applications $S(t)$ may in fact be T-dependent, as S will often represent, say, T-maturity Libor forward rates. In such cases, the question of whether Proposition 7.4.3 leads to a more efficient numerical scheme than Proposition 7.4.1 is settled by comparing the number of strikes and the number of spot levels involved.

7.5 Asymptotic Expansions for General φ

As we have shown, there are a number of "tricks" that can be employed to make the application of finite difference methods a computationally viable approach to pricing a large number of European call options. Nevertheless, there is significant convenience and computer code simplification associated with closed-form pricing formulas, so we now turn to the development of asymptotic approximations for the solution to the generic backward PDE (7.32). There are a number of approaches that can be taken, including the "most likely path" method in Gatheral [2001] (see also Gatheral [2006] and Section 22.1.7, and the singular perturbation techniques in Hagan and Woodward [1999b], Henry-Labordére [2008], Gatheral et al. [2009], to name a few. Our presentation here is based on a fairly straightforward, yet often highly accurate, asymptotic expansion in time to maturity.

7.5.1 Expansion around Displaced Log-Normal Process

As in Proposition 7.4.1, we start by writing $c(t,S;T,K) = g(\tau,S)$, where $\tau = \lambda^2(T-t)$ and g satisfies (7.35). Inspired by the known solution of (7.35) in Proposition 7.2.12 for the case $\varphi(x) = \beta + \zeta x$, $\zeta \neq 0$, let us guess at a solution of (7.35) of the form

$$g(\tau,S) = \left(S + \frac{\beta}{\zeta}\right)\Phi(z_+) - \left(K + \frac{\beta}{\zeta}\right)\Phi(z_-), \tag{7.41}$$

$$z_\pm = \frac{\ln\left(\frac{S+\beta/\zeta}{K+\beta/\zeta}\right) \pm \frac{1}{2}\Omega(\tau,S)^2}{\Omega(\tau,S)},$$

where the function $\Omega(\tau,S)$ is to be determined. In (7.41), note that we obviously must assume that $S,K > -\beta/\zeta$. Substituting (7.41) into (7.35) gives the following PDE for $\Omega(\tau,S)$:

$$\left(S + \frac{\beta}{\zeta}\right)^2 \Omega \frac{\partial \Omega}{\partial \tau} = \frac{1}{2}\varphi(S)^2 \left[\left(S + \frac{\beta}{\zeta}\right)^2 \Omega \frac{\partial^2 \Omega}{\partial S^2} + (1 - h_{-3}) - h_1(1 - h_{-1})\right], \tag{7.42}$$

where

$$h_i \triangleq \left(S + \frac{\beta}{\zeta}\right)\frac{\partial \Omega}{\partial S}\left(\Omega^{-1}\ln\left(\frac{S+\beta/\zeta}{K+\beta/\zeta}\right) + \frac{1}{2}i\Omega\right), \quad i = -3,-1,1.$$

The PDE (7.42) does not generally allow for an explicit solution, so we resort to an asymptotic expansion in τ.

Proposition 7.5.1. *An asymptotic expansion for the solution of (7.35) is given by (7.41), with*

$$\Omega(\tau,S) = \Omega_0(S)\tau^{1/2} + \Omega_1(S)\tau^{3/2} + O\left(\tau^{5/2}\right), \tag{7.43}$$

$$\Omega_0(S) = \ln\left(\frac{S+\beta/\zeta}{K+\beta/\zeta}\right)\left(\int_K^S \varphi(u)^{-1}\,du\right)^{-1}, \tag{7.44}$$

$$\Omega_1(S) = -\frac{\Omega_0(S)}{\left(\int_K^S \varphi(u)^{-1}\,du\right)^2}\ln\left(\Omega_0(S)\left(\frac{(S+\beta/\zeta)(K+\beta/\zeta)}{\varphi(S)\varphi(K)}\right)^{1/2}\right), \tag{7.45}$$

where the parameters β and ζ can be chosen arbitrarily, subject to the constraints $S,K > -\beta/\zeta$ and $\zeta \neq 0$.

Proof. In (7.41) we clearly require $\Omega(\tau, S) \sim \tau^{1/2}$ as $\tau \to 0$, so we seek a small-time solution of the form

$$\Omega(\tau, S) = \sum_{i \geq 0} \tau^{i+1/2} \Omega_i(S). \tag{7.46}$$

Notice that (7.46) omits all integer powers of τ — it turns out that their weights are all identically 0. Substituting (7.46) into (7.42) and matching terms of order $O(1)$ gives

$$(S + \beta/\zeta)^2 \, \Omega_0^2 = \varphi(S)^2 \left(1 - \frac{\Omega_0'}{\Omega_0} (S + \beta/\zeta) \ln \left(\frac{S + \beta/\zeta}{K + \beta/\zeta} \right) \right)^2, \tag{7.47}$$

where the prime denotes differentiation with respect to S. Taking the square root of the above equation and rearranging leads to two first-order ordinary differential equations of the Bernoulli type. Solving (7.47) subject to the boundary condition that the limit of Ω_0 must be finite for $S \to K$ (and discarding the negative solution) leads to (7.44).

Progressing now to the $O(\tau)$ term in (7.42), we get

$$\begin{aligned} 2 (S + \beta/\zeta) \, \Omega_1 = {} & \frac{1}{2} \varphi(S)^2 \left((S + \beta/\zeta) \, \Omega_0'' + \Omega_0' \right) \\ & - \varphi(S) (S + \beta/\zeta) \ln \left(\frac{S + \beta/\zeta}{K + \beta/\zeta} \right) \left(\frac{\Omega_1'}{\Omega_0} - \frac{\Omega_1 \Omega_0'}{\Omega_0^2} \right). \end{aligned}$$

Inserting the result for Ω_0 and rearranging again leads to a Bernoulli-type ODE, the explicit solution of which is (7.45). As before, we have ensured that the limit $S \to K$ is finite. □

Remark 7.5.2. We notice that $\Omega_0(K)$ and $\Omega_1(K)$ in Proposition 7.5.1 exist by construction. Taking the limit $S \to K$ explicitly, we get

$$\begin{aligned} \Omega_0(K) &= \frac{\varphi(K)}{K + \beta/\zeta}, \\ \Omega_1(K) &= \frac{1}{24} \Omega_0(K)^3 \left[1 + (K + \beta/\zeta)^2 \, \varphi(K)^{-2} \left(2\varphi(K)\varphi''(K) - \varphi'(K)^2 \right) \right]. \end{aligned}$$

While Proposition 7.5.1 only includes two terms in the expansion for Ω, it is possible to compute further terms if necessary. Such terms become increasingly cumbersome however, and typically do not add much further accuracy.

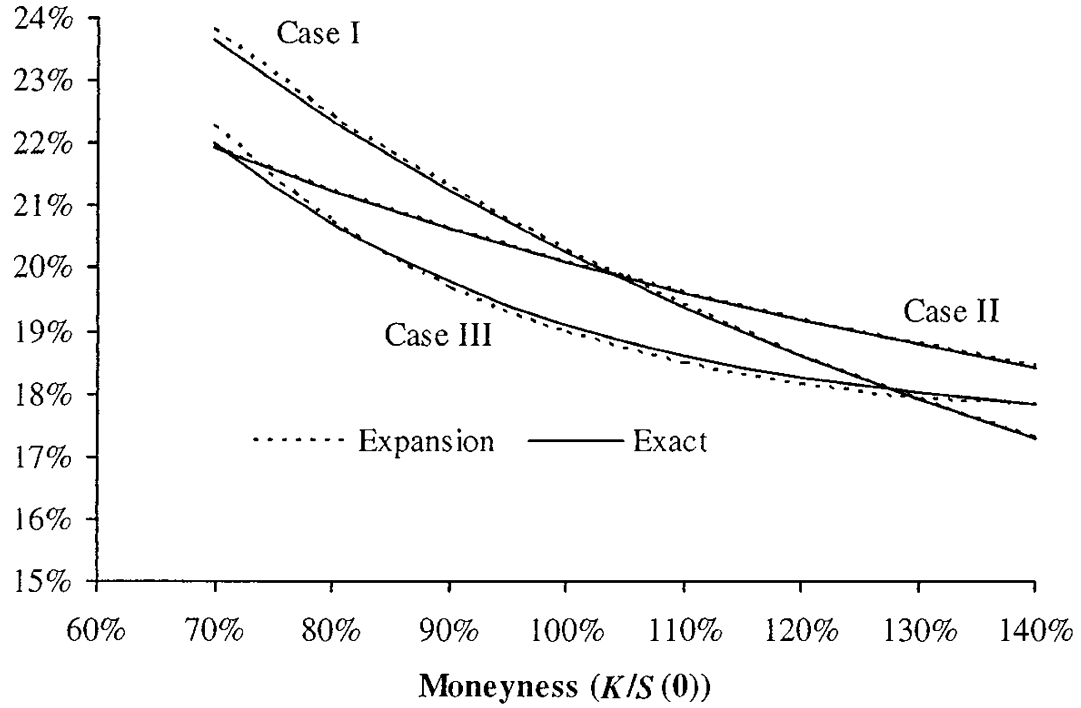
The best choice of the parameters β and ζ is not always obvious. One choice is to use Remark 7.2.14. Alternatively, we could think of a more global approach and, roughly speaking, set them in such a way that the straight line $\beta + \zeta x$ would provide as good a fit to $\varphi(x)$ as possible, over the

statistically relevant range of x. Sometimes, we can use a Taylor expansion around $x = (S + K)/2$, say, and set

$$\zeta = \varphi'\left((S+K)/2\right), \quad \beta = \varphi\left((S+K)/2\right) - \zeta\,(S+K)/2.$$

We note that when $\beta = 0$, $\Omega(\lambda^2(T-t), S(t))/\sqrt{T-t}$ in Proposition 7.5.1 conveniently becomes the time t implied Black volatility σ_{B} discussed earlier. For a few selected φ, Figure 7.1 below compares σ_{B} computed from the expansion in Proposition 7.5.1 (with $\beta = 0$) against exact results. Despite the long option maturity used in the figure, precision of the expansion is excellent, especially for the CEV case.

Fig. 7.1. Implied Volatility



Notes: The graph shows the implied volatility for a 10 year option, as a function of option moneyness $K/S(0)$. The initial value of the underlying is $S(0) = 6\%$. Three DVF models are considered in the figure. Case I: $\varphi(x) = x^{0.1}$, $\lambda = 1.59\%$. Case II: $\varphi(x) = x^{0.5}$, $\lambda = 4.90\%$. Case III: $\varphi(x) = x(1+30e^{-10x})$, $\lambda = 16.75\%$. The "Expansion" numbers in the graph were computed from the result in Proposition 7.5.1 with $\beta = 0$. For Case I and Case II, the "Exact" numbers were computed by the known CEV pricing formula in Proposition 7.2.6; for Case III the "Exact" numbers were computed in a Crank-Nicholson finite difference grid with 150 time steps and 250 spatial steps.

7.5.2 Expansion around Gaussian Process

For cases where φ is close to a constant, one might like to base the asymptotic expansion on $\varphi(x) = \beta$, for some constant β. In this case (which violates

one of the restrictions in Proposition 7.5.1), we use the Gaussian formula (7.16), and write

$$g(\tau, S) = (S - K)\,\Phi(w) - \Psi(\tau, S)\phi(w), \quad w = \frac{S - K}{\Psi(\tau, S)}. \tag{7.48}$$

For completeness, an asymptotic expansion of $\Psi(\tau, S)$ is given below.

Proposition 7.5.3. *An asymptotic expansion for the solution of (7.35) is given by (7.48), with*

$$\begin{aligned}
\Psi(\tau, S) &= \Psi_0(S)\tau^{1/2} + \Psi_1(S)\tau^{3/2} + O\left(\tau^{5/2}\right), \\
\Psi_0(S) &= (S - K)\left(\int_K^S \varphi(u)^{-1}\,du\right)^{-1}, \\
\Psi_1(S) &= -\frac{\Psi_0(S)^3}{(S - K)^2} \ln\left(\Psi_0(S)\left(\varphi(S)\varphi(K)\right)^{-1/2}\right).
\end{aligned}$$

In Proposition 7.5.3, the limit $S \to K$ leads to the following expressions

$$\begin{aligned}
\Psi_0(K) &= \varphi(K), \\
\Psi_1(K) &= \frac{1}{24}\varphi(K)\left(2\varphi(K)\varphi''(K) - \varphi'(K)^2\right).
\end{aligned}$$

The proof of Proposition 7.5.3 is similar to that of Proposition 7.5.1 and is omitted. Note that $\Psi(\lambda^2(T - t), S(t))/\sqrt{T - t}$ can be interpreted as an *implied Normal volatility.*

7.6 Extensions to Time-Dependent φ

So far, we have limited our discussion to the case where the function φ is independent of calendar time t. While there is some danger in making φ a function of t — the model inevitably becomes less time-stationary — there are a number of applications where such an extension is necessary to improve the fit to market data. Unlike the non-parametric approaches in Dupire [1994], Derman and Kani [1994], and Andersen and Brotherton-Ratcliffe [1998] (and many others) where $\varphi(t, S)$ is calibrated to fit a double continuum of call option prices, the applications we have in mind are normally parametric, and are inspired by typical requirements of calibrating term structure models to swaptions and caplets.

By itself, swaption and caplet pricing does not require time-dependent parameters, as only the terminal distribution is relevant. From that point of view, vanilla models with time-dependent local volatility functions may appear to have limited use in fixed income modeling. However, they often arise as describing approximate dynamics of swap or Libor rates in term

structure models. Many examples of such approximations are given in later chapters (see Chapters 13 and 14, for instance), and handling time-dependent parameters in local volatility models is important for term structure model calibration.

7.6.1 Separable Case

Recall the basic SDE (7.1). Its simplest time-dependent extension specifies a time-dependent scaling volatility λ, $\lambda = \lambda(t)$:

$$dS(t) = \lambda(t)\varphi\left(S(t)\right)\, dW(t). \tag{7.49}$$

This is the so-called *separable case*, as the local volatility function is represented as a product of two functions: $\lambda(\cdot)$, a function of the time variable only, and $\varphi(\cdot)$, a function of the state variable only. The separable form allows for application of the following simple time change argument:

Proposition 7.6.1. *Define*

$$\tau(t) = \int_0^t \lambda(u)^2\, du,$$

and define $s(\cdot)$ *by* $S(t) = s(\tau(t))$, *with* $S(t)$ *following (7.49). Then*

$$ds\left(\tau\right) = \varphi\left(s\left(\tau\right)\right)\, d\widetilde{W}(\tau), \quad s(0) = S(0), \tag{7.50}$$

where $\widetilde{W}$ *is a Brownian motion.*

Proof. The result follows directly from standard results for time-changed Brownian motion, see e.g. Karatzas and Shreve [1997]. □

Consider now the valuation of

$$c\left(t, S(t); T, K\right) = \mathrm{E}_t\left(\left(S(T) - K\right)^+\right),$$

which in the notation of Proposition 7.6.1 can be written as

$$c\left(t, s(\tau(t)); T, K\right) = \mathrm{E}\left(\left(s\left(\tau(T)\right) - K\right)^+ \middle| \widetilde{\mathcal{F}}_{\tau(t)}\right),$$

where $\widetilde{\mathcal{F}}$ is the filtration generated by $\widetilde{W}$. As the process for $s(\cdot)$ in (7.50) is of the type (7.1) (with $\lambda = 1$), all results from previous sections hold unchanged after the simple substitutions $\lambda \mapsto 1$ and $(T-t) \mapsto (\tau(T) - \tau(t))$. Equivalently, whenever the European option price results for constant λ involve terms of the form $\lambda^2(T-t)$, they should be replaced with $\int_t^T \lambda(u)^2\, du$ to accommodate a time-varying $\lambda(\cdot)$.

7.6.2 Skew Averaging

While the separable case can be handled quite easily, it is often too restrictive to be truly useful. Consider therefore the general case

$$dS(t) = \varphi(t, S(t))\, dW(t), \tag{7.51}$$

for $\varphi(t,x)$ satisfying the standard regularity conditions. European options could be valued in this model by PDE methods without much difficulty. However, with calibration applications in mind, this may be too slow or insufficiently accurate.

In this section, we develop European option approximations based on the idea of *time averaging*. Given the SDE (7.51), we look for a model with a *time-independent* local volatility function that yields European option prices approximately matching prices from the time-dependent model. The time-independent local volatility function can then be interpreted as a time average of the time-dependent function. This reduces the problem to one we know how to solve.

We have already seen a flavor of the averaging results that we are looking for. As demonstrated in Section 7.6.1, the values of T-expiry European options in the model

$$dS(t) = \lambda(t)\varphi(S(t))\, dW(t) \tag{7.52}$$

are the same as in the model

$$dS(t) = \overline{\lambda}\varphi(S(t))\, dW(t),$$

where $\overline{\lambda}$ is given by

$$\overline{\lambda}^2 = \int_0^T \lambda(u)^2\, du.$$

Thus, $\overline{\lambda}$ is an *effective volatility* for expiry T for the model (7.52).

Given the comments on U-shaped local volatility functions in Section 7.1.3, our initial focus shall be on functions that are monotonic in the state variable (see Section 7.6.3 for extensions). Such functions are typically well-described by two parameters, with the first parameter governing the overall level of volatility and the second the slope of the volatility smile (or skew). In the general case, both parameters are time-dependent. Let us concentrate on finding the averaging result for the time-dependent skew or, equivalently, on finding the *effective skew* formula.

We apply asymptotic expansion techniques with the *slope of the local volatility function* being the small parameter. Let us denote

$$X_0 = S(0), \quad \lambda(t) = \varphi(t, X_0), \quad g(t,x) = \frac{\varphi(t,x)}{\varphi(t,X_0)}.$$

Then (7.51) can be rewritten as

$$dS(t) = \lambda(t) g\left(t, S(t)\right) \, dW(t), \tag{7.53}$$

where

$$g\left(t, X_0\right) = 1. \tag{7.54}$$

Let us fix a time horizon $T > 0$ and attempt to derive conditions that a time-independent function $\overline{g}(x)$ needs to satisfy so that the SDE (7.53) can be replaced with

$$dS(t) = \lambda(t) \overline{g}\left(S(t)\right) \, dW(t) \tag{7.55}$$

for the purposes of valuing T-expiry European options of all strikes. Without loss of generality, the function $\overline{g}(x)$ is assumed to satisfy

$$\overline{g}(X_0) = 1.$$

Choose $\epsilon \geq 0$, the small slope parameter, and define

$$g^{\epsilon}\left(t, x\right) = g\left(t, X_0 + (x - X_0)\,\epsilon\right), \quad \overline{g}^{\epsilon}(x) = \overline{g}\left(X_0 + (x - X_0)\,\epsilon\right).$$

Next, define two sets of processes

$$\begin{aligned} dX^{\epsilon}(t) &= \lambda(t) g^{\epsilon}\left(t, X^{\epsilon}(t)\right) \, dW(t), \quad X^{\epsilon}(0) = X_0, \\ dY^{\epsilon}(t) &= \lambda(t) \overline{g}^{\epsilon}\left(Y^{\epsilon}(t)\right) \, dW(t), \quad Y^{\epsilon}(0) = X_0. \end{aligned}$$

The requirement that the prices of European options on $X^{\epsilon}(T)$ and $Y^{\epsilon}(T)$ across all strikes be close can be reformulated as the requirement that the distributions of $X^{\epsilon}(T)$ and $Y^{\epsilon}(T)$ be close. This can be formalized as finding $\overline{g}(\cdot)$ such that

$$q\left(\epsilon\right) \to \min,$$

where

$$q\left(\epsilon\right) \triangleq \mathrm{E}\left(\left(X^{\epsilon}(T) - Y^{\epsilon}(T)\right)^2\right). \tag{7.56}$$

Considering the small slope limit $\epsilon \to 0$, we expand $q(\epsilon)$ in powers of ϵ to obtain

$$q\left(\epsilon\right) = q(0) + q'(0)\epsilon + \frac{1}{2} q''(0)\epsilon^2 + O\left(\epsilon^3\right).$$

As part of the proof below we will show that $q(0) = q'(0) = 0$. Hence, the minimization problem simplifies to

$$q''(0) \to \min.$$

The (necessary) minimum condition is given in the following result.

Proposition 7.6.2. *Any function $\overline{g}$ that minimizes $q''(0)$ must satisfy the condition*

$$\frac{\partial \overline{g}(X_0)}{\partial x} = \int_0^T \frac{\partial g\left(t, X_0\right)}{\partial x} w_T(t) \, dt, \tag{7.57}$$

where

$$w_T(t) = \frac{v(t)^2 \lambda(t)^2}{\int_0^T v(t)^2 \lambda(t)^2 \, dt}, \quad v(t)^2 \triangleq \mathrm{E}\left(\left(X^0(t) - X_0\right)^2\right). \tag{7.58}$$

Proof. By Theorem 1.1.3, $q(\epsilon)$ as defined by (7.56) must equal

$$q(\epsilon) = \mathrm{E}\left(\int_0^T \left(g^\epsilon(t, X^\epsilon(t)) - \overline{g}^\epsilon(Y^\epsilon(t))\right)^2 \lambda(t)^2\, dt\right).$$

Differentiating with respect to ϵ, we get (omitting arguments on g^ϵ and $\overline{g}^\epsilon$ for brevity)

$$q'(\epsilon) = 2\mathrm{E}\left(\int_0^T (g^\epsilon - \overline{g}^\epsilon) \times \left(\frac{\partial}{\partial\epsilon} g^\epsilon - \frac{\partial}{\partial\epsilon}\overline{g}^\epsilon\right) \lambda(t)^2\, dt\right), \tag{7.59}$$

$$\begin{aligned} q''(\epsilon) = {} & 2\mathrm{E}\left(\int_0^T \left(\frac{\partial}{\partial\epsilon} g^\epsilon - \frac{\partial}{\partial\epsilon}\overline{g}^\epsilon\right)^2 \lambda(t)^2\, dt\right) \\ & + 2\mathrm{E}\left(\int_0^T (g^\epsilon - \overline{g}^\epsilon)\left(\frac{\partial^2}{\partial\epsilon^2} g^\epsilon - \frac{\partial^2}{\partial\epsilon^2}\overline{g}^\epsilon\right) \lambda(t)^2\, dt\right). \end{aligned}$$

Since $g^0(t,x) = \overline{g}^0(x) = 1$, it follows that $q(0)$, $q'(0)$ and the second term in the expression for $q''(0)$ are zero. Hence,

$$q''(0) = 2\mathrm{E}\left(\int_0^T \left(\frac{\partial}{\partial\epsilon} g^\epsilon(t, X^\epsilon(t))\Big|_{\epsilon=0} - \frac{\partial}{\partial\epsilon}\overline{g}^\epsilon(Y^\epsilon(t))\Big|_{\epsilon=0}\right)^2 \lambda(t)^2\, dt\right).$$

Note that

$$\begin{aligned} \frac{\partial}{\partial\epsilon} g^\epsilon(t, X^\epsilon(t)) &= \left[\epsilon\left(\frac{\partial X^\epsilon}{\partial\epsilon}\right)(t) + (X^\epsilon(t) - X_0)\right] \\ &\quad \times \frac{\partial g}{\partial x}(t, X_0 + \epsilon(X^\epsilon(t) - X_0)), \\ \frac{\partial}{\partial\epsilon}\overline{g}^\epsilon(Y^\epsilon(t)) &= \left[\epsilon\left(\frac{\partial Y^\epsilon}{\partial\epsilon}\right)(t) + (Y^\epsilon(t) - X_0)\right] \\ &\quad \times \frac{\partial \overline{g}}{\partial x}(X_0 + \epsilon(Y^\epsilon(t) - X_0)). \end{aligned}$$

In particular, as $Y^0(t) = X^0(t)$,

$$\begin{aligned} \frac{\partial}{\partial\epsilon} g^\epsilon(t, X^\epsilon(t))\Big|_{\epsilon=0} &= (X^0(t) - X_0)\frac{\partial g}{\partial x}(t, X_0), \\ \frac{\partial}{\partial\epsilon}\overline{g}^\epsilon(Y^\epsilon(t))\Big|_{\epsilon=0} &= (X^0(t) - X_0)\frac{\partial \overline{g}}{\partial x}(X_0). \end{aligned}$$

Thus,

$$q''(0) = 2\int_0^T \mathrm{E}\left(\left(X^0(t) - X_0\right)^2 \lambda(t)^2\right)\left(\frac{\partial g}{\partial x}(t, X_0) - \frac{\partial \overline{g}}{\partial x}(X_0)\right)^2 dt$$
$$= 2\int_0^T v(t)^2\lambda(t)^2\left(\frac{\partial g}{\partial x}(t, X_0) - \frac{\partial \overline{g}}{\partial x}(X_0)\right)^2 dt,$$

with $v(t)^2$ defined in (7.58). Differentiating with respect to the slope $\partial\overline{g}(X_0)/\partial x$ and setting the resulting derivative to zero, we obtain a condition for the minimum of $q''(0)$. This gives (7.57). □

It follows from the proposition that for the purposes of (approximately) pricing T-expiry European options, (7.53) can be replaced with (7.55), where $\overline{g}(\cdot)$ is a function whose slope (skew) at-the-money, $\partial\overline{g}(S(0))/\partial x$, is a weighted average of the time-dependent at-the-money slopes (skews) of the original function $\partial g(t, S(0))/\partial x$, $t \in [0, T]$. The weights $w_T(t)$ to be used in forming the slope-average are the weights $w_T(t)$ in (7.58). Once the SDE of the form (7.53) has been approximated with (7.55), various tools developed in the first part of the chapter become available, and European option prices can be computed efficiently.

7.6.2.1 Examples

The time-dependent local volatility function $g(t, x)$ is often defined to be a time-indexed collection of functions from the same family. Examples include the time-dependent displaced log-normal function

$$g(t, x) = b(t)\frac{x}{S(0)} + (1 - b(t)), \quad t \in [0, T], \tag{7.60}$$

or the time-dependent CEV function

$$g(t, x) = \left(\frac{x}{S(0)}\right)^{p(t)}, \quad t \in [0, T]. \tag{7.61}$$

Note that the functions in the formulas have been scaled to satisfy (7.54). The condition (7.57) does not define the function $\overline{g}$ uniquely. To improve the accuracy of the approximation, it is often beneficial to choose $\overline{g}$ from the same family as the functions they approximate. In particular, for g of the type (7.60), the function $\overline{g}$ is best chosen to be of the same displaced log-normal type

$$\overline{g}(x) = \overline{b}\frac{x}{S(0)} + \left(1 - \overline{b}\right). \tag{7.62}$$

In the same vein, for the CEV case (7.61), a natural choice for $\overline{g}$ is

$$\overline{g}(x) = \left(\frac{x}{S(0)}\right)^{\overline{p}}. \tag{7.63}$$

Both the displaced log-normal parameter b and the CEV parameter p are used as a measure of the skew in the implied volatility smile. The next

corollary expressed the averaging result directly in terms of these parameters, and also explicitly derives the averaging weights.

Corollary 7.6.3. *Over the time-horizon $[0,T]$, the effective skew $\overline{b}$ in (7.62) for the model defined by the time-dependent local volatility function (7.60) is given by*

$$\overline{b} = \int_0^T b(t) w_T(t)\, dt, \tag{7.64}$$

where

$$w_T(t) = \frac{v(t)^2 \lambda(t)^2}{\int_0^T v(t)^2 \lambda(t)^2\, dt}, \quad v(t)^2 = \int_0^t \lambda(s)^2\, ds. \tag{7.65}$$

Proof. For $g(t,x)$ and $\overline{g}(x)$ given by (7.60) and (7.62), we have

$$\frac{\partial g}{\partial x}(t, S(0)) = \frac{b(t)}{S(0)}, \quad \frac{\partial \overline{g}}{\partial x}(S(0)) = \frac{\overline{b}}{S(0)}.$$

Thus, (7.64) follows from (7.57). The formula (7.65), and in particular the expression for $v(t)^2$, follows from the definition

$$v(t)^2 = \mathrm{E}\left(\left(X^0(t) - X_0\right)^2\right)$$

and the fact that $X^0(t)$ satisfies

$$dX^0(t) = \lambda(t) g^0\left(t, X^0(t)\right)\, dW(t)$$

with

$$g^0\left(t, X^0(t)\right) \equiv 1.$$

□

Remark 7.6.4. An identical result holds for the effective CEV parameter $\overline{p}$,

$$\overline{p} = \int_0^T p(t) w_T(t)\, dt,$$

where $p(\cdot)$ and $\overline{p}$ are the parameters in (7.61) and (7.63), and $w_T(t)$ is as given in (7.65).

Example 7.6.5. Assuming constant volatility $\lambda(t) \equiv \lambda$, we obtain particularly simple formulas for the effective skew,

$$v(t)^2 = \lambda^2 t, \quad w_T(t) = \frac{t}{\int_0^T t\, dt} = \frac{t}{T^2/2},$$

so that

$$\overline{b} = \frac{1}{T^2/2} \int_0^T t b(t)\, dt.$$

This demonstrates that instantaneous skews $b(t)$ for larger t contribute more to $\overline{b}$ than those for lower t. Intuitively, the process needs to build up its variance before the changes in the instantaneous slopes start having an effect on the effective slope of the local volatility.

7.6.2.2 A Caveat About the Process Domain

Even though the skew averaging result is obtained in the small slope limit, practical experience validates its broad applicability in option pricing problems. Some typical results can be found in Piterbarg [2005c] and Piterbarg [2006]. Still, the equivalence between the original time-dependent model and the time-averaged one should not be taken too far, as we now proceed to demonstrate. For this, we focus on the simple displaced diffusion model from the previous section, i.e. we consider the time-dependent SDE

$$dS(t) = \lambda \left(b(t)S(t) + (1 - b(t))\, S(0)\right)\, dW(t), \tag{7.66}$$

and approximate it with

$$dS(t) = \lambda \left(\bar{b}S(t) + \left(1 - \bar{b}\right)\, S(0)\right)\, dW(t), \tag{7.67}$$

where $\bar{b}$ is set as in Corollary 7.6.3. While the two SDEs (7.66) and (7.67) may have similar properties in the neighborhood of $S(0)$, they generally do not even have the same range for $S(t)$. For the constant parameter case (7.67) with $\bar{b} > 0$, the process $S(t)$ has a lower bound, the root of the local volatility function: $S(t) \in (S(0)(\bar{b} - 1)/\bar{b}, \infty)$. The same is not necessarily true for the time-dependent SDE (7.66), as should be reasonably clear from the following heuristic argument. If at a given time t, $S(t)$ is close to the root of the local volatility function but still above it, i.e.

$$S(t) \gtrsim S(0)\,(b(t) - 1)\,/b(t),$$

it may so happen that at $t + dt$, $S(t + dt)$ is actually *below* the root of the local volatility function,

$$S\,(t + dt) < S(0)\,(b\,(t + dt) - 1)\,/b\,(t + dt)$$

due to the change in the function $b(\cdot)$. The range

$$(-\infty, S(0)\,(b\,(t + dt) - 1)\,/b\,(t + dt))$$

will then be reachable by $S(\cdot)$. The following proposition provides formal justification.

Proposition 7.6.6. *Consider the SDE*

$$dX(t) = (a(t) + b(t)X(t))\, dW(t) \tag{7.68}$$

with $X(0) \geq a(0)$. *If* $a'(u) \leq 0$ *for all* $u \in [0, t]$, *then* $X(t) > a(t)$ *a.s. If there exists* u, $0 < u < t$, *such that* $a'(u) > 0$, *then* $\mathrm{P}(X(t) < l) > 0$ *for any* $l \in \mathbb{R}$.

Proof. Define

$$\zeta(t) = \int_0^t b(u)\, dW(u) - \frac{1}{2}\int_0^t b^2(u)\, du, \quad Z(t) = \exp\left(\zeta(t)\right).$$

Then the solution to the SDE (7.68) is given by

$$X(t) = Z(t)\left[X(0) + \int_0^t a(u)\, d\left(1/Z(u)\right)\right],$$

as can either be checked directly or obtained from Section 5.6.C of Karatzas and Shreve [1997]. Integrating by parts yields

$$X(t) = Z(t)\left(X(0) - a(0)\right) + a(t) - Z(t)\int_0^t \frac{a'(u)}{Z(u)}\, du.$$

With $X(0) \geq a(0)$,

$$Z(t)\left(X(0) - a(0)\right) + a(t)$$

is bounded from below by $a(t)$. If $a'(u) \leq 0$ for all $u \in [0,t]$ then the remaining term

$$-Z(t)\int_0^t \frac{a'(u)}{Z(u)}\, du$$

is positive and $X(t)$ is bounded from below by $a(t)$. If, however, there exists u such that $a'(u) > 0$, this term can be arbitrarily negative with positive probability. □

In practice, the likelihood of actually breaching the lower boundary is typically small and we can often safely ignore this possibility. If needed, one can always "regularize" the time-dependent process to limit its range, along the same lines as done in Section 7.2.3.

7.6.3 Skew and Convexity Averaging by Small-Noise Expansion

The technique used in the previous section to derive Proposition 7.6.2 is not the only route to go. An alternative approach relies on *small-noise expansion*, a concept closely related to the Ito-Taylor expansion in Chapter 3. To illustrate the versatility of this method, we shall use it to derive not only the skew averaging result in Corollary 7.6.3, but also to demonstrate how to compute *average convexity* in a time-dependent quadratic model.

As our starting point, we define, for some constant X_0, the quadratic form

$$\begin{aligned}\varphi\left(t, X(t)\right) &= \varphi\left(b(t), c(t), X(t)\right) \\ &= \left(1 - b(t)\right) X_0 + b(t)X(t) + \frac{1}{2}c(t)\left(X(t) - X_0\right)^2,\end{aligned}$$

and then introduce the following two processes:

$$dX(t) = \lambda(t)\varphi\left(b(t), c(t), X(t)\right)\, dW(t), \quad X(0) = X_0, \tag{7.69}$$

$$dY(t) = \lambda(t)\varphi\left(\overline{b}, \overline{c}, Y(t)\right)\, dW(t), \quad Y(0) = X_0, \tag{7.70}$$

where $W(t)$ is a Brownian motion in some probability measure. We can characterize the process for $X(t)$ as having quadratic local volatility with time-dependent slope $b(t)$ and time-dependent convexity $c(t)$; for a fixed value of T, we are interested in establishing how to set the constants $\overline{b}$ in $\overline{c}$ in the process for Y such that $Y(T)$ is a good approximation to $X(T)$.

We will answer the question posed above in the small-noise limit. For that, set

$$dX^\epsilon(t) = \epsilon\lambda(t)\varphi\left(b(t), c(t), X^\epsilon(t)\right)\, dW(t), \tag{7.71}$$

$$dY^\epsilon(t) = \epsilon\lambda(t)\varphi\left(\overline{b}, \overline{c}, Y^\epsilon(t)\right)\, dW(t), \tag{7.72}$$

with $Y^\epsilon(0) = X^\epsilon(0) = X_0$. Notice that $X^1(t) = X(t)$ and $Y^1(t) = Y(t)$, and that $X^0(t) = Y^0(t) = X_0$.

Lemma 7.6.7. *For the SDE (7.71), we have the formal expansion*

$$X^\epsilon(T) = X_0 + \epsilon A_X(T) + \frac{1}{2}\epsilon^2 B_X(T) + \frac{1}{6}\epsilon^3 C_X(T) + O\left(\epsilon^4\right),$$

where

$$A_X(T) = X_0 \int_0^T \lambda(t)\, dW(t),$$

$$B_X(T) = 2\int_0^T \lambda(t)b(t)A_X(t)\, dW(t),$$

$$C_X(t) = 3\int_0^T \lambda(t)c(t)A_X(t)^2\, dW(t) + 3\int_0^T \lambda(t)b(t)B_X(t)\, dW(t).$$

Proof. We rely on standard asymptotic expansion techniques (e.g. Yoshida [1992]) to construct a Taylor series of $X^\epsilon(T)$ around $\epsilon = 0$. Dropping the arguments of $\varphi(t) = \varphi(b(t), c(t), X^\epsilon(t))$ for brevity, we get

$$\begin{aligned} A_X(T) &= \left.\frac{\partial X^\epsilon(T)}{\partial\epsilon}\right|_{\epsilon=0} \\ &= \left(\int_0^T \lambda(t)\varphi(t)dW(t) + \epsilon\int_0^T \lambda(t)\frac{\partial\varphi(t)}{\partial X^\epsilon(t)}\frac{\partial X^\epsilon(t)}{\partial\epsilon}dW(t)\right)\Bigg|_{\epsilon=0} \\ &= \int_0^T \lambda(t)\varphi\left(b(t), c(t), X_0\right) dW(t) = X_0\int_0^T \lambda(t)dW(t). \end{aligned}$$

Similarly,

$$
\begin{aligned}
B_X(T) &= \left.\frac{\partial^2 X^\epsilon(T)}{\partial \epsilon^2}\right|_{\epsilon=0} \\
&= \left(\int_0^T \lambda(t) \frac{\partial \varphi(t)}{\partial X^\epsilon(t)} \frac{\partial X^\epsilon(t)}{\partial \epsilon} dW(t) + \int_0^T \lambda(t) \frac{\partial \varphi(t)}{\partial X^\epsilon(t)} \frac{\partial X^\epsilon(t)}{\partial \epsilon} dW(t) \right) \Bigg|_{\epsilon=0} \\
&\quad + \epsilon \int_0^T \lambda(t) \frac{\partial^2 \varphi(t)}{\partial X^\epsilon(t)^2} \left(\frac{\partial X^\epsilon(t)}{\partial \epsilon} \right)^2 dW(t) \Bigg|_{\epsilon=0} \\
&\quad + \epsilon \int_0^T \lambda(t) \frac{\partial \varphi(t)}{\partial X^\epsilon(t)} \frac{\partial^2 X^\epsilon(t)}{\partial \epsilon^2} dW(t) \Bigg|_{\epsilon=0} \\
&= 2 \int_0^T \lambda(t) b(t) A_X(t)\, dW(t),
\end{aligned}
$$

where we have used the fact that $\partial \varphi(t)/\partial X^\epsilon(t) = b(t)$ when $X^\epsilon(t) = X_0$. The result for $C_X(T)$ follows in the same fashion. □

For the variable Y^ϵ in (7.72), we get

$$
Y^\epsilon(T) = X_0 + \epsilon A_X(T) + \frac{1}{2}\epsilon^2 B_Y(T) + \frac{1}{6}\epsilon^3 C_Y(T) + O\left(\epsilon^4\right),
$$

where B_Y and C_Y are found by substituting $\bar{b}$ for $b(t)$ and $\bar{c}$ for $c(t)$ in the expressions for B_X and C_X in Lemma 7.6.7. We therefore immediately have the following result.

Lemma 7.6.8. *Consider the ϵ-indexed processes (7.71)–(7.72). Then, for $T > 0$,*

$$
X^\epsilon(T) - Y^\epsilon(T) = \epsilon^2 I_1(\bar{b}; T) + \epsilon^3 I_2(\bar{b}, \bar{c}; T) + O(\epsilon^4),
$$

where we have defined zero-mean random variables

$$
\begin{aligned}
I_1(\bar{b}; T) &= \int_0^T \lambda(t) \left(b(t) - \bar{b}\right) A_X(t)\, dW(t), \\
I_2(\bar{b}, \bar{c}; T) &= \frac{1}{2} \int_0^T \lambda(t) \left(c(t) - \bar{c}\right) A_X(t)^2\, dW(t) \\
&\quad + \frac{1}{2} \int_0^T \lambda(t) b(t) B_X(t)\, dW(t) - \frac{1}{2}\bar{b} \int_0^T \lambda(t) B_Y(t)\, dW(t).
\end{aligned}
$$

There are numerous ways in which we can use the results of the previous section to determine the values of $\bar{b}$ and $\bar{c}$ that will make $Y^\epsilon(T)$ best approximate $X^\epsilon(T)$. Starting with $\bar{b}$, we here elect to set it such that the variance of the $O(\epsilon^3)$ term (the "skew term") in Lemma 7.6.8 is minimized. That is, our optimal choice $\bar{b}^*$ for $\bar{b}$ is characterized by

$$
\bar{b}^* = \operatorname*{argmin}_{\bar{b}} \mathrm{E}\left(I_1\left(\bar{b}; T\right)^2\right). \tag{7.73}
$$

Proposition 7.6.9. *The solution to (7.73) is*

$$\bar{b}^* = \int_0^T b(t) w_T(t) dt, \quad w_T(t) = \frac{\lambda(t)^2 v(t)^2}{\int_0^T \lambda(t)^2 v(t)^2 dt},$$

where $v(\cdot)^2$ is defined in (7.65).

Proof. First, we need to establish the expectation of the random variable $I_1(\bar{b}; T)^2$. From elementary properties of the Ito integral (see Theorem 1.1.3), we know that

$$\begin{aligned} \mathrm{E}\left(I_1\left(\bar{b}; T\right)^2\right) &= \mathrm{E}\left(\left(\int_0^T \lambda(t)\left(b(t) - \bar{b}\right) A_X(t)\, dW(t)\right)^2\right) \\ &= \int_0^T \lambda(t)^2 \left(b(t) - \bar{b}\right)^2 \mathrm{E}\left(A_X(t)^2\right)\, dt. \end{aligned}$$

Since $A_X(t)$ is a Gaussian random variable with mean 0 and variance $X_0^2 v(t)^2$, it follows that

$$\mathrm{E}\left(I_1\left(\bar{b}; T\right)^2\right) = X_0^2 \int_0^T \lambda(t)^2 \left(b(t) - \bar{b}\right)^2 v(t)^2\, dt.$$

The (necessary) condition for a minimum is

$$\frac{1}{X_0^2} \frac{\partial \mathrm{E}\left(I_1\left(\bar{b}; T\right)^2\right)}{\partial \bar{b}} = 2\bar{b} \int_0^T \lambda(t)^2 v(t)^2\, dt - 2 \int_0^T \lambda(t)^2 b(t) v(t)^2\, dt = 0,$$

from which the result in Proposition 7.6.9 follows. □

As advertised, the result of Proposition 7.6.9 is identical to that of Corollary 7.6.3.

It remains to find $\bar{c}$. We fundamentally wish to fix it such that the variance of the $O(\epsilon^4)$ term (the "convexity term") in Lemma 7.6.8 is minimized, given $\bar{b} = \bar{b}^*$. When $\bar{b} = \bar{b}^*$, however, we can observe that

$$I_2(\bar{b}^*, \bar{c}; T)^2 \approx \frac{1}{2} \int_0^T \lambda(t)\left(c(t) - \bar{c}\right) A_X(t)^2\, dW(t),$$

which suggests the simplified condition[8]

$$\bar{c}^* = \underset{\bar{c}}{\operatorname{argmin}}\, \mathrm{E}\left(\left(\frac{1}{2} \int_0^T \lambda(t)\left(c(t) - \bar{c}\right) A_X(t)^2\, dW(t)\right)^2\right). \tag{7.74}$$

[8] More rigorous results can be found in Andersen and Hutchings [2010], but the accuracy of (7.74) is typically sufficient for applications.

Proposition 7.6.10. *The value $\overline{c}^*$ that satisfies (7.74) is*

$$\overline{c}^* = \int_0^T c(t) q_T(t)\, dt, \quad q_T(t) \triangleq \frac{\lambda(t)^2 v(t)^4}{\int_0^T \lambda(t)^2 v(t)^4 dt},$$

where $v(\cdot)^2$ is defined in (7.65).

Proof. We note that

$$\mathrm{E}\left(\left(\frac{1}{2}\int_0^T \lambda(t)\left(c(t)-\overline{c}\right) A_X(t)^2\, dW(t)\right)^2\right)$$
$$= \frac{1}{4}\int_0^T \lambda(t)^2 \left(c(t)-\overline{c}\right)^2 \mathrm{E}\left(A_X(t)^4\right) dt.$$

From a standard property of Gaussian random variables, we have

$$\mathrm{E}\left(A_X(t)^4\right) = 3X_0^4 v(t)^4.$$

Applying this result, the (necessary) condition for a minimum is

$$\frac{1}{2}\overline{c}\int_0^T \lambda(t)^2 v(t)^4\, dt - \frac{1}{2}\int_0^T \lambda(t)^2 c(t) v(t)^4\, dt = 0.$$

The Proposition 7.6.10 follows. □

Remark 7.6.11. For the special case where λ is constant, we have $v(t)^2 = \lambda^2 t$ and therefore

$$\overline{b}^* = \frac{2\int_0^T b(t) t dt}{T^2}, \quad \overline{c}^* = \frac{3\int_0^T c(t) t^2 dt}{T^3}.$$

Note that the contribution of the instantaneous convexity $c(t)$ to the effective local volatility convexity grows with t at a faster rate ($O(t^2)$) than the contribution of $b(t)$ to the effective local volatility skew ($O(t)$).

7.6.4 Numerical Example

A brief numerical example is now in order. To provide a simple setup in which we can test our averaging results, we consider a two-period case where

$$\lambda(t) = \begin{cases} \lambda_0, & t \in [0, T'], \\ \lambda', & t \in (T', T], \end{cases} \quad b(t) = \begin{cases} b_0, & t \in [0, T'], \\ b', & t \in (T', T], \end{cases} \quad c(t) = \begin{cases} 0, & t \in [0, T'], \\ c', & t \in (T', T]. \end{cases} \tag{7.75}$$

The advantage of this setup is that it allows for high precision call option pricing without the need for finite difference grids or Monte Carlo methods. In particular, by having $c(t) = 0$ for $t \in [0, T']$, it follows that

$$dX(t) = (b_0 X(t) + (1 - b_0)X_0)\,\lambda_0\,dW(t), \quad t \in [0, T']$$

such that, from the fact that these dynamics are those of a simple displaced log-normal process,

$$X(T') = \begin{cases} b_0^{-1}\left(X_0 \exp\left(-\frac{1}{2}b_0^2\lambda_0^2 T' + b_0\lambda_0 W(T')\right) - (1-b_0)X_0\right), & b_0 \neq 0 \\ X_0 + X_0\lambda_0 W(T'), & b_0 = 0. \end{cases} \tag{7.76}$$

Let[9] $C(t, x; K, T)$ be the time t price of a K-strike, T-maturity call option when $X(t) = x$. Clearly (assuming zero interest rates)

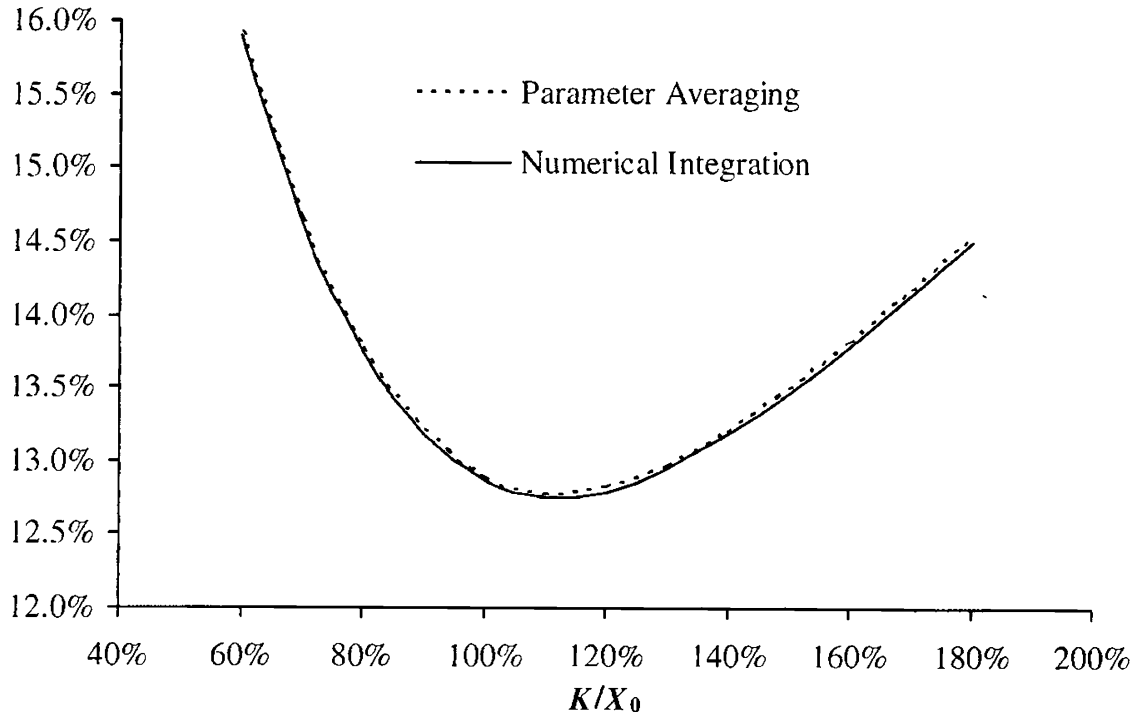
$$C(0, X_0; K, T) = \mathrm{E}\left(C(T', X(T'); K, T)\right).$$

At time T', process parameters switch to constant values λ', c', b' so for any value of $X(T')$ computation of $C(T', X(T'); K, T)$ can be done using the formulas for call options in the quadratic model (see Section 7.3 and Andersen [2010]). From (7.76), computation of $C(0, X_0; K, T)$ can then easily be performed by numerical integration. Figure 7.2 below shows a sample fit for a high-convexity case ($X_0 = 1, c' = 4$).

The constant-parameter approximation here does an excellent job of matching the volatility smile of the true model. For even higher precision — especially for the (rare) case where convexity is very large and rapidly changing in time — additional correction terms may be required; see Andersen and Hutchings [2010] for the details and more numerical tests.

[9] We temporarily use notation C (rather than the usual c) for a call option, to distinguish it from the convexity function $c(t)$.

Fig. 7.2. Implied Volatility Smile



Notes: Parameters are as in (7.75), with $T' = 1$, $T = 2$, $\lambda_0 = 10\%$, $\lambda' = 15\%$, $b_0 = 0$, b'=0.75, $c' = 4$. The x-axis denotes relative strike K/X_0, with $X_0 = 1$. The "Numerical Integration" graph is the 2 year implied volatility smile for the time-dependent model, computed as outlined in the text (100 integration nodes). The "Parameter Averaging" graph computes the 2 year volatility smile from a constant-parameter quadratic model with parameters set as in Propositions 7.6.9 and 7.6.10.

8

Vanilla Models with Stochastic Volatility I

In Chapter 7 we introduced and studied diffusive single-factor vanilla models where the volatility is a deterministic function of the underlying rate. While such level-dependence of volatility is observable in interest rate markets — implied Black volatility is normally higher when rates are low — there is strong empirical evidence for additional sources of randomness in interest rate volatilities. To make our model setup more realistic, and to improve our ability to fit models to market-implied volatility smiles, we continue our investigation of vanilla models by enlarging the DVF models from the previous chapter to allow the volatility to be driven by a separate Brownian motion. The resulting models are said to have *stochastic volatility.*

Beyond raising the dimension of our models dynamics from one to two, the introduction of stochastic volatility brings with it a number of technical complications and, for many important models, the need to work with Fourier transforms when pricing options. We discuss these issues in detail in this chapter, paying particular attention to the *displaced log-normal Heston model* which has good analytical tractability and often provides an excellent fit to market observations.

Stochastic volatility constitutes a large and important topic in contemporary fixed income modeling, and we shall need two chapters of this book to lay the proper foundation for later work. In this chapter, our focus is on basic material and on the development of Fourier integration methods in a time-homogeneous setting. More advanced topics — including many numerical methods and the extension to time-dependent parameters — are postponed to Chapter 9.

8.1 Model Definition

As in Chapter 7, let $S(t)$, the "underlying" as we shall often call it, denote a forward Libor or swap rate. Also, let $Z(t)$, $W(t)$ be two different one-dimensional Brownian motions under a measure P in which $S(t)$ is a

martingale; we assume that $Z(t)$ and $W(t)$ are correlated with constant correlation ρ. As before, we use E instead of E^{P} for the expected value operator under measure P if there is no possibility of confusion. A fairly general family of stochastic volatility models[1] is obtained by specifying

$$dS(t) = \lambda\varphi\left(S(t)\right)\sqrt{z(t)}\,dW(t), \tag{8.1}$$

$$dz(t) = \theta\left(m(t) - z(t)\right)\,dt + \eta\psi\left(z(t)\right)\,dZ(t), \quad z(0) = z_0, \tag{8.2}$$

where λ, θ, η are positive constants, $m(\cdot)$ a positive deterministic function of time, and $\varphi(\cdot)$ and $\psi(\cdot)$ two smooth deterministic functions. In these SDEs, $z(\cdot)$ is a *stochastic variance* process, the square root of which scales a DVF diffusion term similar to that discussed in Chapter 7.

We notice that the drift term of $z(\cdot)$ is such that $z(t)$ gets pulled towards the level $m(t)$ at an exponential rate of θ, known as the *mean reversion speed* (or sometimes just the *mean reversion*). The parameter η is the *volatility of variance*,, and $\psi(z)$ is a skew function for the stochastic variance. We shall later discuss in more detail the roles and effects of the individual parameters in the dynamics for $z(t)$, but before doing so let us try to indicate what constitutes a reasonable model specification. First, since the effect of $z(\cdot)$ on the volatility of $S(\cdot)$ is multiplicative, the initial value z_0 and the value $m(t)$ to which it mean-reverts can be scaled to arbitrary level; for convenience[2] we typically set $m(t) \equiv z_0 = 1$. As for the functions $\psi(\cdot)$ and $\varphi(\cdot)$, there are many empirically reasonable choices, but convenience and efficiency of available valuation algorithms for European options need to be considered. Typically, the function $\psi(\cdot)$ is chosen to be the square root function, making the process for $z(t)$ *affine* and improving analytical tractability. That said, other power functions, nevertheless, can be used and sometimes may be preferred, for reasons explained later (see, e.g., the end of Section 8.3). Analytical tractability also suggests using a linear function for $\varphi(\cdot)$, such that the underlying DVF model is a standard displaced log-normal model, see Section 7.2.4.

It only remains to comment on the correlation parameter ρ. In interest rate applications, the correlation ρ between the Brownian motions driving the stochastic variance and the underlying is often set to 0 due to undesirable effects of common measure changes on the stochastic variance process when correlation is non-zero, see Proposition 8.3.9. This is rarely a limitation,

[1] For non-linear functions $\varphi(x)$ or $\varphi(t, x)$ such models are sometimes called *local stochastic volatility*, or LSV, models. Occasionally the name is also used for models with linear φ.

[2] Note that setting $m(\cdot)$ to a constant different from z_0 defines a model with constant coefficients that has a somewhat richer term volatility structure than with $m(\cdot) \equiv z_0$. The utility of this is limited as we are ultimately interested in time-dependent model extensions anyway.

as the effect of correlation on option prices and their implied volatilities[3] can typically be captured in parameters of the function $\varphi(\cdot)$. From the perspective of matching the implied volatility smile, non-zero correlation is thus largely superfluous. Provided that we define our hedge sensitivities in a certain, natural way, this observation also holds for hedging, a point we shall return to in Section 8.9.2. To keep our discussion general, we nevertheless keep correlation non-zero for much of the discussion that follows.

With the parameter specializations described above, the simplified model we shall concentrate most of our efforts on is defined as

$$dS(t) = \lambda \left(bS(t) + (1-b)\, L\right) \sqrt{z(t)}\, dW(t), \tag{8.3}$$

$$dz(t) = \theta \left(z_0 - z(t)\right)\, dt + \eta \sqrt{z(t)}\, dZ(t), \quad z(0) = z_0 = 1, \tag{8.4}$$

with $\langle dZ(t), dW(t) \rangle = \rho\, dt$. Going forward, this model will be referred to as simply the *SV model*. For the case where $b = 1$, the model becomes identical to the so-called *Heston model*; see Heston [1993]. To avoid degenerate situations, we make the following assumption:

Assumption 8.1.1. *All parameters b, θ, η, λ are strictly positive, and $|\rho| < 1$.*

8.2 Model Parameters

We proceed to a more detailed discussion of the parameters in the model (8.3)–(8.4). First, recall that in the local volatility model of the displaced log-normal type (7.21), the parameter λ is responsible for the overall level of the implied volatility smile, while the parameter b is responsible for its slope. This interpretation of the parameters carries over to the stochastic volatility case (8.3)–(8.4), and we often refer to λ and b as the *SV volatility* and the *skew*, respectively.

The volatility of variance parameter η controls the curvature of the volatility smile, see Section 8.7. The effect of η on the volatility smile is similar to that of the second-order, or convexity, term in a quadratic DVF model of Section 7.3, although the dynamics of the volatility smile are quite different in the two models, a point we shall return to later, in Section 8.8.

The mean reversion of variance, θ, controls the speed at which deviations of $z(\cdot)$ away from z_0 are pulled back towards this level. Increasing θ decreases the long-term variance of $z(\cdot)$ and limits the effect of the stochastic variance process on the volatility smile for medium- and long-dated maturities. In essence, θ controls the speed of decay of the volatility smile convexity.

[3] If the correlation is negative — i.e. if $z(t)$ tends to be high when $S(t)$ is low — the model will imply a downward-sloping volatility smile, as should be intuitively clear.

The local volatility function $\varphi(x) = bx + (1-b)L$ involves a quantity L, the *level* parameter. As discussed in the previous chapter, we normally set this to a number equal or close to[4] $S(0)$, to ensure that λ will have the dimension of implied Black volatility, irrespective of the setting of b. This decoupling of parameters is particularly convenient in a calibration context.

As in the (local volatility) displaced log-normal model, λ is expressed in the units of relative volatility, while the skew b is typically confined to a range between 0% to 100%, although the "super-Normal" ($b < 0$) and "super-log-normal" ($b > 1$) settings may occasionally be useful. For $b < 0$ or $b > 1$, our earlier discussion in Section 7.6.2.2 shows that if $L > 0$, the state space for $S(\cdot)$ is bounded (above or below depending on the sign of b) by the value $-L(1-b)/b$. The existence of such a bound is somewhat unrealistic; however, the advantages of being able to use values of b outside of $[0,1]$ usually outweigh this concern.

The parameter η is expressed in the units of annualized relative volatility of *variance*. Sometimes it is more natural to think in terms of the volatility of *volatility*, i.e. the volatility of the process for $\sqrt{z(t)}$. By Ito's lemma,

$$d\sqrt{z(t)} = O(dt) + \frac{\eta}{2} dZ(t).$$

When $z(t)$ has unit magnitude, $\eta/2$ can loosely be thought of as the volatility of volatility. For example, a value of 100% for η associates the implied Black volatility of the model with an instantaneous relative annualized volatility of about 50%. The related parameter θ, the speed of mean reversion, is expressed in percentage points per year. The inverse quantity θ^{-1} is measured in years and is related to the time over which a volatility shock dissipates. Specifically, the half-life of a volatility shock is

$$t_{1/2} = \frac{\ln 2}{\theta}.$$

All major interest rate markets exhibit high volatility of variance/low mean reversion of variance parameters, with $\eta = 150\%$ and $\theta = 10\%$ being typical parameter settings. While a half-life of a volatility shock of $10 \ln 2 \approx 7$ years may appears quite unrealistic, one should not forget that the pricing measure P will rarely represent real-world probabilities whereby the drift in the process for $z(\cdot)$ will likely contain strong market price of risk adjustment. The impact of measure changes on the speed of mean reversion for the variance is highlighted in Proposition 8.3.9.

[4] The rationale for not letting $L = S(0)$ always is that computation of delta sensitivities $\partial/\partial S(0)$ would then perturb the constant in the linear form $\varphi(x)$ which may or may not be desirable. See Sections 16.1.1 and 16.1.2 for more details.

8.3 Basic Properties

In this section we collect several important facts about the distribution and other relevant characteristics of $z(\cdot)$ and $S(\cdot)$ in the model (8.3)–(8.4). First, we look at the regularity properties of the process for the stochastic variance $z(\cdot)$; the results below should be compared to Proposition 7.2.1.

Proposition 8.3.1. *The SDE (8.4) has a unique solution. If* $2z_0\theta \geq \eta^2$, *i.e. the so-called* Feller condition *holds,* $z = 0$ *is unattainable. If the Feller condition is violated,* $2z_0\theta < \eta^2$, *then* $z = 0$ *is an attainable boundary but is strongly reflecting.*

Proof. See Revuz and Yor [1999] or Andersen and Piterbarg [2007]. □

The transition distribution for the variance process $z(\cdot)$ given by (8.4) was derived in Cox et al. [1985] is listed below.

Proposition 8.3.2. *Let* $\Upsilon(z;\nu,\gamma)$ *be the cumulative distribution function for the non-central chi-square distribution with* ν *degrees of freedom and non-centrality parameter* γ*:*

$$\Upsilon(z;\nu,\gamma) = e^{-\gamma/2}\sum_{j=0}^{\infty}\frac{(\gamma/2)^j}{j!2^{\nu/2+j}\Gamma(\nu/2+j)}\int_0^z y^{\nu/2+j-1}e^{-y/2}\,dy. \tag{8.5}$$

For the process (8.4) define

$$d = 4\theta z_0/\eta^2, \quad n(t,T) = \frac{4\theta e^{-\theta(T-t)}}{\eta^2\left(1-e^{-\theta(T-t)}\right)}, \quad T > t. \tag{8.6}$$

Let $T > t$. *Conditional on* $z(t)$, $z(T)$ *is distributed as* $e^{-\theta(T-t)}/n(t,T)$ *times a non-central chi-square distributed random variable with* d *degrees of freedom and non-centrality parameter* $z(t)n(t,T)$,

$$\mathrm{P}\left(z(T) < x \,|\, z(t)\right) = \Upsilon\left(\frac{x\cdot n(t,T)}{e^{-\theta(T-t)}}; d, z(t)n(t,T)\right).$$

Of particular importance, especially in Monte Carlo methods discussed later in Section 9.5, are the conditional moments of $z(\cdot)$. From the known properties of the non-central chi-square distribution, the following corollary easily follows[5]:

Corollary 8.3.3. *For* $T > t$, $z(T)$ *has the following first two conditional moments:*

$$\mathrm{E}\left(z(T)|z(t)\right) = z_0 + \left(z(t) - z_0\right)e^{-\theta(T-t)}, \tag{8.7}$$

$$\mathrm{Var}\left(z(T)|z(t)\right) = \frac{z(t)\eta^2 e^{-\theta(T-t)}}{\theta}\left(1-e^{-\theta(T-t)}\right) + \frac{z_0\eta^2}{2\theta}\left(1-e^{-\theta(T-t)}\right)^2. \tag{8.8}$$

[5] In Appendix A.A, p.1150, we also derive an expression for, and a numerical approximation to, $\mathrm{E}(\sqrt{z(t)})$.

The transition distribution is useful for setting numerical bounds for PDE and Monte Carlo methods. Because it is somewhat complicated, we often find it convenient to use the stationary distribution of $z(t)$ (that is, the distribution of $z(\infty)$) instead, as an approximation.

Proposition 8.3.4. *The stationary distribution of $z(\cdot)$ in (8.4) is a Gamma distribution, see (3.9), and the stationary density $\pi(z)$ is given by*

$$\pi(z) = \frac{z^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)\,\beta^{-\alpha}},$$

where

$$\alpha = 2\theta z_0/\eta^2, \quad \beta = 2\theta/\eta^2.$$

In particular, the mean of the stationary distribution is given by

$$\int_0^\infty z\pi(z)\,dz = \frac{\alpha}{\beta} = z_0,$$

and the variance by

$$\int_0^\infty (z - z_0)^2\,\pi(z)\,dz = \frac{\alpha}{\beta^2} = \frac{z_0\eta^2}{2\theta}.$$

Proof. Follows directly from Proposition 8.3.2 and Corollary 8.3.3, by taking the limit $T - t \to \infty$.

Now let us look at the properties of the process $S(\cdot)$ for the underlying. The martingale property for $S(\cdot)$ should not be taken for granted in stochastic volatility models, but fortunately holds in our case:

Proposition 8.3.5. *The process $S(\cdot)$ given by (8.3)–(8.4) is a proper martingale.*

Proof. See Andersen and Piterbarg [2007]. □

The SDE (8.3) for $S(\cdot)$ can be integrated explicitly:

Proposition 8.3.6. *In the model (8.3)–(8.4), we have*

$$S(t) = \frac{1}{b}\left[(bS(0) + (1-b)\,L)\,X(t) - (1-b)\,L\right],$$

where

$$dX(t)/X(t) = \lambda b\sqrt{z(t)}\,dW(t), \quad X(0) = 1,$$

i.e.,

$$\ln X(t) = \lambda b\int_0^t \sqrt{z(s)}\,dW(s) - \frac{1}{2}\lambda^2 b^2\int_0^t z(s)\,ds. \tag{8.9}$$

Proof. Follows by applying Ito's lemma to $\ln(bS(t) + (1-b)L)$. □

The moment-generating function of $\ln X(t)$ in (8.9) is of fundamental importance for European option pricing in the model (8.3)–(8.4), and is linked to the moment-generating function of the integrated variance process, as the following proposition demonstrates.

Proposition 8.3.7. *Define*

$$\Psi_X(u;t) \triangleq \mathrm{E}\left(e^{u \ln X(t)}\right) = \mathrm{E}\left(X(t)^u\right). \tag{8.10}$$

In the model (8.3)–(8.4), for any $u \in \mathbb{C}$ for which the right-hand side exists, we have

$$\Psi_X(u;t) = \Psi_{\overline{z}}\left(\frac{1}{2}(\lambda b)^2 u(u-1), u; t\right),$$

where we have denoted

$$\Psi_{\overline{z}}(v,u;t) \triangleq \mathrm{E}^{\widetilde{\mathrm{P}}}\left(e^{v\overline{z}(t)}\right), \quad \overline{z}(t) \triangleq \int_0^t z(s)\,ds. \tag{8.11}$$

Under the new probability measure $\widetilde{\mathrm{P}}$ the process for $z(\cdot)$ is

$$dz(t) = \left(\theta(z_0 - z(t)) + \rho\eta\lambda b u z(t)\right)dt + \eta\sqrt{z(t)}\,d\widetilde{Z}(t), \quad z(0) = z_0, \tag{8.12}$$

with $\widetilde{Z}$ a $\widetilde{\mathrm{P}}$-Brownian motion. If $\rho = 0$, then $\widetilde{\mathrm{P}} = \mathrm{P}$ and $z(\cdot)$ in (8.11) follows (8.4) rather than (8.12).

Proof. From (8.9) we get

$$\mathrm{E}\left(e^{u \ln X(t)}\right) = \mathrm{E}\left(\varsigma(t)\exp\left(\frac{1}{2}u(u-1)\lambda^2 b^2 \int_0^t z(s)\,ds\right)\right),$$

where $\varsigma(t)$ is the exponential martingale

$$\begin{aligned}\varsigma(t) &= \mathcal{E}\left(u\lambda b\int_0^t \sqrt{z(s)}\,dW(s)\right)\\ &= \exp\left(u\lambda b\int_0^t \sqrt{z(s)}\,dW(s) - \frac{1}{2}u^2\lambda^2 b^2\int_0^t z(s)\,ds\right).\end{aligned}$$

Letting $\varsigma(t)$ be the density process for a measure change, Proposition 8.3.7 follows from Girsanov's theorem, see Theorem 1.5.1. □

A version of the proposition above also holds for a more general process (8.2) for $z(\cdot)$, see Andersen and Piterbarg [2007]. What makes the specification (8.4) particularly useful is the availability of a closed-form expression for $\Psi_{\overline{z}}(v,u;t)$.

Proposition 8.3.8. *For $\Psi_{\overline{z}}(v,u;t)$ defined by (8.11) we have that*

$$\begin{aligned}
\ln \Psi_{\overline{z}}(v,u;t) &= A(v,u) + B(v,u)\, z_0, \\
A(v,u) &= \frac{\theta z_0}{\eta^2}\left[2\ln\left(\frac{2\gamma}{\theta'+\gamma-e^{-\gamma T}(\theta'-\gamma)}\right)+(\theta'-\gamma)T\right], \\
B(v,u) &= \frac{2v\left(1-e^{-\gamma T}\right)}{(\theta'+\gamma)\left(1-e^{-\gamma T}\right)+2\gamma e^{-\gamma T}}, \\
\gamma = \gamma(v,u) &= \sqrt{(\theta')^2-2\eta^2 v}, \\
\theta' = \theta'(u) &= \theta-\rho\eta\lambda b u.
\end{aligned}$$

Proof. The process (8.12) is of the form

$$dz(t) = (\theta z_0 - \theta' z(t))\, dt + \eta\sqrt{z(t)}\, d\widetilde{Z}(t), \quad \theta' = \theta - \rho\eta\lambda b u,$$

which is of the same form as the short rate process in Cox et al. [1985], see Section 10.2. As demonstrated by, e.g., Dufresne [2001], the discount bond pricing result from Cox et al. [1985] (derived via PDE methods) immediately establishes the moment-generating function of the time integral of $z(\cdot)$. □

Beyond being useful in the proof of Proposition 8.3.7, measure changes are of primary importance in interest rate modeling, where a stochastic volatility model would typically be "embedded" in a full term structure model. To get a feel for issues that arise in this context, let us consider the impact of measure changes on the stochastic variance process. For this, let $V(t,X(t))$ be the numeraire-deflated price process for some asset in the model (8.3)–(8.4), where $V(t,x)$ is a deterministic function. Implicit in the notation is the assumption that the price does not depend on the stochastic variance process $z(\cdot)$, an assumption that holds true in the cases of interest to us. Assuming the price process is positive, it can be used as a numeraire, defining a new measure $\widetilde{\mathrm{P}}$, see Section 1.3. Since we have

$$dV(t,X(t))/V(t,X(t)) = \lambda b X(t)\frac{\partial \ln(V(t,X(t)))}{\partial x}\sqrt{z(t)}\, dW(t),$$

the process

$$\begin{aligned}
&\left(d\widetilde{W}(t), d\widetilde{Z}(t)\right)^{\top} = (dW(t), dZ(t))^{\top} \\
&-\left(\lambda b X(t)\frac{\partial \ln(V(t,X(t)))}{\partial x}\sqrt{z(t)}, \rho\lambda b X(t)\frac{\partial \ln(V(t,X(t)))}{\partial x}\sqrt{z(t)}\right)^{\top} dt
\end{aligned}$$

is a two-dimensional Brownian motion under the measure $\widetilde{\mathrm{P}}$, see Theorem 1.5.1, and we obtain the following result.

Proposition 8.3.9. *In the model (8.3)–(8.4), the dynamics of the stochastic variance process $z(\cdot)$ under a measure $\widetilde{\mathrm{P}}$ defined by a numeraire $V(t,X(t))$ are given by*

$$dz(t) = \widetilde{\theta}\,(t, X(t))\left(\widetilde{f}\,(t, X(t)) - z(t)\right)\,dt + \eta\sqrt{z(t)}\,d\widetilde{Z}(t),$$

where

$$\widetilde{\theta}(t,x) = \theta - \eta\rho\lambda bx\frac{\partial \ln\left(V(t,x)\right)}{\partial x}, \quad \widetilde{f}(t,x) = \frac{\theta z_0}{\widetilde{\theta}(t,x)},$$

and $\widetilde{Z}(\cdot)$ *is a* $\widetilde{\mathrm{P}}$*-Brownian motion.*

We note that if $\rho \neq 0$, not only do the speed of mean reversion and the mean reversion level get altered by the measure change, they become dependent on the process for the underlying $S(\cdot)$ itself. As mentioned before, this makes it difficult, if not impossible, to relate statistically-observed stochastic variance parameters to the risk-neutral ones. Additionally, non-zero value of the correlation ρ introduces technical complications in interest rate modeling due to the heavy use of measure change machinery, complications that we normally avoid by setting ρ to 0.

Returning to the examination of the properties of the S-process, we note that while $S(\cdot)$ in (8.3)–(8.4) is always a martingale (see Proposition 8.3.5), some of its higher-order moments may become infinite with time. This has important implications in interest rate modeling where values of some common types of contracts require finite second-order moments, see Chapter 16 on convexity derivatives. The following proposition gives sharp conditions on moment existence.

Proposition 8.3.10. *Consider the model (8.3)–(8.4). For a given* $u > 1$, *set* $v = (\lambda b)^2 u(u-1)/2 \geq 0$ *and define*

$$\beta = 2v/\eta^2 > 0, \quad \alpha = 2\left(\rho\eta\lambda bu - \theta\right)/\eta^2, \quad D = \alpha^2 - 4\beta.$$

The moment $\mathrm{E}(S(T)^u)$ *will be finite for* $T < T^*$ *and infinite for* $T \geq T^*$, *where* T^* *is given by*

1. $D \geq 0$, $\alpha < 0$:

$$T^* = \infty;$$

2. $D \geq 0$, $\alpha > 0$:

$$T^* = \gamma_+^{-1}\eta^{-2}\ln\left(\frac{\alpha/2 + \gamma_+}{\alpha/2 - \gamma_+}\right), \ \gamma_+ \triangleq \frac{1}{2}\sqrt{D};$$

3. $D < 0$:

$$T^* = 2\gamma_-^{-1}\eta^{-2} \times \left(\pi 1_{\{\alpha<0\}} + \arctan(2\gamma_-/\alpha)\right), \ \gamma_- \triangleq \frac{1}{2}\sqrt{-D}.$$

Proof. See Andersen and Piterbarg [2007]. □

The problem of moment explosions in the SV model (8.3)–(8.4) can be resolved by replacing (8.4) with a slightly more general specification (8.2) with $\psi(z) = z^p$ for $p < 1/2$, at a cost of losing some analytical tractability.

There are a number of subtle but important issues related to stochastic volatility processes with $\psi(z) = z^p$; the reader is referred to Andersen and Piterbarg [2007] for a comprehensive discussion. While somewhat outside the main focus of our exposition, we list some relevant results in Appendix 8.A.

8.4 Fourier Integration

Having covered the basics, we now turn to the problem of establishing of accurate pricing methods for the SV model. The method we present here is based on the application of Fourier integration methods, and is largely taken from Lewis [2000], with some modifications. Carr and Madan [1999], Lipton [2002], and Lee [2004], among many others, can be consulted for additional details.

8.4.1 General Theory

The following general result shows how to calculate call option prices when a moment-generating function is available for the logarithm of the underlying.

Theorem 8.4.1. *Let ξ be a random variable, and define its moment-generating function by $\chi(u)$,*

$$\chi(u) = \mathrm{E}\left(e^{u\xi}\right).$$

Then for $k \in \mathbb{R}$,

$$\mathrm{E}\left(\left(e^{\xi} - e^{k}\right)^{+}\right) = \chi(1) - \frac{e^{k}}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-k(\alpha+i\omega)}\chi\left(\alpha + i\omega\right)}{\left(\alpha + i\omega\right)\left(1 - \alpha - i\omega\right)}\, d\omega \tag{8.13}$$

for any $0 < \alpha < 1$ for which the right-hand side exists.

Proof. Let

$$c(k) = \mathrm{E}\left(\left(e^{\xi} - e^{k}\right)^{+}\right).$$

To improve regularity of our eventual numerical scheme, we split out a bounded component $\min(e^{\xi-k}, 1)$ from the unbounded function $(e^{\xi} - e^{k})^{+}$, writing

$$\begin{aligned} c(k) &= \mathrm{E}\left(\max\left(e^{\xi} - e^{k}, 0\right)\right) \\ &= \mathrm{E}\left(e^{\xi} - e^{k}\min\left(e^{\xi-k}, 1\right)\right) \\ &= \chi(1) - e^{k}\mathrm{E}\left(\min\left(e^{\xi-k}, 1\right)\right). \end{aligned}$$

Our intention is now to apply Fourier transforms in the computation of $\mathrm{E}(\min(e^{\xi-k}, 1))$. While the function $\min(e^{x-k}, 1)$ is bounded by design, it

is not integrable — it equals 1 for all $x \geq k$. To work around this, we can follow Carr and Madan [1999] and write, with $p(x)$ being the density of ξ,

$$\mathrm{E}\left(\min\left(e^{\xi-k},1\right)\right) = e^{-\alpha k}\int_{-\infty}^{\infty}\left[\min\left(e^{-(k-x)},1\right)e^{\alpha(k-x)}\right]\left[e^{\alpha x}p(x)\right]\,dx,$$

where $\alpha > 0$ is a classical *dampening constant*. Note that this integral is a convolution

$$\left(f_1 * f_2\right)(k) \triangleq \int_{-\infty}^{\infty} f_1\left(k-x\right) f_2(x)\,dx$$

of two functions,

$$f_1(x) = \min\left(e^{-x},1\right)e^{\alpha x}$$

and

$$f_2(x) = e^{\alpha x}p(x),$$

evaluated at k. Let $\mathcal{F}$ be Fourier transform and $\mathcal{F}^{-1}$ its inverse, i.e.,

$$\left(\mathcal{F}f\right)(\omega) \triangleq \int_{-\infty}^{\infty} e^{i\omega x}f(x)\,dx, \tag{8.14}$$

$$\left(\mathcal{F}^{-1}g\right)(x) \triangleq \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega x}g\left(\omega\right)\,d\omega. \tag{8.15}$$

As is well known, the Fourier transform of a convolution is a product of Fourier transforms, so

$$\begin{aligned}&\int_{-\infty}^{\infty}\left[\min\left(e^{-(k-x)},1\right)e^{\alpha(k-x)}\right]\left[e^{\alpha x}p(x)\right]\,dx\\ &\quad = \left(f_1 * f_2\right)(k) = \left(\mathcal{F}^{-1}\left(\mathcal{F}\left(f_1 * f_2\right)\right)\right)(k) = \left(\mathcal{F}^{-1}\left(g_1\left(\omega\right)g_2\left(\omega\right)\right)\right)(k),\end{aligned}$$

where

$$\begin{aligned}g_1\left(\omega\right) &= \int_{-\infty}^{\infty} e^{i\omega x}\min\left(e^{-x},1\right)e^{\alpha x}\,dx,\\ g_2\left(\omega\right) &= \int_{-\infty}^{\infty} e^{i\omega x}e^{\alpha x}p(x)\,dx.\end{aligned}$$

Simple calculations lead to

$$\begin{aligned}g_1\left(\omega\right) &= \int_{-\infty}^{0} e^{x(\alpha+i\omega)}dx + \int_{0}^{\infty} e^{x(-1+\alpha+i\omega)}dx\\ &= \frac{1}{\alpha+i\omega} - \frac{1}{\alpha-1+i\omega}\\ &= \frac{1}{\left(\alpha+i\omega\right)\left(1-\alpha-i\omega\right)},\\ g_2\left(\omega\right) &= \chi\left(\alpha+i\omega\right),\end{aligned}$$

where the convergence of integrals follows from the fact that $0 < \alpha < 1$. Therefore,

$$\mathrm{E}\left(\min\left(e^{\xi-k},1\right)\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-k(\alpha+i\omega)}\chi\left(\alpha+i\omega\right)}{\left(\alpha+i\omega\right)\left(1-\alpha-i\omega\right)}\,d\omega$$

and the theorem follows. □

Remark 8.4.2. The formula (8.13) from Theorem 8.4.1 can be re-written as

$$\mathrm{E}\left(\left(e^{\xi}-e^{k}\right)^{+}\right) = \chi(1) - \frac{e^{k}}{\pi}\int_{0}^{\infty} \mathrm{Re}\left(\frac{e^{-k(\alpha+i\omega)}\chi\left(\alpha+i\omega\right)}{\left(\alpha+i\omega\right)\left(1-\alpha-i\omega\right)}\right) d\omega,$$

a form that is used in, say, Attari [2004] and may yield computational benefits.

Proof. Let $\overline{x}$ be the complex conjugate of x, $x \in \mathbb{C}$. If $H(\omega)$ is such that

$$H\left(-\omega\right) = \overline{H\left(\omega\right)}, \tag{8.16}$$

then

$$\begin{aligned}
\int_{-\infty}^{\infty} H\left(\omega\right) d\omega &= \int_{-\infty}^{0} H\left(\omega\right) d\omega + \int_{0}^{\infty} H\left(\omega\right) d\omega \\
&= \int_{0}^{\infty} H\left(-\omega\right) d\omega + \int_{0}^{\infty} H\left(\omega\right) d\omega \\
&= \overline{\int_{0}^{\infty} H\left(\omega\right) d\omega} + \int_{0}^{\infty} H\left(\omega\right) d\omega \\
&= 2\mathrm{Re}\left(\int_{0}^{\infty} H\left(\omega\right) d\omega\right).
\end{aligned}$$

Since

$$\overline{\chi\left(\alpha+i\omega\right)} = \mathrm{E}\left(\overline{e^{(\alpha+i\omega)\xi}}\right) = \mathrm{E}\left(e^{(\alpha-i\omega)\xi}\right) = \chi\left(\alpha-i\omega\right),$$

the integrand in (8.13) satisfies (8.16) and the result follows. □

A result complimentary to Theorem 8.4.1 holds for a call option on ξ rather than e^{ξ}.

Theorem 8.4.3. *In the notations of Theorem 8.4.1,*

$$\mathrm{E}\left(\left(\xi-k\right)^{+}\right) = \left.\frac{d\chi(k)}{dk}\right|_{k=0} - k + \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-k(-\alpha+i\omega)}\chi\left(-\alpha+i\omega\right)}{\left(-\alpha+i\omega\right)^{2}} d\omega$$

for any $\alpha > 0$ for which the right-hand side exists.

Proof. As in the proof of Theorem 8.4.1, denote

$$c(k) = \mathrm{E}\left((\xi - k)^+\right).$$

While not strictly necessary, to keep the presentation consistent with the proof of Theorem 8.4.1, we manipulate this expression to obtain a bounded payoff inside the expected value,

$$\begin{aligned} c(k) &= \mathrm{E}\left(\max\left(\xi - k, 0\right)\right) \\ &= \mathrm{E}\left(\xi - \min\left(\xi, k\right)\right) \\ &= \chi'(0) - k - \mathrm{E}\left(\min\left(\xi - k, 0\right)\right), \end{aligned}$$

where $\chi'(\cdot)$ is the first-order derivative of the moment-generating function. Choosing $\alpha > 0$ and dampening the integrand with an exponential function, we obtain

$$\begin{aligned} \mathrm{E}\left((\xi - k)^+\right) = \chi'(0) - k \\ - e^{\alpha k} \int_{-\infty}^{\infty} \left[\min\left(-\left(k - x\right), 0\right) e^{-\alpha(k-x)}\right] \left[e^{-\alpha x} p(x)\right] dx, \end{aligned}$$

where $p(x)$ is the density of ξ. By the same arguments as in the proof of Theorem 8.4.1,

$$\mathrm{E}\left((\xi - k)^+\right) = \chi'(0) - k - e^{\alpha k}\left(\mathcal{F}^{-1}\left(g_1\left(\omega\right) g_2\left(\omega\right)\right)\right)(k),$$

where

$$\begin{aligned} g_1\left(\omega\right) &= \int_{-\infty}^{\infty} e^{i\omega x} \min\left(-x, 0\right) e^{-\alpha x}\, dx, \\ g_2\left(\omega\right) &= \int_{-\infty}^{\infty} e^{i\omega x} e^{-\alpha x} p(x)\, dx. \end{aligned}$$

Simple calculations lead us to

$$\begin{aligned} g_1\left(\omega\right) &= -\int_0^{\infty} x e^{x(-\alpha + i\omega)} dx = -\frac{1}{\left(-\alpha + i\omega\right)^2}, \\ g_2\left(\omega\right) &= \chi\left(-\alpha + i\omega\right), \end{aligned}$$

and the theorem follows. □

8.4.2 Applications to SV Model

Combining Theorem 8.4.1 with the closed-form expression for the moment-generating function in the SV model (Propositions 8.3.6, 8.3.7, and 8.3.8), we obtain an efficient formula for pricing European call and put options in

the model (8.3)–(8.4). As suggested in Andersen and Andreasen [2002], its numerical properties can be enhanced by a type of control variate method where we add the Black formula and subtract its Fourier representation, reducing the discretization errors in the process. We present the call price result in this form.

Theorem 8.4.4. *The price of a call option $c_{\mathrm{SV}}(0,S;T,K)$ in the SV model (8.3)–(8.4) is given by*

$$c_{\mathrm{SV}}(0,S;T,K) = \frac{1}{b}c_{\mathrm{B}}(0,S';T,K',\lambda b) - \frac{K'}{2\pi b}\int_{-\infty}^{\infty}\frac{e^{(1/2+i\omega)\ln(S'/K')}q(1/2+i\omega)}{\omega^2+1/4}\,d\omega, \quad (8.17)$$

where $c_{\mathrm{B}}(0,S';T,K',\sigma)$ is the Black formula for spot S', strike K', expiry T and volatility σ, with

$$S' = bS + (1-b)L, \quad K' = bK + (1-b)L.$$

Also, we have defined

$$q(u) = \Psi_{\overline{z}}\left(\frac{1}{2}(\lambda b)^2 u(u-1), u; T\right) - e^{\frac{1}{2}\lambda^2 b^2 z_0 T u(u-1)}, \quad (8.18)$$

with $\Psi_{\overline{z}}$ given in Proposition 8.3.8.

Remark 8.4.5. In (8.17) we use volatility λb in the Black model. As a further refinement, one can use the ATM volatility implied by the SV model instead. The ATM volatility can, for instance, be approximated by an expansion approach, as explained in Sections 8.7 and 9.2.

Proof. From Proposition 8.3.6,

$$\begin{aligned} c_{\mathrm{SV}}(0,S;T,K) &= \mathrm{E}(S(T)-K)^+ \\ &= \frac{1}{b}\mathrm{E}\left(S'e^{\ln X(T)} - K'\right)^+ \\ &= \frac{S'}{b}\mathrm{E}\left(e^{\ln X(T)} - e^{\ln(K'/S')}\right)^+. \end{aligned}$$

By Theorem 8.4.1 and the definition (8.10) of $\Psi_X(u;t)$,

$$c_{\mathrm{SV}}(0,S;T,K) = \frac{1}{b}\left(S' - \frac{K'}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-(\alpha+i\omega)\ln(K'/S')}\Psi_X(\alpha+i\omega;T)}{(\alpha+i\omega)(1-\alpha-i\omega)}d\omega\right), \quad (8.19)$$

where we have used the fact that $\Psi_X(1;T) = 1$. Applying this result to the SV model with $\eta = 0$, we find that the value of the option in the displaced log-normal model $c_{\mathrm{DLN}}(0,S;T,K)$ is given by

$$c_{\mathrm{DLN}}(0,S;T,K)=\frac{1}{b}\left(S'-\frac{K'}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-(\alpha+i\omega)\ln(K'/S')}\Psi_X^0(\alpha+i\omega;T)}{(\alpha+i\omega)(1-\alpha-i\omega)}d\omega\right), \tag{8.20}$$

where

$$\Psi_X^0(u;T)\triangleq \mathrm{E}\left(e^{u\left(\lambda b\sqrt{z_0}W(T)-\frac{1}{2}\lambda^2b^2z_0T\right)}\right)=e^{\frac{1}{2}\lambda^2b^2z_0T\left(u^2-u\right)}. \tag{8.21}$$

On the other hand,

$$c_{\mathrm{DLN}}(0,S;T,K)=\frac{1}{b}c_{\mathrm{B}}(0,S';T,K',\lambda b),$$

so that

$$\frac{1}{b}c_{\mathrm{B}}(0,S';T,K',\lambda b)$$
$$-\frac{1}{b}\left(S'-\frac{K'}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-(\alpha+i\omega)\ln\left(K'/S'\right)}\Psi_X^0(\alpha+i\omega;T)}{(\alpha+i\omega)(1-\alpha-i\omega)}d\omega\right)=0.$$

Adding the left-hand side of this identity, which is zero, to the right-hand side of (8.19), we obtain

$$c_{\mathrm{SV}}(0,S;T,K)=\frac{1}{b}c_{\mathrm{B}}(0,S';T,K',\lambda b)$$
$$-\frac{K'}{2\pi b}\int_{-\infty}^{\infty}\frac{e^{-(\alpha+i\omega)\ln\left(K'/S'\right)}q(\alpha+i\omega)}{(\alpha+i\omega)(1-\alpha-i\omega)}d\omega,$$

where

$$q(u)=\Psi_X(u;T)-\Psi_X^0(u;T).$$

Using Propositions 8.3.7 and 8.3.8 for $\Psi_X(u;T)$ and (8.21) for $\Psi_X^0(u;T)$, and setting $\alpha=1/2$, the result follows. □

Remark 8.4.6. The choice of $\alpha=1/2$ in Theorem 8.4.4 is common in practice (see Lipton [2002]) and appears to give robust and stable results in most situations. As first pointed out by Lewis [2001], the value of α can be seen to define an integration contour in the complex plane, and values of α other than $1/2$ can be used as long as $\alpha+i\omega$ for all $\omega\in\mathbb{R}$ lie in the so-called *strip of convergence*[6]. One can attempt to optimize α to improve the numerical properties of the integral, see, e.g., Lee [2004] or Lord and Kahl [2007] for details. Moreover, integration contours are not restricted to straight lines. Lucic [2007] shows that all singularities of the function $q(u)$ are real (for our definition of q), paving the way for finding better — curvilinear — contours.

[6]The region of $u\in\mathbb{C}$ for which the moment-generating function $\chi(u)$ exists. Heston [1993] and Lewis [2000] establish the strip of convergence for the Heston model. The strip is directly related to moment existence, for the latter see Proposition 8.3.10.

Remark 8.4.7. Integrating complex values functions, such as $q(\alpha + i\omega)$, in a complex domain typically requires some care. Particularly troublesome are multi-valued functions such as the complex logarithm, as present in the expression for $\Psi_{\bar{z}}$ in Proposition 8.3.8. Should an integration contour cross a branching cut of such a function, the value will jump to a different branch, typically leading to wrong results. Fortunately the moment-generating function as presented in Proposition 8.3.8 is free of such problems. This is not the case for other, mathematically equivalent, expressions, such as, say, the one given in the original paper Heston [1993] — the reader is referred to Albrecher et al. [2007] for proofs and a detailed discussion of related issues.

Remark 8.4.8. By Assumption 8.1.1, Theorem 8.4.4 does not cover the case $b = 0$. If needed, this case can be handled by utilizing Theorem 8.4.3 instead of Theorem 8.4.1.

8.4.3 Numerical Implementation

The Fourier integral in (8.17) can be evaluated directly by any numerical integration scheme, in what is sometimes called the *direct integration approach*, see Kilin [2007]. With suitable restrictions on the integration technique and the integration grid spacing, one can formulate the pricing formula as a *discrete Fourier transform* (DFT), allowing for the usage of the *Fast Fourier Transform* (FFT) method, see Press et al. [1992]. The FFT method is developed in Section 8.4.5 below for applications requiring calculations of option prices for multiple strikes — such as volatility smile calibration or evaluation of European payoffs beyond simple puts and calls. The FFT method is certainly not competitive for calculating a *single* call option price, so here we focus on the direct integration method.

A direct numerical integration of (8.17) involves a scheme to discretize the integral and to handle the infinite integration domain. Many algorithms of varying degrees of sophistication have been proposed, some of which involve adaptive error control, optimal choice of dampening parameter α, and the mapping of the infinite integration domain on to a finite one. Lee [2004], Kilin [2007], Kahl and Jäckel [2005], Lord and Kahl [2007] contain sample algorithms, none of which employ the Black control variate inherent in our formulation (Theorem 8.4.4). As the control variate produces powerful error cancellations, we find that its inclusion allows for excellent results even when much simpler integration schemes are employed. We outline one such approach here.

Turning first to the integration bounds, we focus on the behavior of the integrand in (8.17) for large $|\omega|$; in fact, by Remark 8.4.2, only the limit $\omega \to +\infty$ needs to be explored. It turns out that the function $q(1/2 + i\omega)$ decays exponentially for large ω. In particular, as we can write

$$|q(1/2 + i\omega)| = e^{\mathrm{Re}\,(\ln(q(1/2+i\omega)))},$$

we have the following result for $\ln(q(1/2+i\omega)$.

Proposition 8.4.9. *Under our standard assumption that $|\rho|<1$, for $q(\cdot)$ defined as in Theorem 8.4.4 we have*

$$\lim_{\omega\to+\infty}\frac{1}{\omega}\ln\left(q\left(1/2+i\omega\right)\right)=-q_\infty,$$

where we have defined

$$q_\infty\triangleq\frac{\lambda bz_0}{\eta}\left(\sqrt{1-\rho^2}+i\rho\right)\left(1+\theta T\right). \tag{8.22}$$

Proof. The proof is obtained by applying simple calculus to formulas from Proposition 8.3.8; here we merely sketch it following the ideas of Kahl and Jäckel [2005]. We consider the limit of large positive ω. Let us denote

$$u\left(\omega\right)=1/2+i\omega,\quad v\left(\omega\right)=\frac{1}{2}\left(\lambda b\right)^2u\left(\omega\right)\left(u\left(\omega\right)-1\right)=-\frac{1}{2}\left(\lambda b\right)^2\left(\omega^2+1/4\right).$$

Using the notations of Proposition 8.3.8, we have (we use "$\sim$" to denote equivalence in the limit $\omega\to+\infty$),

$$\theta'\left(u\left(\omega\right)\right)\sim-i\rho\eta\lambda b\omega,\quad\gamma(v\left(\omega\right),u\left(\omega\right))\sim\rho^c\eta\lambda b\omega,$$

where

$$\rho^c\triangleq(1-\rho^2)^{1/2}. \tag{8.23}$$

From the asymptotic behavior of $\gamma(\cdot,\cdot)$ it follows that the term $e^{-\gamma T}$ in the expressions for $A(\cdot,\cdot)$, $B(\cdot,\cdot)$ in Proposition 8.3.8 tends to zero as $\omega\to+\infty$. Therefore,

$$B\left(v\left(\omega\right),u\left(\omega\right)\right)\sim-\frac{\lambda b}{\eta}\left(\rho^c+i\rho\right)\omega,$$

and the logarithm in the definition of $A(\cdot,\cdot)$ tends to a constant,

$$\lim_{\omega\to+\infty}\ln\left(\frac{2\gamma}{\theta'+\gamma-e^{-\gamma T}\left(\theta'-\gamma\right)}\right)=\ln\left(\frac{2\rho^c}{\rho^c-i\rho}\right).$$

Therefore, only the term $(\theta'-\gamma)T$ in the expression for $A(\cdot,\cdot)$ grows with ω, and thus

$$A\left(v\left(\omega\right),u\left(\omega\right)\right)\sim-\frac{\lambda bz_0}{\eta}\theta\left(i\rho+\rho^c\right)T\omega.$$

Hence,

$$\begin{aligned}-\frac{1}{\omega}\ln\left(\Psi_X\left(1/2+i\omega;T\right)\right)&=-\frac{1}{\omega}\ln\left(\Psi_{\overline{z}}\left(v\left(\omega\right),u\left(\omega\right);T\right)\right)\\&=\frac{1}{\omega}\left(A\left(v\left(\omega\right),u\left(\omega\right)\right)+z_0B\left(v\left(\omega\right),u\left(\omega\right)\right)\right)\\&\to\frac{\lambda bz_0}{\eta}\left(\rho^c+i\rho\right)\left(1+\theta T\right)\end{aligned}$$

as $\omega \to +\infty$. Clearly, $\Psi_X^0(1/2+i\omega;T)$ decays faster than that, as $e^{-\text{const}\times\omega^2}$, so $q(\cdot)$ inherits its tail behavior from $\Psi_X(\cdot;T)$, and the result follows. □

The indefinite integral in Theorem 8.4.4 needs to be truncated before it can be evaluated numerically. Let $\omega_{\max} > 0$ be the upper truncation limit. We have the following simple tail estimate,

$$\begin{aligned}
&\left| \int_{\omega_{\max}}^{\infty} \frac{e^{(1/2+i\omega)\ln(S'/K')} q\left(1/2+i\omega\right)}{\omega^2+1/4}\, d\omega \right| \\
&\quad \le \int_{\omega_{\max}}^{\infty} \left| e^{(1/2+i\omega)\ln(S'/K')} \right| \frac{\left|q\left(1/2+i\omega\right)\right|}{\omega^2}\, d\omega \\
&\quad \le \sqrt{\frac{S'}{K'}} e^{-\mathrm{Re}\,(q_\infty)\omega_{\max}} \int_{\omega_{\max}}^{\infty} \frac{d\omega}{\omega^2} \\
&\quad = \sqrt{\frac{S'}{K'}} \frac{e^{-\mathrm{Re}\,(q_\infty)\omega_{\max}}}{\omega_{\max}}.
\end{aligned}$$

If $\varepsilon_\omega > 0$ is the absolute tolerance for computing the option price via (8.17) (a value of $\varepsilon_\omega = 10^{-3}$ to 10^{-6} is a reasonable choice), then we set the upper truncation limit $\omega_{\max}$ by the condition

$$\frac{e^{-\mathrm{Re}\,(q_\infty)\omega_{\max}}}{b\omega_{\max}} = \varepsilon_\omega, \tag{8.24}$$

where q_∞ is as given in Proposition 8.4.9. With Remark 8.4.2 in mind and a computational budget of N_ω points (N_ω is usually of the order of 100), we proceed to discretize uniformly over $[0, \omega_{\max}]$ and apply the rectangular rule

$$\begin{aligned}
&\mathrm{Re}\left(\int_0^{\infty} \frac{e^{(1/2+i\omega)\ln(S'/K')} q\left(1/2+i\omega\right)}{\omega^2+1/4}\, d\omega \right) \\
&\quad \approx \frac{\omega_{\max}}{N_\omega} \sum_{n=0}^{N_\omega-1} \frac{e^{\ln(S'/K')/2}}{\omega_n^2+1/4} \mathrm{Re}\left(e^{i\omega_n \ln(S'/K')} q\left(1/2+i\omega_n\right) \right),
\end{aligned} \tag{8.25}$$

where

$$\omega_n = \omega_{\max} n/N_\omega, \quad n = 0, \ldots, N_\omega - 1.$$

Other quadrature rules (e.g. the trapezoidal rule) can, of course, be used instead of the rectangular one.

8.4.4 Refinements of Numerical Implementation

While the method of Section 8.4.3 is simple and robust, numerical experiments show that the integration interval $[0, \omega_{\max}]$, with $\omega_{\max}$ obtained in (8.24), is often too wide, in the sense that a large proportion of the N_ω integration points are located in the region of integration where the integrand is so small that contributions to the integral are immaterial. To rectify

this, we can contemplate using an adaptive integration scheme, which by design would focus the computational work in regions where the integrand is material. Alternatively, we can refine our analysis of the integrand to provide guidance for where an ordinary integration scheme should spend its time. The latter is more involved but also more illuminating, so we pursue this approach here. Much of the material is based on Bang [2009], which can be consulted for additional details. As noted earlier, the ultimate benefit of sophisticated integration schemes (including the one proposed here) tends to be rather limited in practice, as long as the Black-Scholes control variate is properly employed.

We start by stating the following refinement of Proposition 8.4.9.

Proposition 8.4.10. *Let $q(\cdot)$ be defined as in Theorem 8.4.4 and assume, as always, that $|\rho| < 1$. Then for any $\epsilon > 0$ there exists $\Omega_\epsilon > 0$ such that, for any ω that satisfies*

$$\omega \geq \max\left(\Omega_\epsilon, \frac{5}{\eta\lambda b \rho^c T}\right),$$

we have

$$\frac{1-\epsilon}{\omega^2} \leq \left|\ln\left(q\left(1/2+i\omega\right)\right) - \left(-q_\infty\omega + q_0 - \frac{q_{-1}}{\omega}\right)\right| \leq \frac{1+\epsilon}{\omega^2}, \tag{8.26}$$

where (compare to (8.22))

$$\begin{aligned}
q_\infty &= \frac{\lambda b z_0}{\eta}\left(\rho^c + i\rho\right)\left(1+\theta T\right),\\
q_0 &= \frac{z_0}{\rho^c\eta^2}\left(\rho^c+i\rho\right)\widehat{\theta}(1+\theta T) + \frac{2\theta z_0}{\eta^2}\left(\ln\left(2\rho^c\right) + i\arctan\left(\frac{\rho}{\rho^c}\right)\right),\\
q_{-1} &= \frac{\theta z_0}{\eta^2}\left(T\mu\eta\lambda b + 2\widehat{\theta}\frac{\rho^c+i\rho}{\left(\rho^c\right)^2\eta\lambda b}\right) + \mu\frac{\lambda b z_0}{\eta},\\
\mu &= \frac{\widehat{\theta}^2}{2\eta^2(\lambda b)^2\left(\rho^c\right)^3} + \frac{1}{8\rho^c}.
\end{aligned}$$

Here $\rho^c = (1-\rho^2)^{1/2}$ is given by (8.23) and $\widehat{\theta} = \theta'(1/2)$, where $\theta'(u) = \theta - \rho\eta\lambda b u$ is defined in Proposition 8.3.8.

Proof. The proof is by expanding $\ln(q(1/2+i\omega))$ into a series in $1/\omega$ for small values of $1/\omega$, along the lines of the proof of Proposition 8.4.9. Full details are available in Bang [2009]. □

Let us denote

$$r(\omega) = \ln\left(q\left(1/2+i\omega\right)\right)$$

and by $r_\infty(\omega)$ its expansion to the zeroth order for large ω (see (8.26)),

$$r_\infty(\omega) = -q_\infty\omega + q_0.$$

Consider the integral on the left-hand side of (8.25), and let us split out a part that covers the region of (approximate) validity for the asymptotic approximation $\ln(q(1/2+i\omega)) \approx r_\infty(\omega)$. To define this region, let us choose $\varepsilon'_\omega > 0$ reasonably small (of the order 10^{-2}) and pick $\omega'_{\max} > 0$ such that the following two conditions are simultaneously met:

$$\omega'_{\max} > \max\left(\frac{5}{\eta\lambda b\rho^c T}\right) \tag{8.27}$$

and, for any $\omega > \omega'_{\max}$,

$$\frac{|q_{-1}|}{\omega} \leq |r_\infty(\omega)|\,\varepsilon'_\omega. \tag{8.28}$$

Then, from Proposition 8.4.10,

$$\frac{|\ln(q(1/2+i\omega)) - r_\infty(\omega)|}{|r_\infty(\omega)|} \approx \frac{|q_{-1}|}{\omega\,|r_\infty(\omega)|} \leq \varepsilon'_\omega$$

and, thus, for $\omega > \omega'_{\max}$, the function $\ln(q(1/2+i\omega))$ is indeed well-approximated by $r_\infty(\omega)$. Accordingly, we write

$$\int_0^\infty \frac{e^{(1/2+i\omega)\ln(S'/K')}q(1/2+i\omega)}{\omega^2+1/4}\,d\omega = I_1 + I_2 + I_3, \tag{8.29}$$

where

$$I_1 = \int_0^{\omega'_{\max}} \frac{e^{(1/2+i\omega)\ln(S'/K')}q(1/2+i\omega)}{\omega^2+1/4}\,d\omega, \tag{8.30}$$

$$I_2 = \int_{\omega'_{\max}}^\infty \frac{e^{(1/2+i\omega)\ln(S'/K')}}{\omega^2+1/4}\left(q(1/2+i\omega) - e^{r_\infty(\omega)}\right)d\omega, \tag{8.31}$$

$$I_3 = \int_{\omega'_{\max}}^\infty \frac{e^{(1/2+i\omega)\ln(S'/K')}}{\omega^2+1/4}e^{r_\infty(\omega)}\,d\omega. \tag{8.32}$$

As it turns out, the integral I_3 in (8.32) can be expressed through special functions. Let $E_1(z)$ be the so-called *exponential integral* (see Abramowitz and Stegun [1965]), i.e. an analytic continuation of the integral

$$E_1(z) = \int_1^{+\infty} \frac{e^{-zk}}{z}\,dk$$

to the complex plane. We then have the following result.

Lemma 8.4.11. *Let a and c be two non-negative real numbers and let z be a complex number such that* $\mathrm{Re}(z) > 0$. *Then*

$$R(z,a,c) \triangleq \int_c^\infty \frac{e^{-zk}}{k^2+a^2}\,dk$$
$$= \frac{1}{2ia}\left(e^{-iaz}E_1\left(z\left(c-ia\right)\right) - e^{iaz}E_1\left(z\left(c+ia\right)\right)\right).$$

Proof. Follows by standard contour integration methods of complex analysis. Details are in Bang [2009]. □

Remark 8.4.12. The function $E_1(\cdot)$ can be evaluated numerically using an algorithm from Press et al. [1992]. Bang [2009] also recommends an efficient algorithm available from `http://jin.ece.uiuc.edu`.

With the help of Lemma 8.4.11, we can rewrite I_3 in (8.32) as

$$\int_{\omega'_{\max}}^\infty \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)}}{\omega^2+1/4} e^{r_\infty(\omega)}\,d\omega = e^{q_0+\ln\left(S'/K'\right)/2}$$
$$\times R(q_\infty - i\ln\left(S'/K'\right), 1/2, \omega'_{\max}),$$

and calculate it efficiently using Remark 8.4.12.

Turning next to the integral I_2 in (8.29), we wish to employ a quadrature rule designed to handle the oscillations of the integrand in (8.31). To that end, and following Bang [2009], we introduce a step size

$$\delta_\omega = \frac{\left(\lambda b z_0 \sqrt{T}\right)^{-1}}{2N_{\text{stdev}}},$$

where N_{stdev} is a user-specified range in standard deviations[7] (typically 5–6), set the number of points to be N''_ω (to be specified shortly), define

$$\omega''_n = \omega'_{\max} + \delta_\omega n, \quad n = 0, \ldots, N''_\omega,$$

and write

$$I_2 \approx e^{q_0}\sqrt{\frac{S'}{K'}} \int_{\omega'_{\max}}^{\omega'_{\max}+\delta_\omega N''_\omega} \frac{e^{\omega(-q_\infty + i\ln\left(S'/K'\right))}}{\omega^2+1/4}\left(e^{r(\omega)-r_\infty(\omega)}-1\right)d\omega$$
$$= e^{q_0}\sqrt{\frac{S'}{K'}} \sum_{n=0}^{N''_\omega-1} \int_{\omega''_n}^{\omega''_{n+1}} \frac{e^{\omega(-q_\infty + i\ln\left(S'/K'\right))}}{\omega^2+1/4}\left(e^{r(\omega)-r_\infty(\omega)}-1\right)d\omega,$$

so that

[7] This step size in Fourier space is inspired by a Fourier transform of a Gaussian distribution. If the "width" of a Gaussian PDF is given by its standard deviation σ, then the "width" of its characteristic function is given by $1/\sigma$.

$$I_2 \approx e^{q_0}\sqrt{\frac{S'}{K'}}\sum_{n=0}^{N''_\omega-1}\frac{e^{r(\omega''_n)-r_\infty(\omega''_n)}-1}{(\omega''_n)^2+1/4}\int_{\omega''_n}^{\omega''_{n+1}} e^{\omega(-q_\infty+i\ln(S'/K'))}\,d\omega$$

$$= e^{q_0}\sqrt{\frac{S'}{K'}}\sum_{n=0}^{N''_\omega-1}\frac{e^{r(\omega''_n)-r_\infty(\omega''_n)}-1}{(\omega''_n)^2+1/4}$$

$$\times\frac{e^{\omega''_{n+1}(-q_\infty+i\ln(S'/K'))}-e^{\omega''_n(-q_\infty+i\ln(S'/K'))}}{-q_\infty+i\ln(S'/K')}. \tag{8.33}$$

Note how we integrated analytically the oscillatory part of the integrand on the last step. With this scheme in place, we calculate I_2 using the quadrature rule (8.33) with N''_ω terms of the sum where we choose N''_ω adaptively by stopping when incremental changes from new terms in the sum are small enough..

Finally, let us discuss the term I_1 in (8.29), defined by (8.30). Here nothing special[8] is needed and we can just use a quadratic or trapezoidal rule with a given budget of N'_ω points (say, around 50 or so) along the same lines as we did in (8.25).

In conclusion, let us summarize the complete algorithm for calculating the integral in (8.29). First we choose a small $\varepsilon'_\omega > 0$ (of the order 10^{-2}) and find the cutoff point $\omega'_{\max}$ that satisfies (8.27)–(8.28). Then we decompose the integral in (8.29) into three parts. The first integral I_1 is calculated by the standard quadratic or trapezoidal rule, similarly to (8.25). The second integral I_2 is calculated by the quadrature rule (8.33) with the number of points determined by the convergence criteria (relative or absolute). Finally the term I_3 is calculated per Remark 8.4.12. We note that while this scheme is more complex than what we described in Section 8.4.3, it does result in a faster and more accurate algorithm with a better utilization of the computational budget.

8.4.5 Fourier Integration for Arbitrary European Payoffs

Consider the problem of computing prices of European-style options with arbitrary payoffs. In particular, let $f(x)$ be a payoff function, and consider the problem of computing the following expected value,

$$\mathrm{E}\left(f\left(S(T)\right)\right).$$

Clearly,

[8] Of the two terms in the definition of $q(1/2+i\omega)$ in (8.18), the (second) one related to the Gaussian distribution decays much faster than the (first) one related to the SV model, as we already noted. Hence, we can stop sampling the second term for smaller values of ω, to save a bit on calculation time. This is described in Bang [2009].

$$\mathrm{E}\left(f\left(S(T)\right)\right) = \int f(K)\mathrm{P}\left(S(T) \in dK\right)$$

and, by (7.5),

$$\mathrm{E}\left(f\left(S(T)\right)\right) = \int f(K)\frac{\partial^2 c\left(0, S(0); T, K\right)}{\partial K^2} dK, \tag{8.34}$$

where $c(0, S; T, K)$ is the European call option value for the process $S(\cdot)$. Integrating by parts, we obtain a useful representation of a general European payoff in terms of European calls and puts.

Proposition 8.4.13. *For any twice-continuously differentiable*[9] $f(x)$, *the value of a European option with payoff* $f(\cdot)$ *and expiry* T *is equal to the weighted integral of call and put options with weights equal to the second derivative of* $f(\cdot)$,

$$\begin{aligned}\mathrm{E}\left(f\left(S(T)\right)\right) &= f\left(K^*\right) + f'\left(K^*\right)\left(S(0) - K^*\right) \\ &+ \int_{-\infty}^{K^*} p\left(0, S(0); T, K\right) f''(K)\, dK + \int_{K^*}^{\infty} c\left(0, S(0); T, K\right) f''(K)\, dK,\end{aligned} \tag{8.35}$$

for any K^*.

Proof. Follows by integration by parts of (8.34). □

A combination of the suitably-discretized integral representation from Proposition 8.4.13 and Theorem 8.4.4 gives us an algorithm for computing values of European-style options with arbitrary payoffs. With the need to simultaneously compute call option prices of different strikes, the FFT method may deserve a closer look. In order to apply FFT, the discretization scheme of the integrals in (8.35) should be chosen carefully. From Theorem 8.4.4, the integrals to evaluate are

$$I\left(K'\right) = \int_{-\infty}^{\infty} \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)}}{\omega^2 + 1/4} q\left(1/2 + i\omega\right) d\omega \tag{8.36}$$

for various K'. We set $K^* = S(0)$ in (8.35) and discretize K in such a way that $\ln(S'/K')$ in (8.36) are equidistant. In particular, we choose $\delta > 0$, the discretization step, and define

$$x_n = \delta n, \quad K'_n = S' e^{x_n}, \quad n = -N, \ldots, N.$$

This leads to

$$bK_n + (1-b)L = \left(bS + (1-b)L\right) e^{x_n},$$

or

[9] But see Section 16.6.1 for extensions.

$$K_n = \left(S + \frac{1-b}{b}L\right) e^{x_n} - \frac{1-b}{b}L.$$

Then

$$I_n \triangleq I\left(K_n'\right) = \int_{-\infty}^{\infty} \frac{e^{-(1/2+i\omega)\delta n}}{\omega^2 + 1/4} q\left(1/2 + i\omega\right) d\omega = e^{-0.5\delta n} J_n,$$

$$J_n \triangleq \int_{-\infty}^{\infty} e^{-i\omega\delta n} \frac{q\left(1/2 + i\omega\right)}{\omega^2 + 1/4} d\omega.$$

At a computational effort of $O(N \ln N)$, all J_n's can now be evaluated by applying (inverse) FFT to the function

$$\frac{q\left(1/2 + i\omega\right)}{\omega^2 + 1/4}.$$

Once the J_n are computed, all

$$p_{\mathrm{SV}}\left(0, S; T, K_n\right), \quad c_{\mathrm{SV}}\left(0, S; T, K_n\right), \quad n = -N, \ldots, N,$$

can be calculated easily. The value of the option with any payoff $f(\cdot)$ is then obtained by discretizing the integrals in (8.35). We state the result as a proposition.

Proposition 8.4.14. *Fix $\delta > 0$. Let K_n, K_n', $n = -N, \ldots, N$, be defined by*

$$K_n = \left(S + \frac{1-b}{b}L\right) e^{\delta n} - \frac{1-b}{b}L, \quad K_n' = S' e^{\delta n}.$$

Then the value of a call option with payoff $f(\cdot)$ at time T in the SV model (8.3)-(8.4), is approximately given by

$$\mathrm{E}\left(f\left(S(T)\right)\right) \approx f\left(S(0)\right) + \sum_{n=-N}^{-1} p_{\mathrm{SV}}\left(0, S; T, K_n\right) f''\left(K_n\right)\left(K_{n+1} - K_n\right) + \sum_{n=0}^{N-1} c_{\mathrm{SV}}\left(0, S; T, K_n\right) f''\left(K_n\right)\left(K_{n+1} - K_n\right),$$

where

$$c_{\mathrm{SV}}\left(0, S; T, K_n\right) = \frac{1}{b} c_{\mathrm{B}}\left(0, S'; T, K_n', \lambda b\right) - \frac{K_n'}{2\pi b} e^{-0.5\delta n} J_n,$$

$$p_{\mathrm{SV}}\left(0, S; T, K_n\right) = \frac{1}{b} p_{\mathrm{B}}\left(0, S'; T, K', \lambda b\right) - \frac{K_n'}{2\pi b} e^{-0.5\delta n} J_n,$$

with $\{J_n\}_{n=-N}^{N}$ evaluated by an inverse FFT transform of the function

$$\frac{q\left(1/2 + i\omega\right)}{\omega^2 + 1/4},$$

and $q(u)$ given in Theorem 8.4.4.

Using FFT to compute the $2N+1$ J_n-integrals improves numerical effort of a direct integration scheme, from $O(N^2)$ to $O(N \ln N)$. On the other hand, FFT has several potential drawbacks, including the fact that it imposes quite onerous requirements on the discretization of the strike domain, requiring that N be a power of 2 and that the grid be equidistant in $\ln(S'/K')$. Also, by the nature of FFT, an equidistant grid of the same size is then used to discretize the frequency domain. Both choices are often suboptimal — for example, we may want to choose a strike grid to take into account particular features of the payoff $f(\cdot)$, and we may want to discretize the frequency domain with a different number of grid points and/or non-equidistant spacing. In fact, Kilin [2007] observes that the integration effort is dominated by the calculation of the values of $q(1/2+i\omega)$ for different ω and that they, critically, do not depend on strike. Kilin [2007] convincingly demonstrates that a careful implementation of the direct integration method of (8.17), even for multiple strikes, is often more efficient than FFT, provided that i) the values of $q(\cdot)$ are cached and reused when valuing different options, and ii) better discretization schemes are employed in the strike/frequency domains than those required by the FFT method.

8.5 Integration in Variance Domain

Under the assumption $\rho = 0$, a well-known "mixing" result (see e.g. Hull and White [1987]) represents the value of a European call option in the SV model (8.3)–(8.4) as an integral of the values of call options under the displaced log-normal model against the distribution of integrated variance. Specifically, the following lemma holds.

Lemma 8.5.1. *In the SV model (8.3)–(8.4) with $\rho = 0$, the value of a call option is given by*

$$c_{\mathrm{SV}}(0,S;T,K) = \frac{1}{b}\mathrm{E}\left(c_{\mathrm{B}}\left(0,S;T,K,\lambda b\sqrt{\overline{z}(T)/T}\right)\right), \tag{8.37}$$

where (see (8.11))

$$\overline{z}(T) = \int_0^T z(t)\,dt$$

and $c_{\mathrm{B}}(\cdot,\cdot;\cdot,\cdot,\sigma)$ is the value of a call option in the Black model with volatility σ.

Proof. Follows by conditioning on the trajectory of $z(\cdot)$ and using the independence of the Brownian motion $W(\cdot)$ of $z(\cdot)$. □

Remark 8.5.2. An extension of this result to non-zero correlation ρ exists, see Proposition A.3.7 and in particular equation (A.39). Unfortunately it cannot be used for our purposes here, as the more general formula involves not only the time integral of $z(\cdot)$ but also other random variables.

It is natural to treat the function under the expected value operator in (8.37) as a function of $\overline{z}(T)$,

$$c_{\mathrm{SV}}(0,S;T,K) = \mathrm{E}\left(C\left(\overline{z}(T)\right)\right), \quad C(U) = \frac{1}{b}c_{\mathrm{B}}\left(0,S;T,K,\lambda b\sqrt{U/T}\right). \tag{8.38}$$

As the moment-generating function $\Psi_{\overline{z}}(u,0;T)$ of $\overline{z}(T)$ is known from Proposition 8.3.8, the expected value in (8.38) can be computed by Fourier integration. In particular, denoting by $p_{\overline{z}}(U)$ the probability density function of $\overline{z}(T)$, consider using (8.38) to argue that

$$\begin{aligned} c_{\mathrm{SV}}(0,S;T,K) &= \int_0^\infty C(U)p_{\overline{z}}(U)\,dU \\ &= \frac{1}{2\pi}\int_0^\infty C(U)\int_{-\infty}^\infty e^{-i\omega U}\Psi_{\overline{z}}(i\omega,0;T)\,d\omega\,dU \\ &= \frac{1}{2\pi}\int_{-\infty}^\infty \Psi_{\overline{z}}(i\omega,0;T)\left(\int_0^\infty C(U)e^{-i\omega U}\,dU\right)d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^\infty \Psi_{\overline{z}}(i\omega,0;T)\left(\mathcal{F}C\right)(-\omega)\,d\omega, \end{aligned}$$

where

$$\left(\mathcal{F}C\right)(\omega) \triangleq \int_0^\infty e^{i\omega U}C(U)\,dU \tag{8.39}$$

is the Fourier transform of $C(U)$ and we have used in the second equality the fact that $\Psi_{\overline{z}}$ is the Fourier transform of $p_{\overline{z}}$.

This argument demonstrates the main idea behind Fourier integration in the variance domain, but suffers from the fundamental problem that the function $C(\cdot)$ is not integrable, whereby the Fourier transform (8.39) is not well-defined. Fortunately we can solve the problem by the standard remedy of introducing a dampening function $e^{-\alpha U}$ in the integrand, as the following proposition demonstrates.

Proposition 8.5.3. *For $\alpha > 0$ such that $\Psi_{\overline{z}}(\alpha,0;T)$ exists, the following holds,*

$$c_{\mathrm{SV}}(0,S;T,K) = \frac{1}{2\pi}\int_{-\infty}^\infty \Psi_{\overline{z}}(\alpha+i\omega,0;T)\left(\mathcal{F}\widehat{C}\right)(-\omega)\,d\omega,$$

where

$$\widehat{C}(U) = C(U)e^{-\alpha U}, \tag{8.40}$$

and $\Psi_{\overline{z}}(u,0;T)$ is given in Proposition 8.3.8.

Proof. We have

$$
\begin{aligned}
c_{\mathrm{SV}}(0,S;T,K) &= \int_0^\infty C(U)e^{-\alpha U}\left(e^{\alpha U}p_{\overline{z}}(U)\right)\,dU \\
&= \frac{1}{2\pi}\int_0^\infty C(U)e^{-\alpha U}\left(\int_{-\infty}^\infty e^{-i\omega U}\Psi_{\overline{z}}(\alpha+i\omega,0;T)\,d\omega\right)\,dU \\
&= \frac{1}{2\pi}\int_{-\infty}^\infty \Psi_{\overline{z}}(\alpha+i\omega,0;T)\left(\int_0^\infty C(U)e^{-\alpha U}e^{-i\omega U}\,dU\right)\,d\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^\infty \Psi_{\overline{z}}(\alpha+i\omega,0;T)\left(\mathcal{F}\widehat{C}\right)(-\omega)\,d\omega.
\end{aligned}
$$

□

It is probably the case that the numerical method based on the result of Proposition 8.5.3 is not as speedy as the direct integration method in Section 8.4, but it allows for interesting generalizations to arbitrary payoff functions and arbitrary skew functions, a setup where it compares favorably to Monte Carlo or PDE methods. With this generalization in mind, consider the general model specification (8.1)–(8.2), where we have the following counterpart to Lemma 8.5.1.

Lemma 8.5.4. *For a positive constant v, let $g(t,S;v)$ satisfy the PDE*

$$
\frac{\partial g(t,S;v)}{\partial S} + \frac{1}{2}v\varphi(S)^2\frac{\partial^2 g(t,S;v)}{\partial S^2} = 0, \tag{8.41}
$$

subject to the terminal boundary condition $g(T,S;v) = f(S)$. For the general stochastic volatility model dynamics (8.1)–(8.2) with $\rho = 0$ we have

$$
\mathrm{E}\left(f\left(S(T)\right)\right) = \mathrm{E}\left(g\left(0,S(0);T^{-1}\lambda^2\overline{z}(T)\right)\right). \tag{8.42}
$$

Consistent with (8.39) and (8.40), we proceed to introduce a Fourier transform of a dampened function g,

$$
\left(\mathcal{F}\widehat{g}\right)(\omega) = \int_{-\infty}^\infty e^{i\omega U}e^{-\alpha U}g\left(0,S(0);T^{-1}\lambda^2 U\right)dU, \tag{8.43}
$$

where $\alpha > 0$ is as in Proposition 8.5.3. Then we have the following generalization of Proposition 8.5.3.

Proposition 8.5.5. *Consider the system (8.1)–(8.2), with $\psi(z) = \sqrt{z}$. Let $g(t,S;v)$ be as in (8.41) and $\left(\mathcal{F}\widehat{g}\right)$ as in (8.43) for $\alpha > 0$ such that $\Psi_{\overline{z}}(\alpha + i\omega,0;T)$ is finite for all ω. Then*

$$
\mathrm{E}\left(f\left(S(T)\right)\right) = \frac{1}{2\pi}\int_{-\infty}^\infty \left(\mathcal{F}\widehat{g}\right)(-\omega)\Psi_{\overline{z}}(\alpha+i\omega,0;T)\,d\omega,
$$

where $\Psi_{\overline{z}}(u,0;T)$ is given in Proposition 8.3.8.

The proposition gives us a way to compute values of arbitrary European options in a model with an essentially arbitrary volatility function $\varphi(\cdot)$. In calculating the integral in (8.43), we need a way to efficiently compute the function $g(0, S(0); v)$ from (8.41) for many different values of v. Fortunately, in Chapter 7 we developed many such methods, ranging from analytical expressions, to expansions and finite difference methods[10]. We note that if the function $\varphi(\cdot)$ is complicated enough to require finite difference methods, it is crucial that we use the "trick" of Section 7.4.1 to ensure that only a single finite difference grid is solved.

Remark 8.5.6. It can be verified that the moment-generating function $\Psi_{\overline{z}}(u, 0; T)$ is finite in a neighborhood around $u = 0$. Moments of arbitrary order of $\overline{z}(T)$ consequently exist and can be computed by differentiation

$$\mathrm{E}\left(\overline{z}(T)^n\right) = \frac{d^n}{du^n}\Psi_{\overline{z}}\left(u, 0; T\right)\Big|_{u=0}, \quad n = 1, 2, \ldots .$$

Among other things, these moments can be used to dimension the U-grid used for the integration algorithm. For instance, for a given confidence multiplier γ (e.g. 5 or 10) we can, somewhat crudely, set

$$U_{\max} = \mathrm{E}\left(\overline{z}(T)\right) + \gamma\sqrt{\mathrm{Var}\left(\overline{z}(T)\right)}, \quad U_{\min} = \left(\mathrm{E}\left(\overline{z}(T)\right) - \gamma\sqrt{\mathrm{Var}\left(\overline{z}(T)\right)}\right)^+ .$$

More elaborate schemes are also possible.

We note that Proposition 8.5.5 can also be applied to the case $\psi(z) = \sqrt{z - v}$, where $v > 0$ is a constant and where we enforce the additional constraint that $v < z_0$. To see this, consider the SDE

$$dz(t) = \theta\left(z_0 - z(t)\right)dt + \eta\sqrt{z(t) - v}\, dZ(t),$$

and set $z^*(t) = z(t) - v$. Then

$$dz^*(t) = \theta\left(z_0^* - z^*(t)\right)dt + \eta\sqrt{z^*(t)}\, dZ(t), \quad z_0^* = z_0 - v > 0, \tag{8.44}$$

and

$$\begin{aligned}\mathrm{E}\left(e^{u\int_0^T z(t)\,dt}\right) &= \mathrm{E}\left(e^{u\int_0^T (z^*(t)+v)dt}\right)\\ &= e^{uvT}\mathrm{E}\left(e^{u\int_0^T z^*(t)\,dt}\right) = e^{uvT}\Psi_{\overline{z}}\left(u, 0; T\right),\end{aligned}$$

where $\Psi_{\overline{z}}(u, 0; T)$ is computed as in Proposition 8.3.8 with the substitution $z_0 \to z_0 - v$. The form $\psi(x) = \sqrt{x - v}$ is useful if we wish to keep the process

[10]Many of the methods in Chapter 7 were specific to calls, for which the boundary condition on the PDE is $f(S(T)) = (S(T) - K)^+$. Not only is this case by far the most important in practice, but also helps with pricing of other payouts via the replication approach (Proposition 8.4.13).

$z(t)$ away from $z = 0$: it easily follows from (8.44) and $z(t) = z^*(t) + v$ that $z(t)$ will never go below v. According to Proposition 8.A.1, another way to keep $z(\cdot)$ away from the origin is to use $\psi(x) = x^p$, $1/2 < p < 1$. This case, however, has no analytical tractability.

For general $\psi(\cdot)$, let us consider ways to characterize the function $\Psi_{\overline{z}}(u, 0; T)$ that we now define by (8.11) for a general $z(\cdot)$ in (8.2). A useful starting point is the following result, easily proven from the Feynman-Kac formula in Section 1.8.

Lemma 8.5.7. *Let*

$$dz(t) = \theta\left(z_0 - z(t)\right) dt + \eta\psi\left(z(t)\right) dZ(t).$$

Then $\Psi_{\overline{z}}(u, 0; T) = L(0, z_0; u)$, *where* $L(t, z; u)$ *satisfies the PDE*

$$\frac{\partial L}{\partial t} + \theta\left(z_0 - z\right)\frac{\partial L}{\partial z} + \frac{1}{2}\eta^2\psi(z)^2\frac{\partial^2 L}{\partial z^2} + uzL = 0,$$

subject to the boundary condition $L(T, z; u) = 1$.

Solution of the PDE in Lemma 8.5.7 can, of course, be done by finite difference methods, but at considerable numerical expense. An asymptotic expansion approach with decent precision is possible, however, and shall be demonstrated in Section 9.2 for the more general case of time-dependent λ. As it turns out, for many choices of $\psi(\cdot)$ — most notably for $\psi(z) = z^p$ — naively writing

$$\psi\left(z(t)\right) \approx \sqrt{z(t)}\psi\left(z(0)\right)/\sqrt{z(0)}$$

and then using the expression for $\Psi_{\overline{z}}(u, 0; T)$ from Proposition 8.3.8 often gives good results. Indeed, as shown in Andersen and Brotherton-Ratcliffe [2005], for call options, the dependence of option values on p in the specification $\psi(z) = z^p$ is quite mild across a reasonably wide range of strikes.

For complicated functions $\varphi(\cdot)$ and $\psi(\cdot)$ — and for the case where $\rho \neq 0$ — we always have the option of abandoning Fourier methods altogether and instead opting for more generally applicable numerical techniques, such as Monte Carlo and two-dimensional finite difference methods. We cover the application of these schemes to stochastic volatility models later on, in Sections 9.5 and 9.4, respectively.

8.6 CEV-Type Stochastic Volatility Models and SABR

As discussed earlier, certain choices of $\varphi(\cdot)$ and $\psi(\cdot)$ introduce technical problems, such as exploding higher-order moments of $S(\cdot)$, non-zero probability of generating negative $S(\cdot)$, or non-zero probability of the variance process $z(\cdot)$ being absorbed at zero. In practice, moment explosion is often the thorniest of these issues, as it has the potential to produce severe errors

for certain common securities (see Section 16.9). As it turns out, a simple switch from a linear function for $\varphi(\cdot)$ to a CEV-type specification prevents moment explosions that exist (Proposition 8.3.10) in the SV model. This is a useful result, so let us state it formally below. The proof is in Andersen and Piterbarg [2007].

Proposition 8.6.1. *Consider the model (8.1)–(8.2) with* $\varphi(x) = x^c$ *and* $\psi(z) = z^p$, *with* $0 < c < 1$ *and* $p > 0$. *Then for all* $T \geq 0$ *and* $u \geq 0$,

$$\mathrm{E}\,(S(T)^u) < \infty.$$

A particular CEV-type stochastic volatility model that has gained popularity with many practitioners is the so-called *SABR model*, see Hagan et al. [2002]. In Hagan et al. [2002], the SABR model is defined as

$$dS(t) = S(t)^c u(t)\, dW(t), \tag{8.45}$$
$$du(t) = \nu u(t)\, dZ(t), \tag{8.46}$$

with $\langle dW(t), dZ(t)\rangle = \rho\, dt$ and $0 < c < 1$. Note that the stochastic volatility $u(\cdot)$ is here modeled as simple geometric Brownian motion with zero drift. To translate the SDE (8.45)–(8.46) into more familiar terms, set $u(t) = \lambda\sqrt{z(t)}$, where $\lambda = u(0)/\sqrt{z_0}$. Then, with $\eta = 2\nu$,

$$dS(t) = \lambda S(t)^c \sqrt{z(t)}\, dW(t),$$
$$dz(t) = \frac{1}{4}\eta^2 z(t)\, dt + \eta z(t)\, dZ(t).$$

We recognize this as a special case of our set-up (8.1)–(8.2) with $\psi(z) = z$, $m(t) = 0$, and *negative* mean reversion speed $\theta = -\eta^2/4$. The drift term in the process for $z(\cdot)$ is rather unattractive but allows for some tractability, as we shall see below. While higher-order moments can be very large in the SABR model, it follows from Proposition 8.6.1 that all positive moments of $S(t)$ exist (the fact that the mean reversion is negative can be shown to not influence the result in the proposition). Notice also that in the SABR model $S(\cdot)$ cannot go negative (although absorption at zero is a possibility) and that the variance process is strictly positive.

The main justification for the form of the equations (8.1)–(8.2) is that it allows for fairly accurate asymptotic expansions for European option prices. Hagan et al. [2002] obtained the first such expansion result by combining classical perturbation methods with, in the words of Obloj [2008], "impressive intuition". Still, the result in Hagan et al. [2002] suffers from an internal inconsistency as $c \to 1$ and has later been revised by authors relying on more formal approaches. The result we list below is proven in Obloj [2008], based on earlier theoretical results in Berestycki et al. [2004] and Henry-Labordére [2005]. A similar result has been proven by Osajima [2007], using the small-noise expansion technique that we employed in Section 7.6.3.

Proposition 8.6.2. *For the model (8.45)-(8.46), the implied volatility smile is*

$$\sigma_{\mathrm{B}}(t, S(t); K, T) = I^0 \left(1 + (T-t)I^1\right) + O\left((T-t)^2\right),$$

where

$$I^0 = \frac{-\nu \ln\left(K/S(t)\right)}{\ln\left(\frac{\sqrt{1-2\rho q+q^2}+q-\rho}{1-\rho}\right)}, \quad q = \frac{\nu}{u(t)} \frac{S(t)^{1-c} - K^{1-c}}{1-c},$$

$$I^1 = \frac{(c-1)^2}{24} \frac{u(t)^2}{(S(t)K)^{1-c}} + \frac{1}{4} \frac{\rho\nu u(t) c}{(S(t)K)^{(1-c)/2}} + \frac{2-3\rho^2}{24}\nu^2.$$

Due to its lack of a mean reversion parameter, the SABR model often has difficulty matching smiles at different maturities when only a single set of calibration parameters $(\nu, c, \rho, u(0))$ is used. In practice, many financial institutions therefore maintain T-indexed vectors of these parameters, using the model primarily as a tool to interpolate and extrapolate the volatility smile. Some care must be exercised here, since the expansion listed above is not necessarily arbitrage-free; indeed, it is known that the expansion above may imply negative state price densities for low strikes and large maturities[11]. These issues could potentially be rectified by ad-hoc methods for modifying the density, see Section 16.9 for an example.

8.7 Numerical Examples: Volatility Smile Statics

Having established a valuation formula for European options in the SV model, let us proceed to put it to work on some concrete model parameterizations. In doing so, we pay special attention to the way the various parameters of the SV model effect the implied volatility smile $\sigma_B(0, S(0); K, T)$. The results here provide additional color to the qualitative parameter discussion in Section 8.2. To aid our discussion, we start by listing a small-T expansion for the implied volatility of the SV model. The expansion is not particularly precise for medium and long-dated securities, but it suffices for the largely qualitative analysis in this section. As the expansion relies on techniques that we discuss in detail later (in Section 9.2) we skip the proof and also omit, for now, a precise characterization of the approximation convergence as $T \to 0$.

Lemma 8.7.1. *Define log-moneyness* $\chi \triangleq \ln(K/S(0))$ *and consider writing the implied Black volatility as*

[11] Relative to the original SABR expansion in Hagan et al. [2002], the expansion in Proposition 8.6.2 is more robust in the low-strike tail; see Obloj [2008] for some numerical comparisons.

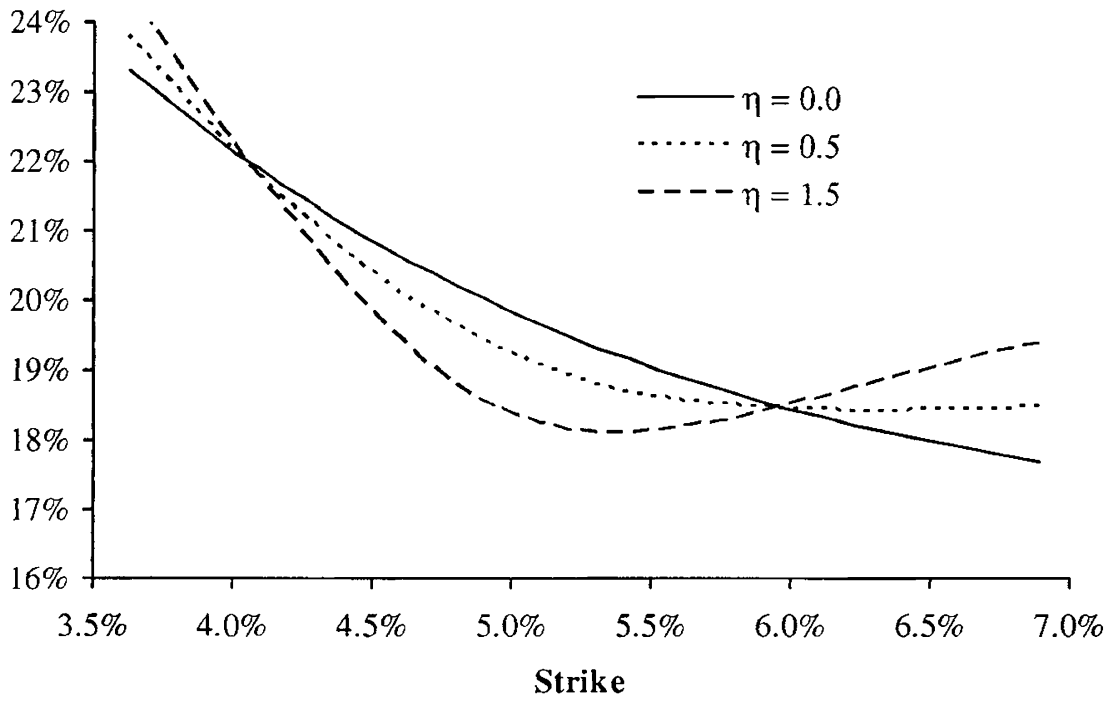
$$\sigma_{\mathrm{B}}(0, S(0); T, K) = \sigma_{\mathrm{ATM}} + R \cdot \chi + \frac{1}{2} B \cdot \chi^2 + \dots$$

for certain constants R and B. For small T and small χ, in the SV model (8.3)-(8.4) with $L = S(0)$ we have

$$\sigma_{\mathrm{ATM}} \approx \lambda, \quad R \approx \frac{\lambda}{2}\left(-(1-b) + \frac{\eta\rho}{2\lambda}\right),$$
$$B \approx \lambda\left(\frac{1-b^2}{6} + \frac{\eta^2\left(2-5\rho^2\right)}{24\lambda^2}\right).$$

Armed with Lemma 8.7.1, we start out with an example of how the volatility of variance parameter η affects the convexity of the volatility smile. As discussed previously, η serves to generate convexity in the volatility smile, an effect that is obvious from the approximation for B in Lemma 8.7.1 and also clearly visible in Figure 8.1.

Fig. 8.1. 1 Year Volatility Smile

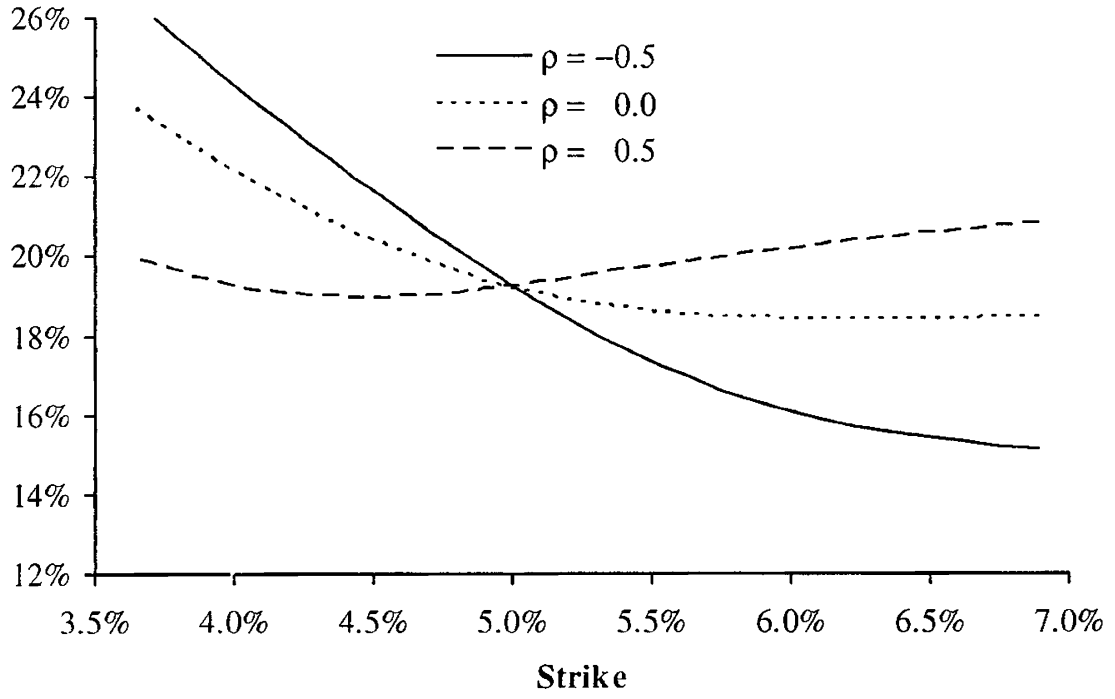


Notes: Implied volatility smile for SV model with $T = 1$, $S(0) = L = 5\%$, $z_0 = 1$, $b = 0.1$, $\lambda = 20\%$, $\theta = 0.1$, and $\rho = 0$. The volatility of variance parameter η varies as shown in the graph.

In Figure 8.1, the variance process is uncorrelated to the rate process, whereby Lemma 8.7.1 tells us that the slope (or skew) of the volatility smile at the at-the-money strike (5%) is generated solely by the slope parameter $b = 0.1$ in the local volatility function of the SV model. The stochastic volatility process can, of course, contribute to the skew if we use non-zero

correlation; see Figure 8.2 for a numerical example. As expected, lowering correlation rotates the smile clockwise, qualitatively similar to the impact of b. Another effect is also evident in Figure 8.2: when ρ moves away from zero, the convexity of the smile around the ATM strike is reduced. This effect is consistent with the expression for B in Lemma 8.7.1 which shows that the convexity (approximately) scales with[12] $2 - 5\rho^2$.

Fig. 8.2. 1 Year Volatility Smile



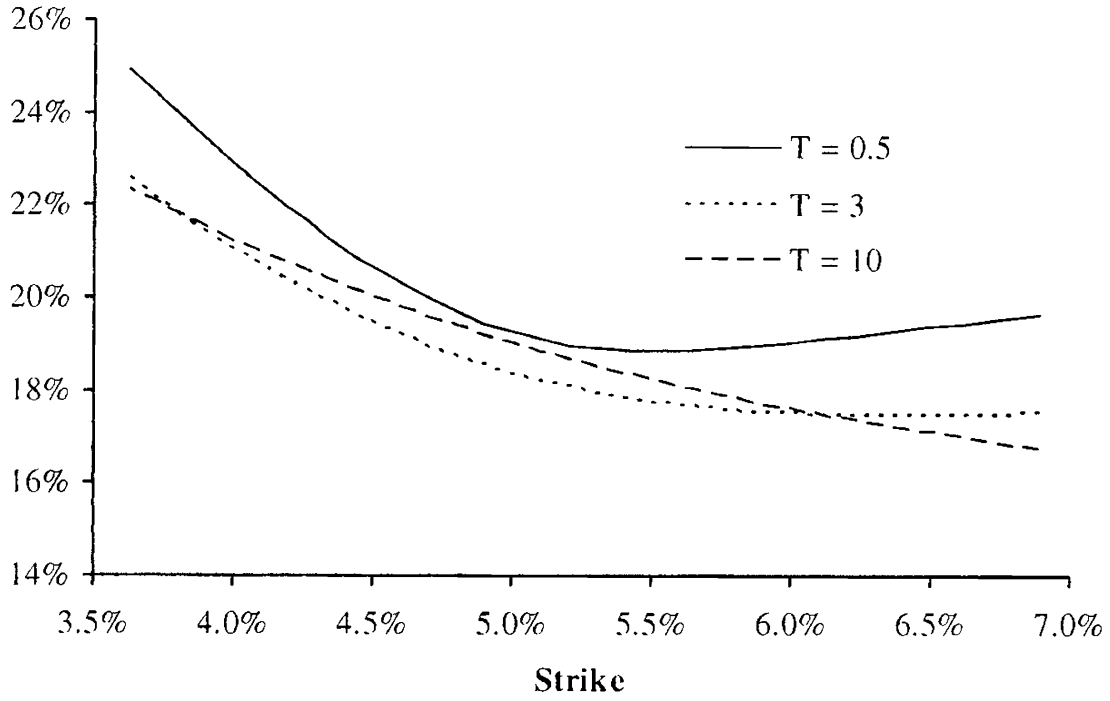
Notes: Implied volatility smile for SV model with $T = 1$, $S(0) = L = 5\%$, $z_0 = 1$, $b = 0.1$, $\lambda = 20\%$, $\theta = 0.1$, and $\eta = 1$. The correlation parameter ρ varies as shown in the graph.

The examples shown in Figures 8.1 and 8.2 both list the 1 year volatility smile only. To examine how the volatility smile $\sigma_B(0, S(0); K, T)$ in the SV model depends on T, consider first the case where $\rho = 0$; representative data are shown in Figure 8.3. The convexity of the smile, which originates with the stochastic volatility process, here clearly decays away as maturity is increased. As hinted at by Lemma 8.5.1, the convexity of the smile at time T is roughly proportional to the variance of the normalized realized variance $T^{-1} \int_0^T z(t)\, dt$. The convexity decay can therefore be interpreted as a mean reversion effect, since the variance of the normalized realized variance itself decays to a long-term (stationary) level, as can be seen from Corollary 8.3.3.

[12]Indeed, according to Lemma 8.7.1 the (short-maturity) smile convexity originating from stochastic volatility can become *negative* is $|\rho| > \sqrt{2/5} \approx 0.632$. This is easily verified numerically.

The speed of the decay is controlled by manipulating mean reversion speed θ; the higher θ is, the quicker the smile convexity decays in the T-direction.

Fig. 8.3. Term Structure of Volatility Smiles



Notes: Implied volatility smile for SV model with $S(0) = L = 5\%$, $z_0 = 1$, $\rho = 0.0$, $\lambda = 20\%$, $\theta = 0.5$, $b = 0.1$, and $\eta = 1.5$. The smile maturity 'T' varies as shown in the graph.

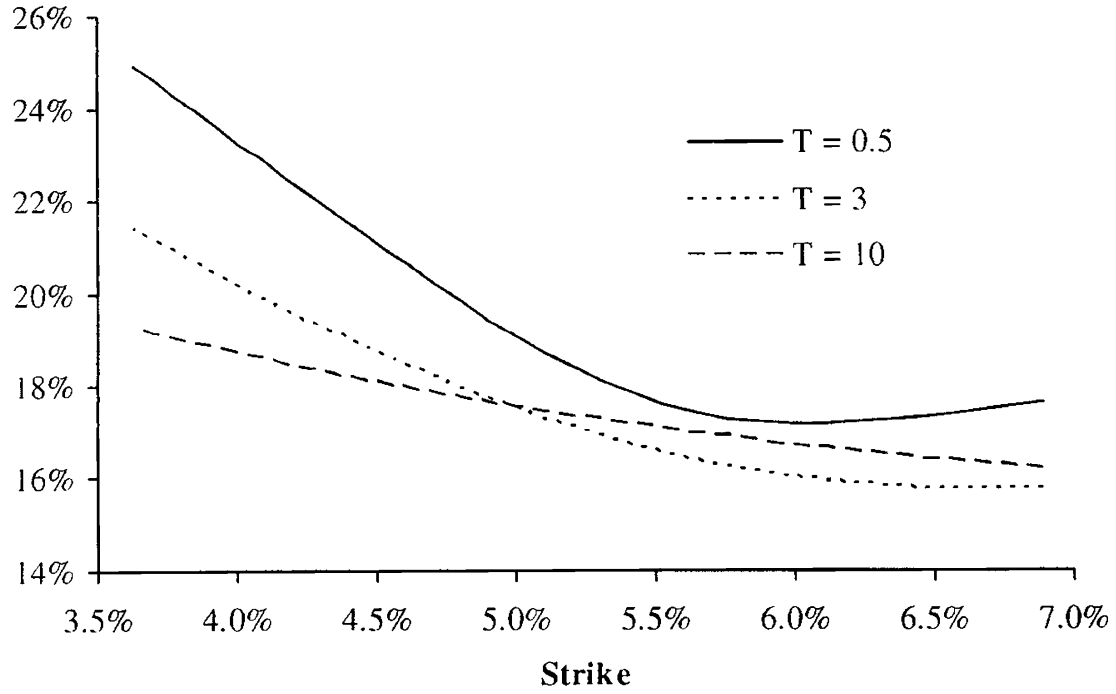
We note in passing that the ATM volatility of a constant parameter SV model is not a monotonic function of option maturity, as a quick glance at Figure 8.3 will confirm. For an analysis of the ATM volatility level and its dependence on maturity, see Lewis [2000].

In Figure 8.3 the slope of the smile around the ATM point is generated only from the parameter 'b' in the 'local volatility function and consequently shows little decay in T. If, on the other hand, we had used a negative variance-spot correlation to generate the skew, we would expect the volatility smile to flatten out in T, for the same reason that the smile convexity decays. Figure 8.4 confirms this intuition.

8.8 Numerical Examples: Volatility Smile Dynamics

As we mentioned earlier, one rationale for introducing stochastic volatility into an LV model is the desire to generate realistic *smile dynamics*. In Section 7.1.3, we listed some qualitative reasons for the failure of LV models to generate reasonable model dynamics in certain cases; we are now in a

Fig. 8.4. Term Structure of Volatility Smiles



Notes: Implied volatility smile for SV model with $S(0) = 5\%$, $z_0 = 1$, $\rho = -0.5$, $\lambda = 20\%$, $\theta = 0.5$, $b = 1$, and $\eta = 1.5$. The smile maturity T varies as shown in the graph.

position to expand on this discussion and to show some concrete results. Specifically, we here wish to compare how the volatility smile moves with the underlying rate process, for two models: i) an ordinary (log-normal) Heston model obtained by setting $b = 1$ in the SV model (8.3)–(8.4); and ii) a pure LV model with quadratic volatility,

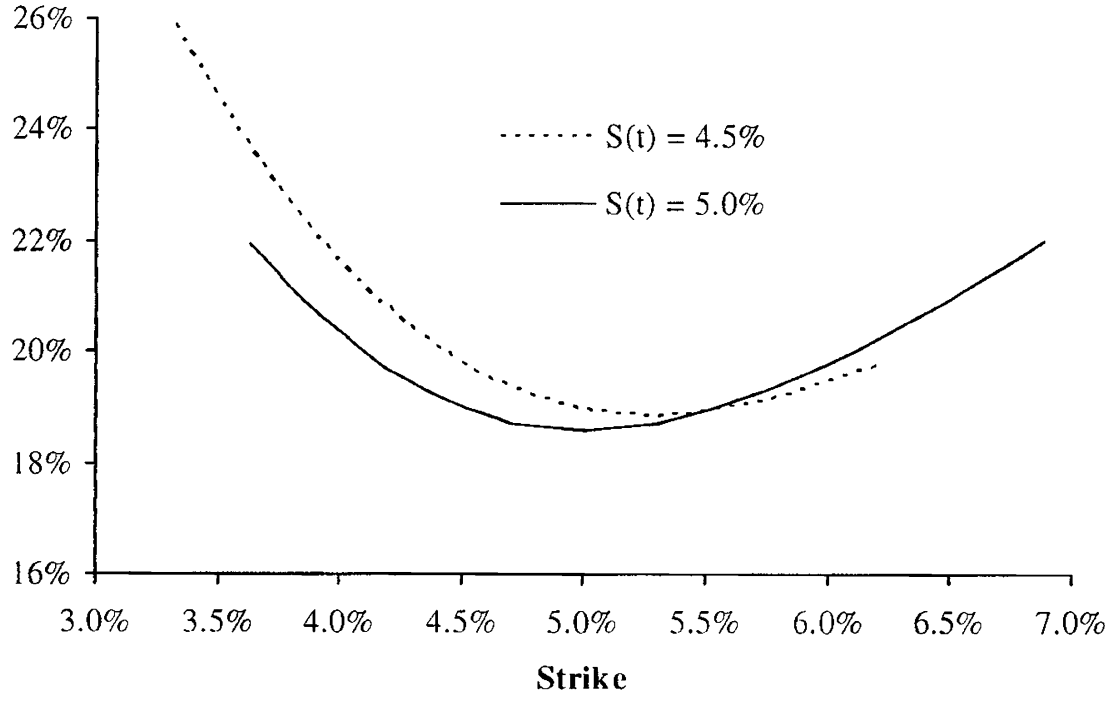
$$dS(t) = \lambda \left((1-b)L + bS(t) + \frac{1}{2}c(S(t)-L)^2 \right) dW(t). \tag{8.47}$$

For our numerical experiments, we move calendar time forward to some arbitrary value t and examine how the smile looks for several levels of $S(t)$. In performing this analysis for the Heston model, we shall initially assume that $z(t)$ stays equal to its initial value z_0, but we relax this assumption later.

First, we consider the case of a (near) symmetric smile which in the local volatility model (8.47) can be obtained by setting $b = 1$. The effect of a 50 bps downward move in $S(0)$ (i.e. $S(t) = S(0) - 0.5\%$) on a specific LV model is shown in Figure 8.5. Starting from a symmetric smile when the forward rate $S(t) = S(0) = 5\%$, a shift down to 4.5% causes an overall increase in volatility levels, as well as a clock-wise tilt of the previously symmetric smile. This is readily understood. as the quadratic local volatility function

will itself increase and loose its symmetry when $S(t)$ is reduced from 5% to 4.5%.

Fig. 8.5. Volatility Smile Dynamics in Quadratic LV Model

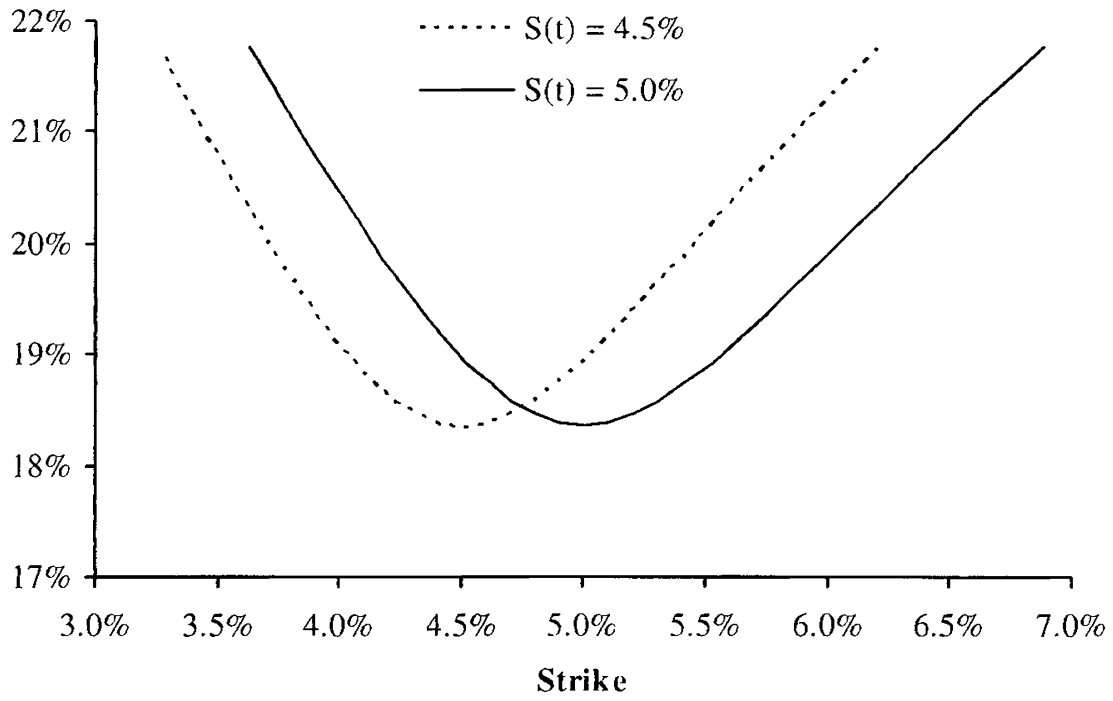


Notes: Time t implied volatility smile for quadratic LV model with $T = t + 1$, $S(0) = L = 5\%$, $b = 1$, $\lambda = 18\%$, and $c = 0.6$. Two different values for the forward rate $S(t)$ are used, as indicated in the graph.

Turning now to the Heston model, we first make the observation from Theorem 8.4.4 that European put and call option values normalized by spot S in both the Heston and Black models — and thereby the implied volatility smile of the Heston model — depend on strike K and forward rate $S(t)$ only through the ratio $K/S(t)$. Specifically, we have $\sigma_B(t, S(t); K, T) = g(K/S(t), T - t)$, for some function $g(\cdot, \cdot)$. In trader lingo, this is known as a "sticky delta" volatility smile[13], and implies that the $T = t + \Delta$ volatility smile expressed in moneyness $K/S(t)$, or log-moneyness $\ln(K/S(t))$, is independent of t and $S(t)$, as long as $z(t)$ remains unchanged at its initial value z_0. This fact makes it easy to construct the Heston model dynamics of the volatility smile in strike space; Figure 8.6 shows an example for a case where the correlation ρ has been set to zero to make the smile is symmetric in log-moneyness. Notice that as $S(t)$ drops from 5% to 4.5%, the volatility smile floats to the left, in tandem with the move in $S(t)$ such that the bottom of the smile remains centered at the forward rate.

[13] A reflection of the fact that the delta in the Black model, i.e. $\partial c_B/\partial S$, only depends on K/S.

Fig. 8.6. Volatility Smile Dynamics in Heston SV Model



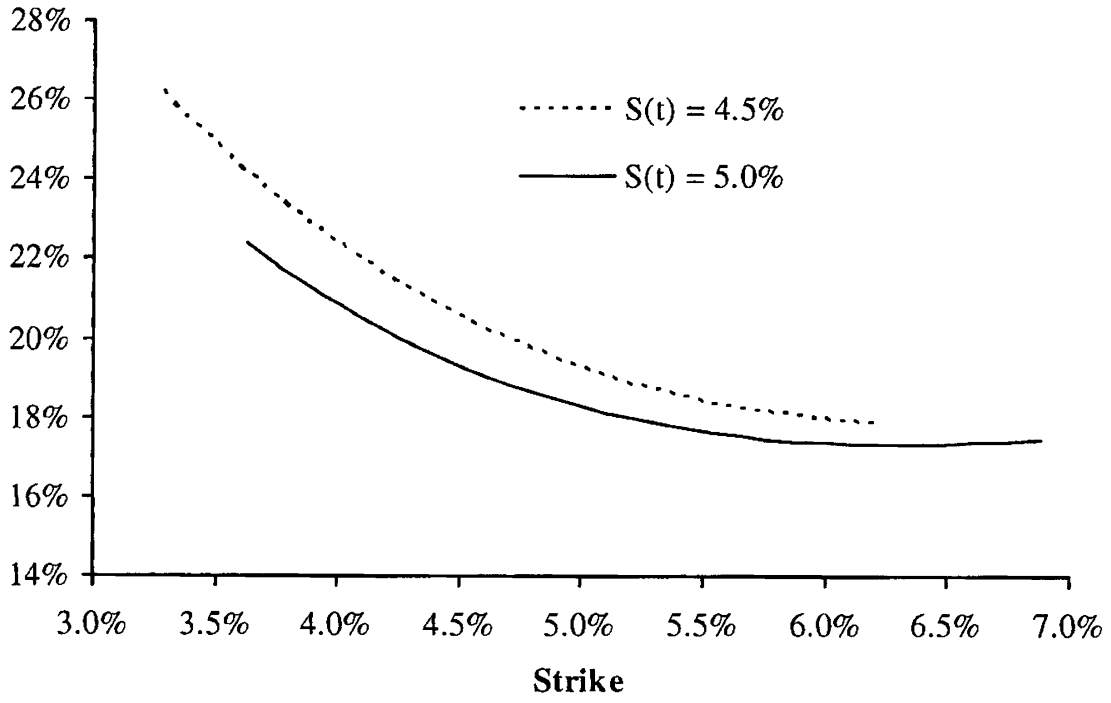
Notes: Time t implied volatility smile for SV model with $T = t + 1$, $S(0) = 5\%$, $z(t) = z_0 = 1$, $b = 1$, $\lambda = 20\%$, $\theta = 0.1$, $\rho = 0$, and $\eta = 1.5$. Two different values for the forward rate $S(t)$ are used, as indicated in the graph.

While Figures 8.5 and 8.6 are interesting and highlight some important differences between local and stochastic volatility models, it is more relevant in an interest rate setting to consider the case where the volatility smile has significant skew. First, we consider the local volatility case, see Figure 8.7. A shift down in $S(t)$ will increase the level of the local volatility function and raise the level of the smile; alternatively, we can interpret the move as a slide to the right. As convexity is relatively low in the graph relative to the skew, the move in $S(t)$ has little effect on the slope of the graph.

In Figure 8.8 we examine the smile dynamics of a Heston model with a significant downward skew, induced by a non-zero correlation ρ. The sticky-delta dynamics of the smile are still in effect here, causing a slide to the left when $S(t)$ is lowered, in a manner identical to that of the symmetric case in Figure 8.6.

The dynamics on display in Figures 8.7 and 8.8 appear to be diametrically opposite of each other: the smile shifts to the right in the local volatility model and to the left in the stochastic volatility model. In reality, however, differences in model dynamics are less dramatic than these graphs show. In particular, we recall that when we computed Figure 8.8, we kept $z(t)$ constant at the value z_0. However, as $z(t)$ and $S(t)$ are negatively correlated in the model used in Figure 8.8, keeping one process constant while the other moves will clearly be wrong "on average". A more representative

Fig. 8.7. Volatility Smile Dynamics in Quadratic LV Model



Notes: Time t implied volatility smile for quadratic LV model with $T = t + 1$, $S(0) = L = 5\%$, $b = 0.1$, $\lambda = 18\%$, and $c = 0.25$. Two different values for the forward rate $S(t)$ are used, as indicated in the graph.

characterization of the smile dynamics of the Heston process would move the variance process to its most likely outcome, given the move in the underlying. That is, we wish to set $z(t)$ equal to

$$\mathrm{E}\left(z(t)|S(t)\right)$$

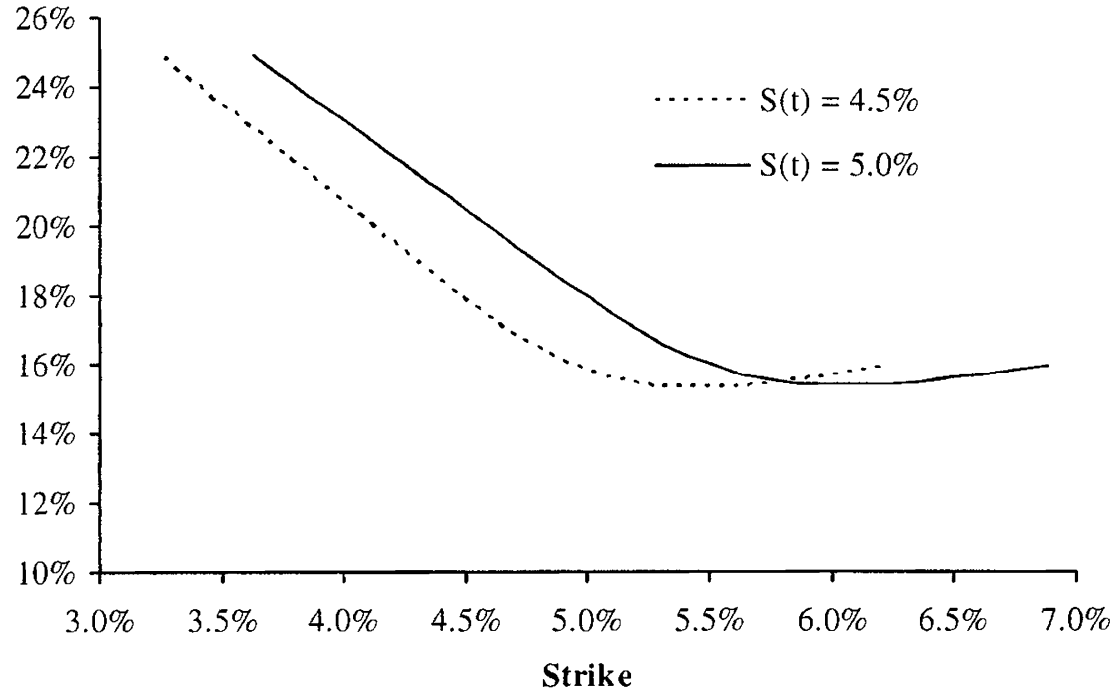
which we here compute by a simple Gaussian approximation that ignores mean reversion,

$$\mathrm{E}\left(z(t)|S(t)\right) \approx z_0 + \frac{\eta\rho}{\lambda}\frac{S(t) - S(0)}{S(0)}. \tag{8.48}$$

Performing this modification on the data in Figure 8.8 results in the data in Figure 8.9.

With the rule in (8.48), the volatility smile shift of Figure 8.8 has reversed direction in Figure 8.9 and now looks quite similar to that of the local volatility dynamics of Figure 8.7. In other words, for volatility smiles that are "skew-dominated", i.e. the skew is significant and the convexity is modest, smile dynamics of local and stochastic volatility models are quite similar on average. This observation is emphasized by Dupire [2006] and to some extent goes against common wisdom (see e.g. Hagan et al. [2002]) which tends to emphasize the sticky strike behavior of the stochastic volatility model. Of course, while the behavior in Figure 8.9 may be more likely

Fig. 8.8. Volatility Smile Dynamics in Heston SV Model



Notes: Time t implied volatility smile for SV model with $T = t + 1$, $z(t) = z_0 = 1$, $S(0) = 5\%$, $b = 1$, $\lambda = 20\%$, $\theta = 0.1$, $\rho = -0.6$, and $\eta = 1.5$. Two different values for the forward rate $S(t)$ are used, as indicated in the graph.

than that of Figure 8.8, both are feasible in a stochastic variance setting, depending on what value $z(t)$ happens to take. For derivatives that have convexity with respect to volatility smile moves[14], what most reasonably represents "average" smile behavior is obviously less important than the fact that variance is random.

We finish this section by noting that the ideas behind (8.48) are also relevant for hedge construction in presence of stochastic volatility. We return to this topic in Section 8.9.2.

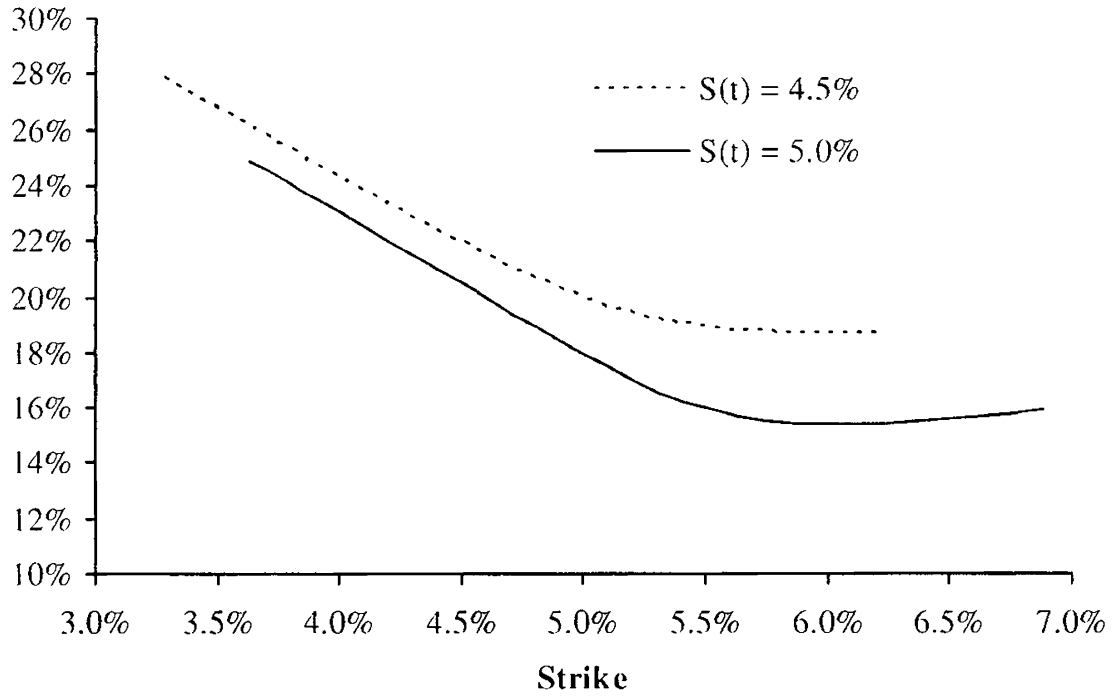
8.9 Hedging in Stochastic Volatility Models

8.9.1 Hedge Construction, Delta and Vega

Having now treated the subject of option pricing with stochastic volatility in quite some detail, let us make a foray into the topic of hedge construction. With their two generally non-collinear sources of randomness W and Z, it

[14] An option on implied volatility is an obvious example, although somewhat esoteric in an interest rate setting. A fairly common interest rate product with some volatility convexity is a barrier option. Many examples exist in other asset classes, such as reverse cliquets and Napoleons, see Jeffery [2004].

Fig. 8.9. Volatility Smile Dynamics in Heston SV Model



Notes: Time t implied volatility smile for the SV model in Figure 8.8, but now with $z(t)$ set as computed from formula (8.48).

should be clear that stochastic volatility models of the type (8.1)–(8.2) are not complete (in the sense defined in Section 1.4) if we limit ourselves to simple delta hedging with positions only in $S(t)$ itself. However, if options with volatility sensitivity are available for trading, these can be included into the hedge portfolio to complete the market.

Assuming general dynamics (8.1)–(8.2), we proceed to consider hedging of a contingent claim $V(t)$ that depends on both $S(t)$ and $z(t)$, i.e. we write $V(t) = V(t, S(t), z(t))$. We assume existence of two traded securities $U_1(t) = U_1(t, S(t), z(t))$ and $U_2(t) = U_2(t, S(t), z(t))$. Using the framework of Section 1.7, we associate $U_1(t)$, $U_2(t)$ with the elements of the asset vector $X(t)$ from that section. Forming a hedging portfolio Π consisting of $-\pi_1(t)$ units of $U_1(t)$ and $-\pi_2(t)$ units of $U_2(t)$, we obtain from (1.26) that

$$\pi_i = \frac{\partial V}{\partial U_i}, \quad i = 1, 2.$$

A bit of calculus leads us to expressions for the hedge ratios in terms of sensitivities to the primitives S, z of the model, and the following result follows.

Lemma 8.9.1. *The portfolio* $\Pi(t) = V(t) - \pi_1(t)U_1(t) - \pi_2(t)U_2(t)$ *is locally riskless if*

$$\pi_1 = \left(\frac{\partial V}{\partial S}\frac{\partial U_2}{\partial z} - \frac{\partial U_2}{\partial S}\frac{\partial V}{\partial z}\right)\left(\frac{\partial U_1}{\partial S}\frac{\partial U_2}{\partial z} - \frac{\partial U_2}{\partial S}\frac{\partial U_1}{\partial z}\right)^{-1}, \tag{8.49}$$

$$\pi_2 = \left(\frac{\partial V}{\partial S}\frac{\partial U_1}{\partial z} - \frac{\partial U_1}{\partial S}\frac{\partial V}{\partial z}\right)\left(\frac{\partial U_2}{\partial S}\frac{\partial U_1}{\partial z} - \frac{\partial U_1}{\partial S}\frac{\partial U_2}{\partial z}\right)^{-1}. \tag{8.50}$$

Remark 8.9.2. In practice, the first security U_1 would often be chosen to not depend on z — for example the swap from which $S(t)$ is computed could be used as U_1 — in which case the hedge weights simplify. In particular,

$$\pi_1 = \frac{\partial V/\partial S}{\partial U_1/\partial S} - \frac{\partial V/\partial z}{\partial U_2/\partial z}\frac{\partial U_2/\partial S}{\partial U_1/\partial S}, \quad \pi_2 = \frac{\partial V/\partial z}{\partial U_2/\partial z},$$

as one would expect.

Remark 8.9.3. The sensitivity of a given security to volatility is often called its *vega*. Even for a model with non-stochastic volatility, such as the Black model, a vega can be computed, but will not enter the hedge balance equation (1.28). In a stochastic volatility model, a vega can conveniently[15] be defined to be $\partial/\partial z$ — which *will* enter the hedge balance equation. It follows that the choice (8.49)–(8.50) ensures that the hedged portfolio Π is delta-neutral, in the sense that

$$\frac{\partial \Pi(t)}{\partial S} = 0,$$

as well as *vega-neutral,*

$$\frac{\partial \Pi(t)}{\partial z} = 0.$$

8.9.2 Minimum Variance Delta Hedging

While the theoretical notion of "delta" assumes that the stochastic variance process z is kept fixed under perturbations of S, we saw earlier in Section 8.8 (see, in particular, Figure 8.9 and the discussion around it) that it sometimes might be more natural to let z float along with S, in a manner determined by the correlation between these quantities. Indeed, to the extent that our hedging strategy were to employ a position in S only, and not to separately hedge the exposure to z, the "best" hedging strategy — in the sense of locally minimizing hedging errors — is one based on such a joint move in z and S. We proceed to present this idea, using rather ad-hoc (or "deceptively simple", to paraphrase Ewald et al. [2007]) techniques; for a full account and for a connection to the concept of the *minimal martingale measure*, see Follmer and Schweizer [1990] and Ewald et al. [2007].

First, let us return to the model (8.1)–(8.2), but now use a Cholesky decomposition to rewrite the process for $z(t)$ as

[15] From a theoretical viewpoint. More practical definitions of vega are covered later in the book, see Chapter 26 in particular.

$$dz(t) = O(dt) + \sigma_z(t)\left(\rho\, dW(t) + \sqrt{1-\rho^2}\, dB(t)\right),$$

where B is a Brownian motion that is *independent* of W, and we use $\sigma_z(t) = \eta\psi(z(t))$ and $\sigma_S(t) = \lambda\varphi(S(t))\sqrt{z(t)}$ for notational clarity. Consider now a claim

$$V(t) = V(t, S(t), z(t)),$$

where, by Ito's lemma,

$$dV(t) = O(dt) + \frac{\partial V(t)}{\partial S}\, dS(t) + \frac{\partial V(t)}{\partial z}\sigma_z(t)\left(\rho\, dW(t) + \sqrt{1-\rho^2}\, dB(t)\right).$$

Let us form a portfolio Π of the claim V and a position of $-\pi(t)$ in $S(t)$; that is,

$$d\Pi(t) = -\pi(t)\, dS(t) + dV(t). \tag{8.51}$$

We wish to set $\pi(t)$ such that $\mathrm{Var}_t(d\Pi(t))$ is minimized.

Lemma 8.9.4. *With $d\Pi(t)$ defined in (8.51), the variance $\mathrm{Var}_t(d\Pi(t))$ is minimized by setting $\pi(t) = \pi_{\mathrm{mv}}(t)$, where*

$$\pi_{\mathrm{mv}}(t) = \frac{\partial V(t)}{\partial S} + w(t), \quad w(t) = \frac{\partial V(t)}{\partial z}\frac{\rho\sigma_z(t)}{\sigma_S(t)},$$

and $\sigma_z(t) = \eta\psi(z(t))$, $\sigma_S(t) = \lambda\varphi(S(t))\sqrt{z(t)}$.

Proof. It is easily seen that

$$\mathrm{Var}_t(d\Pi(t)) = \left(-\pi(t)\sigma_S(t) + \frac{\partial V(t)}{\partial S}\sigma_S(t) + \frac{\partial V(t)}{\partial z}\sigma_z(t)\rho\right)^2 dt + \left(\frac{\partial V(t)}{\partial z}\right)^2 \sigma_z(t)^2\left(1-\rho^2\right) dt.$$

The first-order condition for the minimum is therefore

$$0 = -2\sigma_S(t)\left(-\pi(t)\sigma_S(t) + \frac{\partial V(t)}{\partial S}\sigma_S(t) + \frac{\partial V(t)}{\partial z}\sigma_z(t)\rho\right),$$

from which the lemma follows. □

We notice that $w(t)$ in Lemma 8.9.4 can be written informally as

$$w(t) = \frac{\partial V(t)}{\partial z}\frac{\mathrm{E}_t\left(dz(t)|dS(t) = dS\right)}{dS}$$

which shows that the *minimum-variance* (MV) hedge ratio is obtained, in effect, by moving the z-process to its expected value, given an infinitesimal perturbation in the S-process. In other words, the hedge represents our best guess for a position in the underlying that will hedge moves in $V(t)$ caused by changes in *both* $S(t)$ and $z(t)$, as in Figure 8.9.

To further characterize the properties of the MV hedge weight, we insert the result of Lemma 8.9.4 into (8.51), which yields

$$d\Pi(t) = O(dt) + \frac{\partial V(t)}{\partial z}\sigma_z(t)\sqrt{1-\rho^2}\,dB(t).$$

In other words, the MV hedge produces a portfolio that is not exposed to $W(t)$ but only to the orthogonal Brownian motion $B(t)$. If one thinks of $W(t)$ as "market" noise, we can say — in the language of the classical CAPM[16] analysis — that the hedged portfolio has no *beta*. For this reason, the hedge construction in Lemma 8.9.4 is also sometimes known as a *zero-beta hedge*.

8.9.3 Minimum Variance Hedging: an Example

To better understand the practical ramifications of MV hedging, let us do a concrete example based on the SABR model from Section 8.6, which we here parameterize as

$$\begin{aligned} dS(t) &= \lambda\sqrt{z(t)}S(t)^c\,dW(t), \\ dz(t) &= \frac{1}{4}\eta^2 z(t)dt + \eta z(t)\left(\rho\,dW(t) + \sqrt{1-\rho^2}\,dB(t)\right), \quad z(0) = 1. \end{aligned}$$

According to Lemma 8.9.4, the MV hedge ratio in SABR is

$$\pi_{\mathrm{mv}}(t) = \frac{\partial V(t)}{\partial S} + \eta\sqrt{z(t)}\rho\frac{\partial V(t)}{\partial z}\frac{1}{\lambda S(t)^c}.$$

In a typical interest rate application $z(t) \approx 1$, $\lambda S(t)^c \approx 0.01$ and $\eta \approx 1$, such that, as a rule of thumb,

$$\pi_{\mathrm{mv}}(t) \approx \frac{\partial V(t)}{\partial S} + 100 \times \rho\frac{\partial V(t)}{\partial z}.$$

For call and put options, the hedge adjustment to the "pure" delta $\partial V/\partial S$ is here typically negative, as we have $\partial V/\partial z > 0$ and, in normal market conditions, $\rho < 0$. This is consistent with Figure 8.8.

We now perform the following small experiment: we lock the correlation parameter at a pre-fixed value and then least-squares calibrate the SABR model to an actual market Black volatility smile. For a range of correlation parameters, we then compute "pure" deltas ($\partial V/\partial S$) and MV deltas (π_{mv}) for swaptions with different strikes. Using market data roughly consistent with the 5y×5y swaption volatility smile in the summer of 2005, the calibration results are in Table 8.1.

As one would expect, making correlation progressively more negative causes the skew power c to increase, from about 20% at $\rho = 0$ to nearly

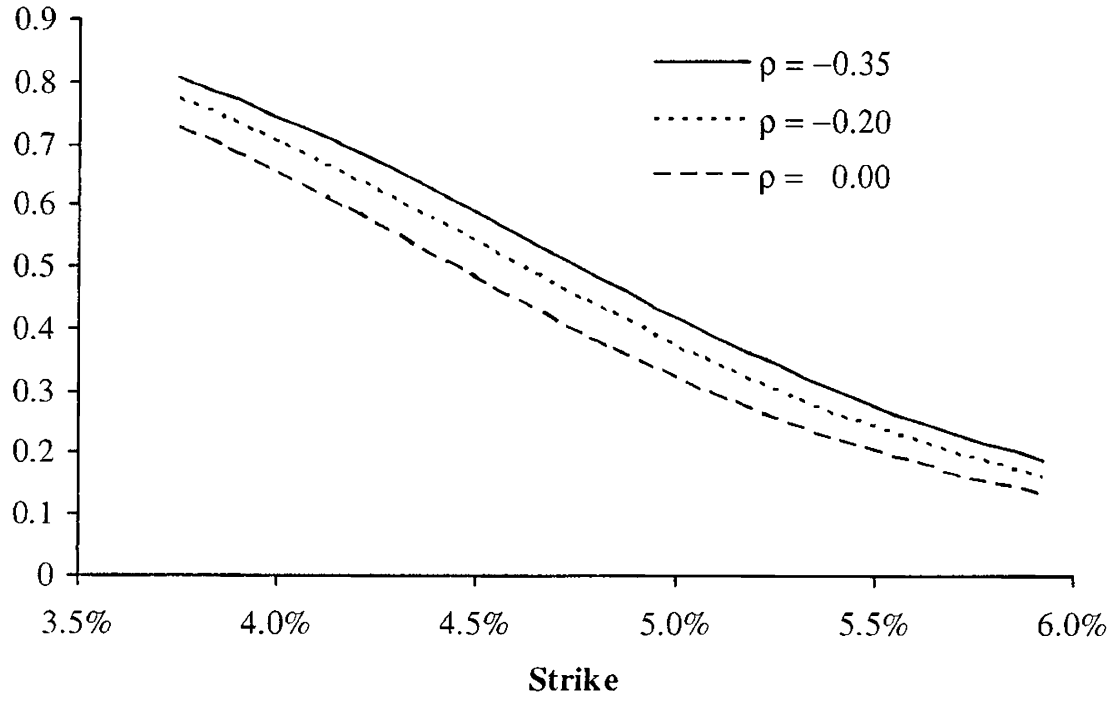
[16] Capital Asset Pricing Model, see Sharpe [1964].

ρ	0	-0.1	-0.2	-0.3	-0.35
$\lambda S(0)^{1-c}$	0.135	0.136	0.137	0.139	0.140
c	0.223	0.432	0.648	0.877	0.999
η	0.684	0.686	0.696	0.712	0.726

Table 8.1. SABR Calibration Results

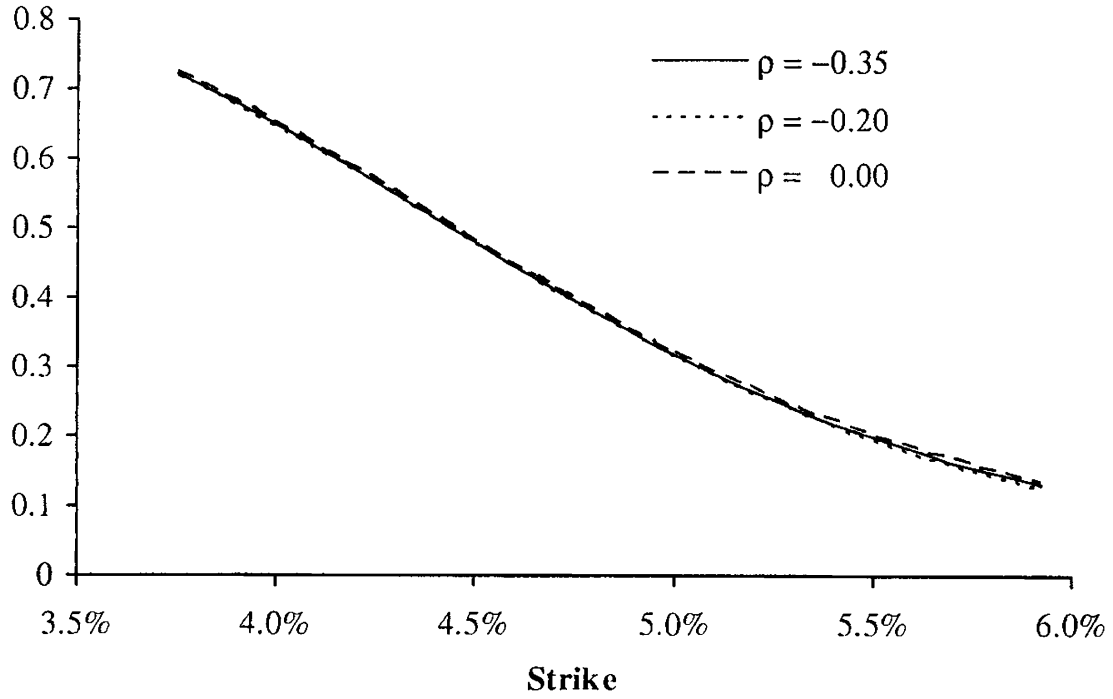
90% at $\rho = -0.3$, with other parameters being quite stable across different correlation choices. Figures 8.10 and 8.11 show the pure delta $\partial V/\partial S$ and the minimum variance delta π_{mv} for selected strikes and correlations. Clearly, the MV delta is here virtually independent of the choice of ρ, whereas the pure delta can increase quite substantially as correlation becomes more negative. It is clear from the figures that as long as hedge ratios are computed to be MV deltas, rather than pure deltas, the precise blend of local and stochastic volatility may not be critical, at least not for vanilla-like options in a skew-dominated market. This confirms a point we made earlier, in Section 8.1.

Fig. 8.10. Pure Delta



Notes: The figure shows the pure delta for the SABR models in Table 8.1.

Fig. 8.11. Minimum Variance Delta



Notes: The figure shows the minimum variance (MV) delta for the SABR models in Table 8.1.

8.A Appendix: Martingale Characterization, Moment Stability, and Other Fundamental Properties for General Variance Processes

As explained in Section 8.3, it is sometimes beneficial to consider a specification of the stochastic volatility model that is more general than (8.3)–(8.4). Let us consider a general power function for $\psi(z)$ in (8.2),

$$dS(t) = \lambda\left(bS(t) + (1-b)\,L\right)\sqrt{z(t)}\,dW(t), \tag{8.52}$$

$$dz(t) = \theta\left(z_0 - z(t)\right)\,dt + \eta z(t)^p\,dZ(t), \tag{8.53}$$

with $\langle dZ(t), dW(t)\rangle = \rho\,dt$. We assume $p > 0$. In this section we briefly outline important properties of such models. For more comprehensive treatment the reader is referred to Andersen and Piterbarg [2007]. Our first result spells out the boundary behavior of the stochastic variance process.

Proposition 8.A.1. *For the process (8.53), the following holds:*

1. *0 is always an attainable boundary for* $0 < p < 1/2$.
2. *0 is an attainable boundary for* $p = 1/2$, *if* $2z_0\theta < \eta^2$.
3. *0 is an unattainable boundary for* $p > 1/2$.
4. ∞ *is an unattainable boundary for all values of* $p > 0$.

When $0 < p < 1/2$, the origin is always accessible and we need to impose a boundary condition at $z = 0$ to make the process unique. To ensure that the process for $z(\cdot)$ has a stationary distribution, we make the following natural choice:

Assumption 8.A.2. *For $0 < p < 1/2$, the process (8.53) for $z(\cdot)$ is reflected at the origin.*

The marginal one-dimensional distribution of $z(t)$ can in principle be computed numerically by various methods, such as PDE methods or by Fourier inversion of a characteristic function. It is often convenient, however, to have an easily-computable approximation. For that purpose, a stationary distribution, if one exists, can be useful. A stationary distribution for $z(\cdot)$ does indeed exist and can be easily computed.

Proposition 8.A.3. *Let $\pi(y)$ be the stationary distribution density for $z(\cdot)$ in (8.53). Under the assumptions listed above,*

$$\pi(y) = C(p) y^{-2p} e^{Q(y;p)}, \quad C(p)^{-1} \triangleq \int_0^\infty y^{-2p} e^{Q(y;p)} dy,$$

where the function $Q(y;p)$ is given by

1. *$0 < p < 1/2$ or $1/2 < p < 1$ or $p > 1$:*

$$Q(y;p) = \frac{2\theta}{\eta^2} \left(\frac{z_0 y^{1-2p}}{1-2p} - \frac{y^{2-2p}}{2-2p} \right).$$

2. *$p = 1/2$:*

$$Q(y;p) = \frac{2\theta}{\eta^2} \left(z_0 \ln y - y \right).$$

3. *$p = 1$:*

$$Q(y;p) = \frac{2\theta}{\eta^2} \left(-z_0/y - \ln y \right).$$

A-priori, $S(\cdot)$ defined by (8.52)–(8.53) is only a local martingale. In fact, under some circumstances, $S(\cdot)$ is a strict local martingale, , usually a significant technical complication. Specifically, we have the following result.

Proposition 8.A.4. *When $p \leq 1/2$ or $p > 3/2$, $S(\cdot)$ is a proper martingale. When $1/2 < p < 3/2$, $S(\cdot)$ is a martingale for $\rho \leq 0$ and a strict supermartingale for $\rho > 0$. For $p = 3/2$, $S(\cdot)$ is a martingale for $\rho \leq \frac{1}{2}\eta(\lambda b)^{-1}$ and a strict supermartingale for $\rho > \frac{1}{2}\eta(\lambda b)^{-1}$.*

What this proposition states is that the set of parameters $1/2 < p < 3/2$, $\rho > 0$, should be avoided in practical modeling. The SV model (8.3)–(8.4), as already noted, has no issues in this regard. If we use $\rho = 0$ — a typical choice in interest rate modeling as explained previously — all values of p

between 0 and 3/2 are acceptable, at least as far as the martingale property is concerned.

In the model with $p = 1/2$, some moments of $S(\cdot)$ can become infinite, as stated in Proposition 8.3.10. With $p < 1/2$, this is no longer an issue:

Proposition 8.A.5. *In the model (8.52)–(8.53), if* $p < 1/2$, *moments* $\mathrm{E}(S(T)^u)$ *of all orders* $u \geq 1$ *for all times* T *are finite.*

On the other hand, if $p > 1/2$ moments may be unstable. For instance:

Proposition 8.A.6. *In the model (8.52)–(8.53), if* $p > 1/2$ *and* $\rho = 0$, *all moments* $\mathrm{E}(S(T)^u)$ *of all orders* $u > 1$ *for all times* T *are infinite.*

The case of non-zero correlation and $p > 1/2$ is more complicated; we refer the reader to Andersen and Piterbarg [2007].

9

Vanilla Models with Stochastic Volatility II

Having covered stochastic volatility models with time-homogeneous dynamics in Chapter 8, we are now ready to proceed with an analysis of the time-dependent case. As we shall see many examples of later in this book, stochastic volatility models with time-dependent parameters emerge naturally when vanilla models are used to approximate interest rate dynamics in a full term structure model.

In this chapter, we start out by modifying the Fourier analysis of Chapter 8 to cover time-dependent model parameters. We then proceed to introduce several approximation techniques that can speed up the calibration of model parameters to observable option prices. In particular, we continue our development of parameter averaging techniques, extending their scope to cover stochastic volatility and outlining in detail their usage in model calibration. Finally, the chapter gives detailed coverage of PDE and MC methods for general derivatives pricing; both of these numerical techniques are, as it turns out, rather tricky to apply to models with stochastic volatility, and an efficient implementation requires careful attention to detail.

9.1 Fourier Integration with Time-Dependent Parameters

As a start, let us consider extending the basic SV model (8.3)–(8.4) to allow for time-dependence of the volatility parameter[1] λ. That is, we now consider the P-measure dynamics

$$dS(t) = \lambda(t)\left(bS(t) + (1-b)\,L\right)\sqrt{z(t)}\,dW(t), \tag{9.1}$$

$$dz(t) = \theta\left(z_0 - z(t)\right)\,dt + \eta\sqrt{z(t)}\,dZ(t), \tag{9.2}$$

[1] A further extension to time-dependence in η, ρ, and θ is trivial, and is covered in Remark 9.1.3.

where $\langle dZ(t), dW(t)\rangle = \rho\, dt$.

The model (9.1)–(9.2) still allows for call option pricing by the Fourier integration method of Section 8.4, provided that we can establish the moment-generating function (mgf) of $\ln X(t)$, with $X(t)$ being the linear function of $S(t)$ defined in Proposition 8.3.6. Let us retain the notation $\Psi_X(u;t)$ for

$$\Psi_X(u;t) = \mathrm{E}\left(e^{u \ln X(t)}\right),$$

where the process for $X(t)$ now is modified from that of Proposition 8.3.6 to include time-dependence in λ:

$$dX(t)/X(t) = b\lambda(t)\sqrt{z(t)}\, dW(t), \quad X(0) = 1.$$

The following counterpart to Proposition 8.3.7 is easily proven.

Proposition 9.1.1. *In the model (9.1)–(9.2), for any $u \in \mathbb{C}$ for which the right-hand side exists, we have*

$$\Psi_X(u;t) = \Psi_{\overline{z\lambda^2}}\left(\frac{1}{2}b^2 u(u-1), u; t\right),$$

where we have defined

$$\Psi_{\overline{z\lambda^2}}(v,u;t) \triangleq \mathrm{E}^{\widetilde{\mathrm{P}}}\left(e^{v\overline{z\lambda^2}(t)}\right), \quad \overline{z\lambda^2}(t) \triangleq \int_0^t z(s)\lambda(s)^2\, ds, \tag{9.3}$$

and under the new probability measure $\widetilde{\mathrm{P}}$ the process for $z(t)$ is

$$dz(t) = \left(\theta\left(z_0 - z(t)\right) + \rho\eta\lambda(t)buz(t)\right) dt + \eta\sqrt{z(t)}\, d\widetilde{Z}(t), \quad z(0) = z_0, \tag{9.4}$$

with $\widetilde{Z}(t)$ a $\widetilde{\mathrm{P}}$-Brownian motion. If $\rho = 0$, $\widetilde{\mathrm{P}} = \mathrm{P}$ and $z(t)$ in (9.3) follows (9.2) rather than (9.4).

The following proposition demonstrates how to compute the moment-generating function of $\overline{z\lambda^2}(T)$.

Proposition 9.1.2. *The function $\Psi_{\overline{z\lambda^2}}(v,u;T)$ defined by (9.3) is given by*

$$\Psi_{\overline{z\lambda^2}}(v,u;T) = \exp\left(A(0,T) + z_0 B(0,T)\right),$$

where $(A(t,T), B(t,T))$ solve the system of Riccati *ODEs*

$$\frac{d}{dt}A(t,T) + \theta z_0 B(t,T) = 0, \tag{9.5}$$

$$\frac{d}{dt}B(t,T) - \left(\theta - \rho\eta bu\lambda(t)\right)B(t,T) + \frac{\eta^2}{2}B(t,T)^2 + v\lambda(t)^2 = 0, \tag{9.6}$$

with the terminal conditions

$$B(T,T) = A(T,T) = 0.$$

Proof. Let us define

$$G(t,z) \triangleq \mathrm{E}^{\widetilde{\mathrm{P}}}\left(e^{v\int_t^T \lambda(s)^2 z(s)\,ds}\,\middle|\, z(t)=z\right).$$

Clearly,

$$\Psi_{z\lambda^2}(v,u;T) = G(0,z_0).$$

On the other hand, by the Feynman-Kac formula, $G(t,z)$ satisfies the following PDE,

$$\frac{\partial}{\partial t}G(t,z) + (\theta z_0 - (\theta - \rho\eta b u\lambda(t))\,z)\frac{\partial}{\partial z}G(t,z) + \frac{\eta^2}{2}z\frac{\partial^2}{\partial z^2}G(t,z) + v\lambda(t)^2 z G(t,z) = 0, \quad (9.7)$$

with the terminal condition

$$G(T,z) = 1, \quad z \geq 0. \quad (9.8)$$

The PDE (9.7) is affine in z, i.e. all coefficients are linear functions of z. To solve it, we make the *ansatz* that the solution $G(t,z)$ is of the exponential form

$$G(t,z) = \exp\left(A(t,T) + zB(t,T)\right).$$

Substituting this conjectured solution into the PDE (9.7) and dividing by G, we get

$$\frac{d}{dt}A(t,T) + z\frac{d}{dt}B(t,T) + (\theta z_0 - (\theta - \rho\eta b u\lambda(t))\,z)\,B(t,T) + \frac{\eta^2}{2}zB(t,T)^2 + v\lambda(t)^2 z = 0.$$

By collecting the coefficients on different powers of z, the two ODEs (9.5)–(9.6) emerge. Boundary conditions follow from (9.8). □

The system of ODEs (9.5)–(9.6) can be solved numerically using the *Runge-Kutta method*, see e.g. Press et al. [1992]. In practice, it is common for the time-dependent volatility $\lambda(t)$ to be piecewise constant,

$$\lambda(t) = \lambda_i, \quad t \in (t_{i-1}, t_i],$$

for some $0 = t_0 < t_1 < \ldots < t_I = T$. In this case, on each of the intervals $(t_{i-1}, t_i]$, the ODEs (9.5)–(9.6) can be solved in closed form, using the formulas from Proposition 8.3.8. By piecing these solutions together[2], we obtain the exact solution to the ODEs over the whole time interval $[0,T]$. However, for a given tolerance on accuracy, the Runge-Kutta method may still be faster than exact solution of the ODEs, as it avoids expensive evaluations of functions exp, ln, etc.

[2]The full procedure is described in Section 10.2.2.2.

Remark 9.1.3. So far, we assumed that η, ρ, and θ were constants. However, it follows easily from the proof of Proposition 9.1.2 that incorporation of time-dependence in η, ρ and θ is merely a matter of changing the ODEs (9.5)–(9.6) to

$$\frac{d}{dt}A(t,T) + \theta(t)z_0 B(t,T) = 0.$$

$$\frac{d}{dt}B(t,T) - (\theta(t) - \rho(t)\eta(t)bu\lambda(t))\,B(t,T) + \frac{\eta(t)^2}{2}B(t,T)^2 + v\lambda(t)^2 = 0.$$

No matter which scheme is ultimately used to solve (9.5)–(9.6), combining the integration method of Theorem 8.4.4 with the integrand in Proposition 9.1.2 — possibly extended as in Remark 9.1.3 — allows for the pricing of call options by the Fourier methods in Section 8.4.

9.2 Asymptotic Expansion with Time-Dependent Volatility

As demonstrated in previous sections, the Fourier method constitutes a powerful tool for establishing a pricing algorithm for European options, provided that the underlying stochastic volatility process is of a sufficiently simple form. Should, say, the volatility function $\psi(z)$ for $z(t)$ be something other than $\sqrt{z}$, or should the skew function $\varphi(x)$ be more complicated than a linear form, analytic tractability (as in Proposition 9.1.2) is often lost and the Fourier method may not be feasible. However, asymptotic expansion methods can still be used in some situations and may, even for cases where Fourier methods do apply, offer a compelling (and very fast) approach to European option pricing.

To develop the asymptotic expansion approach, we return to the general skew functions $\varphi(x)$ and $\psi(z)$ in (8.1)–(8.2), under the simplifying (yet practically relevant) assumption that $\rho = 0$. As in the previous section, we will assume that the volatility $\lambda(t)$ is time-dependent. To summarize, the SDE system under consideration will be

$$dS(t) = \lambda(t)\varphi(S(t))\sqrt{z(t)}\,dW(t), \tag{9.9}$$

$$dz(t) = \theta(z_0 - z(t))\,dt + \eta\psi(z(t))\,dZ(t), \quad z(0) = z_0, \tag{9.10}$$

where $\langle dZ(t), dW(t)\rangle = 0$. The form of the time-dependence — as introduced here exclusively in $\lambda(t)$ — allows us to use time-change arguments similar to those in Section 7.6.1 to show that Lemma 8.5.4 as well as Proposition 8.5.5 still apply.

Lemma 9.2.1. *For the system (9.9)–(9.10) the results of Lemma 8.5.4 and Proposition 8.5.5 hold unchanged, provided we redefine (8.43) to*

$$(\mathcal{F}g)(\omega) = \int_{-\infty}^{\infty} e^{i\omega U} g\left(0, S(0); T^{-1}U\right) dU,$$

and make the substitutions

$$\lambda^2 \overline{z}(T) \to \overline{z\lambda^2}(T), \quad \lambda^2 U \to U, \quad \Psi_{\overline{z}} \to \Psi_{\overline{z\lambda^2}}.$$

For the special case $\psi(z) = \sqrt{z}$, Proposition 9.1.2 derives the expression for $\Psi_{\overline{z\lambda^2}}(u, 0; T)$. For more general choices of $\psi(z)$, we can rely on the PDE from Lemma 8.5.7, appropriately extended to time-dependent $\lambda(t)$. Specifically, $\Psi_{\overline{z\lambda^2}}(u, 0; T) = L(0, z_0; u)$, where $L(t, z; u)$ satisfies the PDE

$$\frac{\partial L}{\partial t} + \theta(z_0 - z)\frac{\partial L}{\partial z} + \frac{\eta^2}{2}\psi(z)^2\frac{\partial^2 L}{\partial z^2} + u\lambda(t)^2 z L = 0, \tag{9.11}$$

subject to the boundary condition $L(T, z; u) = 1$. The equation can be solved numerically, or we can attempt to derive approximations. For the latter, we first introduce a centered transform

$$l(t, z; u) \triangleq L(t, z; u)e^{-u\mu_{\overline{z\lambda^2}}(t,z)}, \tag{9.12}$$

where, under mild regularity conditions on $\psi(z)$,

$$\begin{aligned}
\mu_{\overline{z\lambda^2}}(t, z) &\triangleq \mathrm{E}\left(\int_t^T \lambda(s)^2 z(s)\, ds \,\middle|\, z(t) = z\right) \\
&= \int_t^T \lambda(s)^2 \mathrm{E}\left(z(s) | z(t) = z\right) ds \\
&= \int_t^T \lambda(s)^2 \left(z_0 + (z - z_0)e^{-\theta(s-t)}\right) ds.
\end{aligned}$$

Introduction of $l(t, z; u)$ focuses attention on deviations of $\overline{z\lambda^2}(t)$ away from its mean, which can be expected to be small if η is small — a limit that we shall shortly examine. Insertion of (9.12) into (9.11) reveals that $l(t, z; u)$ satisfies

$$\frac{\partial l}{\partial t} + \theta(z_0 - z)\frac{\partial l}{\partial z} + \frac{\eta^2}{2}\psi(z)^2\left\{\frac{\partial^2 l}{\partial z^2} + lu^2 p(t)^2 + 2up(t)\frac{\partial l}{\partial z}\right\} = 0, \tag{9.13}$$

where

$$p(t) = \int_t^T \lambda(s)^2 e^{-\theta(s-t)}\, ds \tag{9.14}$$

and $l(T, z; u) = 1$.

Lemma 9.2.2. *Let $p(t)$ be as in (9.14), and define $\widetilde{\psi}(z) = \frac{1}{2}\psi(z)^2$ and $h(s, z) = z_0 + (z - z_0)e^{\theta(t-s)}$. An asymptotic expansion for the solution to (9.13) in terms of η^2 is given by*

$$l(t,z;u) = 1 + \eta^2 l_1(t,z;u) + \eta^4 l_2(t,z;u) + O\left(\eta^6\right),$$

where

$$l_1(t,z;u) = u^2 l_{1,2}(t,z),$$
$$l_2(t,z;u) = u^2 l_{2,2}(t,z) - u^3 l_{2,3}(t,z) + \frac{1}{2}u^4 \left(l_{1,2}(t,z)\right)^2,$$

and

$$l_{1,2}(t,z) = \int_t^T p(s)^2 \widetilde{\psi}\left(h(s,z)\right) ds,$$
$$l_{2,2}(t,z) = \int_t^T e^{2\theta s} \widetilde{\psi}\left(h(s,z)\right) \int_s^T e^{-2\theta v} p(v)^2 \widetilde{\psi}''\left(h(v,z)\right) dv\, ds,$$
$$l_{2,3}(t,z) = -2 \int_t^T e^{\theta s} p(s) \widetilde{\psi}\left(h(s,z)\right) \int_s^T e^{-\theta v} p(v)^2 \widetilde{\psi}'\left(h(v,z)\right) dv\, ds.$$

Proof. Let

$$l(t,z;u) = 1 + \sum_{i \geq 1} l_i(t,z;u)\eta^{2i}.$$

Notice that odd powers of η are not used in the expansion, as only η^2 figures in the PDE (9.13). Inserting into (9.13) and collecting terms of order η^2 gives

$$\frac{\partial l_1}{\partial t} + \theta\left(z_0 - z\right)\frac{\partial l_1}{\partial z} + \frac{1}{2}u^2 p(t)^2 \psi(z)^2 = 0,$$

with terminal condition $l_1(T,z) = 0$. This simple PDE can be solved in closed form, yielding the solution listed in the lemma. The result for l_2 is established by collecting terms of order η^4 and proceeding as for l_1. □

While somewhat complicated in appearance, the expressions for the integrals $l_{1,2}$, $l_{2,2}$, and $l_{2,3}$ are trivial to implement on a computer. Indeed, due to the nested nature of the double integrals $l_{2,2}$ and $l_{2,3}$, all integrals can be computed in a single numerical integration loop, at negligible computational cost. In doing the integrals we start from the back, at time T, allowing us at each integration step to update the outer integral, as well as to resolve the inner integrals. In some cases of practical interest it is also possible to evaluate the integrals analytically.

Apart from potential direct application in the Fourier technique in Proposition 8.5.5, the result of Lemma 9.2.2 allows us to compute central moments as follows:

$$\mathrm{E}\left(\left(\overline{z\lambda^2}(T) - \mu_{\overline{z\lambda^2}}(0,z_0)\right)^n\right) = \left.\frac{\partial^n l(0,z_0;u)}{\partial u^n}\right|_{u=0}, \quad n = 1,2,\ldots. \tag{9.15}$$

There are many ways to turn these moments into an option price expression. For instance, we could rely on a classical Gram-Charlier expansion (see

Ochi [1990]) or perhaps some parametric density family to express the full density of $\overline{z\lambda^2}(T)$, to be used directly in (time-dependent generalizations of) equations (8.37) or (8.42). Alternatively, we can use Taylor expansions for a closed-form asymptotic result. Specifically, if the function g is defined as in Lemma 8.5.4, we can write

$$\mathrm{E}\left(f\left(S(T)\right)\right) = g\left(0, S(0); \overline{v}\right) + \sum_{n=1}^{\infty} \frac{1}{n! T^n} \left.\frac{\partial^n g}{\partial v^n}\right|_{v=\overline{v}} \mathrm{E}\left(\left(\overline{z\lambda^2}(T) - \mu_{\overline{z\lambda^2}}(0, z_0)\right)^n\right),$$

where the derivatives are to be evaluated at $\overline{v} \triangleq \mu_{\overline{z\lambda^2}}(0, z_0)/T$.

From (9.15) and the expansion formula in Lemma 9.2.2, a few manipulations give the required result.

Lemma 9.2.3. *With $g(t, S; v)$ defined as in Lemma 8.5.4, we have to order $O(\eta^4)$*

$$\mathrm{E}\left(f\left(S(T)\right)\right) = g\left(0, S(0); \overline{v}\right) + T^{-2}\left(\eta^2 l_{1.2} + \eta^4 l_{2.2}\right)\frac{\partial^2 g}{\partial v^2} - \eta^4 T^{-3} l_{2,3}\frac{\partial^3 g}{\partial v^3} + \frac{1}{2}\eta^4 T^{-4} l_{1,2}^2 \frac{\partial^4 g}{\partial v^4},$$

where all derivatives are evaluated at $\overline{v} = \mu_{\overline{z\lambda^2}}(0, z_0)/T$.

To show an application of this lemma, consider the important special case of a call option $f(x) = (x - K)^+$.

Proposition 9.2.4. *Define the log-moneyness $k = \ln(K/S(0))$ and set $\tau = \int_0^T \lambda(s)^2 ds$. Also set*

$$q_1 = \mu_{\overline{z\lambda^2}}(0, z_0)/T + \alpha_0\eta^2 + \alpha_1\eta^2 k^2 + O\left(\eta^4\right), \tag{9.16}$$

$$q_2 = \mu_{\overline{z\lambda^2}}(0, z_0)/T + \left(\alpha_0\eta^2 + \beta_0\eta^4\right) + \left(\alpha_1\eta^2 + \beta_1\eta^4\right)k^2 + \beta_2\eta^4 k^4 e^{-\Lambda\eta^2 k^2} + O\left(\eta^6\right), \tag{9.17}$$

where Λ is an arbitrary positive number and the coefficients $\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2$ are given in Appendix 9.B. Then the value of a European call option in the model (9.9)–(9.10) is given by

$$c(0, S; T, K) \approx S(0)\Phi(d_+) - K\Phi(d_-), \tag{9.18}$$

$$d_\pm = \frac{-k \pm \sigma_{\mathrm{imp}}^2 T/2}{\sigma_{\mathrm{imp}}\sqrt{T}},$$

where, to order η^2,

$$\sigma_{\mathrm{imp}} = \Omega_0\sqrt{q_1} + \Omega_1 q_1^{3/2} T + O\left(T^2\right),$$

or, to order η^4,

$$\sigma_{\text{imp}} = \Omega_0\sqrt{q_2} + \Omega_1 q_2^{3/2} T + O\left(T^2\right).$$

Also, we have

$$\Omega_0 = \frac{-k}{\int_K^{S(0)} \varphi(u)^{-1} du},$$

$$\Omega_1 = -\frac{\Omega_0}{\left(\int_K^{S(0)} \varphi(u)^{-1} du\right)^2} \ln\left(\Omega_0 \left(\frac{KS(0)}{\varphi(K)\varphi\left(S(0)\right)}\right)^{1/2}\right).$$

Proof. (Sketch). For the case of a call option, the function g can be approximated using the small-time expansion result in Proposition 7.5.1; we here choose to expand around a log-normal model, so $\beta = 0$ in the proposition. Using the resulting expression to evaluate the terms in Lemma 9.2.3 yields, after some work, a direct expansion for the call option price. It is often more accurate to convert the price expansion into an expansion in implied "skew variance" v^* , where v^* satisfies

$$\mathrm{E}\left((S(T) - K)^+\right) = g\left(0, S(0); v^*\right). \tag{9.19}$$

We write

$$v^* = \overline{v} + \eta^2 v_1^* + \eta^4 v_2^* + \ldots, \tag{9.20}$$

insert this expression into (9.19) and Taylor-expand around $\overline{v}$. Matching the resulting expression against the direct expansion for the call option price yields closed-form expressions for v_1^* and v_2^*. These results are such that

$$\overline{v} + \eta^2 v_1^* = q_1, \quad \overline{v} + \eta^2 v_1^* + \eta^4 v_2^* = q_2,$$

where q_1 and q_2 are defined in (9.16) and (9.17), respectively. Another application of Proposition 7.5.1 turns the skew variance into an implied Black volatility,

$$\sigma_{\text{imp}}\sqrt{T} = \Omega_0\sqrt{v^* T} + \Omega_1 \left(v^* T\right)^{3/2} + \ldots.$$

The proposition follows. □

Remark 9.2.5. Full details for the proof of Proposition 9.2.4 and tests of the precision of the expansion can be found in Andersen and Brotherton-Ratcliffe [2005].

9.3 Averaging Methods

The Fourier integration method from Section 9.1 involves numerical integration of a function that itself is calculated numerically by solving a coupled system of ODEs. If both the integral and the ODEs are discretized with N steps, the complexity of the scheme $O(N^2)$, which could be costly. On the other hand, the asymptotic expansion method from Section 9.2 is fast but may not be accurate enough for certain values of model parameters, especially high η. In this section we develop the parameter averaging approach to time-dependent model parameters that is both fast and accurate. We have seen applications of the method to local volatility models already, in Section 7.6.2.

9.3.1 Volatility Averaging

We initially work with the model (9.1)–(9.2) with zero correlation, $\rho = 0$. Our goal is to replace the time-dependent $\lambda(t)$ with a constant $\overline{\lambda}$ in such a way that pricing of vanilla options at a given maturity T is preserved to good approximation. For this, we first notice that a European option price can be represented as an integral of a known function against the distribution of the term stochastic variance, a representation we have already fruitfully used in Sections 8.5 and 9.2. In particular, for an at-the-money option, where $K = S(0)$,

$$\mathrm{E}\left((S(T) - S(0))^+\right) = \mathrm{E}\left(\mathrm{E}\left((S(T) - S(0))^+ \middle| \{z(t),\, t \in [0,T]\}\right)\right). \tag{9.21}$$

Because the Brownian motion that drives $z(t)$ is independent of the Brownian motion that drives $S(t)$, the distribution of $S(T)$ in the model (9.2) is displaced log-normal when conditioned on a particular path of $z(t)$. Hence, the inner conditional expectation in (9.21) can be evaluated easily to yield

$$\mathrm{E}\left((S(T) - S(0))^+\right) = \mathrm{E}\left(h\left(\overline{z\lambda^2}(T)\right)\right), \tag{9.22}$$

where $\overline{z\lambda^2}(T)$ is defined by (9.3) and the function $h(x)$ is the displaced log-normal at-the-money option value as function of variance:

$$h(x) = \frac{bS(0) + (1-b)L}{b}\left(2\Phi\left(b\sqrt{x}/2\right) - 1\right). \tag{9.23}$$

Given the practical importance of correctly pricing at-the-money options, the problem of finding the effective, time-independent model volatility can be cast into the problem of finding such $\overline{\lambda}$ that

$$\mathrm{E}\left(h\left(\int_0^T \lambda(t)^2 z(t)\,dt\right)\right) = \mathrm{E}\left(h\left(\overline{\lambda}^2 \int_0^T z(t)\,dt\right)\right) \tag{9.24}$$

or, in our notations,

$$\mathrm{E}\left(h\left(\overline{z\lambda^2}(T)\right)\right) = \mathrm{E}\left(h\left(\overline{\lambda}^2\overline{z}(T)\right)\right).$$

Neither of the expected values in (9.24) is available in closed form. However, the moment-generating functions of both $\overline{z\lambda^2}(T)$ and $\overline{z}(T)$ are available in closed form and as a solution to a system of ODEs, respectively (see Propositions 8.3.8 and 9.1.2). This observation suggests approximating $h(x)$ with a function of exponential form

$$h(x) \approx a + be^{cx}. \tag{9.25}$$

We choose the coefficients a, b, c to get the best local second-order fit at the mean of $\overline{z\lambda^2}(T)$,

$$h\left(\zeta_T\right) = a + be^{c\zeta_T}, \quad h'\left(\zeta_T\right) = bce^{c\zeta_T}, \quad h''\left(\zeta_T\right) = bc^2e^{c\zeta_T}, \tag{9.26}$$

where

$$\zeta_T = \mathrm{E}\left(\overline{z\lambda^2}(T)\right) = \mu_{\overline{z\lambda^2}}(0, z_0) = z_0\int_0^T \lambda(t)^2\,dt.$$

Clearly

$$c = \frac{h''\left(\zeta_T\right)}{h'\left(\zeta_T\right)}, \tag{9.27}$$

and the problem (9.24) can be approximated with

$$a + b\mathrm{E}\left(e^{c\overline{z\lambda^2}(T)}\right) = a + b\mathrm{E}\left(e^{c\overline{\lambda}^2\overline{z}(T)}\right) \quad \Rightarrow \quad \mathrm{E}\left(e^{c\overline{z\lambda^2}(T)}\right) = \mathrm{E}\left(e^{c\overline{\lambda}^2\overline{z}(T)}\right), \tag{9.28}$$

which gives us an *effective volatility* approximation result that we formulate as a theorem.

Theorem 9.3.1. *Values of European options with expiry T in the model (9.1)–(9.2) are well approximated by their values in the model (8.3)–(8.4) with λ set to the effective SV volatility $\overline{\lambda}$, which solves the equation*

$$\Psi_{\overline{z}}\left(\frac{h''\left(\zeta_T\right)}{h'\left(\zeta_T\right)}\overline{\lambda}^2, 0; T\right) = \Psi_{\overline{z\lambda^2}}\left(\frac{h''\left(\zeta_T\right)}{h'\left(\zeta_T\right)}, 0; T\right), \tag{9.29}$$

where

$$\zeta_T = z_0\int_0^T \lambda(t)^2\,dt,$$

the function $h(x)$ is given by (9.23), and the moment-generating functions $\Psi_{\overline{z}}$ and $\Psi_{\overline{z\lambda^2}}$ are given by Propositions 8.3.8 and 9.1.2, respectively.

Proof. Follows after replacing the problem (9.24) with (9.28), using the expression (9.27) for c. □

Remark 9.3.2. The expression on the left-hand side of (9.29) can be computed in closed form; the right-hand side is straightforward to calculate from Proposition 9.1.2 and the accompanying remarks. Equation (9.29) can be solved for $\overline{\lambda}^2$ in just a couple of Newton-Raphson iterations, starting from an initial guess of $T^{-1}\int_0^T \lambda(t)^2\,dt$.

Remark 9.3.3. The effective volatility $\overline{\lambda}$ as given by Theorem 9.3.1 is second-order accurate in the sense that the approximation (9.25) is second-order accurate with the choice of parameters in (9.26). We note that the method does not readily lend itself to higher-order approximations but this is of little relevance as the quality of the approximation is excellent as is.

9.3.2 Skew Averaging

The slope of the volatility smile in the SV model (8.3)–(8.4) is controlled by the skew parameter b. In this section we make the skew parameter time-dependent, and consider a model driven by the SDEs

$$dS(t) = \lambda(t)\left(b(t)S(t) + (1-b(t))\,L\right)\sqrt{z(t)}\,dW(t), \tag{9.30}$$

$$dz(t) = \theta\left(z_0 - z(t)\right)\,dt + \eta\sqrt{z(t)}\,dZ(t), \tag{9.31}$$

with $\langle dZ(t), dW(t)\rangle = 0$. In Section 7.6.2 we derived the formula for the effective, or average, skew for local volatility models, see Proposition 7.6.2 and Corollary 7.6.3. The extension of these results to stochastic volatility models is straightforward, leading to a similar expression with somewhat more complicated averaging weights, as the following proposition demonstrates.

Proposition 9.3.4. *The effective skew $\overline{b}$ for the equation*

$$dS(t) = \lambda(t)\left(b(t)S(t) + (1-b(t))\,S(0)\right)\sqrt{z(t)}\,dW(t)$$

over a time horizon $[0,T]$ is given by

$$\overline{b} = \int_0^T b(t)w_T(t)\,dt, \tag{9.32}$$

where the weights $w_T(t)$ are given by

$$w_T(t) = \frac{v(t)^2\lambda(t)^2}{\int_0^T v(t)^2\lambda(t)^2\,dt}, \tag{9.33}$$

$$v(t)^2 = z_0^2\int_0^t \lambda(s)^2\,ds + z_0\eta^2 e^{-\theta t}\int_0^t \lambda(s)^2\frac{e^{\theta s} - e^{-\theta s}}{2\theta}\,ds.$$

The result in Proposition 9.3.4 can be derived by the same technique that lead to Proposition 7.6.2 and Corollary 7.6.3. Alternatively, it can be found by the small-noise expansion method in Section 7.6.3. We leave the details of these derivations to the reader and, for instructional value, instead list a third proof based on Markovian semi-groups in Appendix 9.A, see also Piterbarg [2005b]. The fact that the same result is obtained as a solution to a number of differently posed problems of skew averaging suggests robustness and general applicability.

It will be useful for the next section to derive an extension of Proposition 9.3.4 to cover the process $z(t)$ with time-dependent volatility of variance. Specifically, let us use the following dynamics for the stochastic variance process

$$dz(t) = \theta\left(z_0 - z(t)\right)\,dt + \eta(t)\sqrt{z(t)}\,dZ(t). \tag{9.34}$$

Corollary 9.3.5. *The effective skew $\overline{b}$ for the equation*

$$dS(t) = \lambda(t)\left(b(t)S(t) + \left(1 - b(t)\right)S(0)\right)\sqrt{z(t)}\,dW(t)$$

with $z(t)$ following (9.34) over a time horizon $[0,T]$ is given by

$$\overline{b} = \int_0^T b(t)w_T(t)\,dt, \tag{9.35}$$

where the weights $w_T(t)$ are given by

$$w_T(t) = \frac{\widehat{v}(t)^2\lambda(t)^2}{\int_0^T \widehat{v}(t)^2\lambda(t)^2\,dt}, \tag{9.36}$$

$$\widehat{v}(t)^2 = z_0^2\int_0^t \lambda(s)^2\,ds + z_0 e^{-\theta t}\int_0^t \lambda(s)^2 e^{-\theta s}\int_0^s \eta(u)^2 e^{2\theta u}du\,ds.$$

Proof. The proof or the corollary proceeds as the proof (in Appendix 9.A) of Proposition 9.3.4, but using

$$\mathrm{E}\left(z(t)^2\right) = z_0^2 + z_0\int_0^t \eta(u)^2 e^{-2\theta(t-u)}du \tag{9.37}$$

instead of (9.100) in (9.101) for $z(t)$ given by (9.34). □

9.3.3 Volatility of Variance Averaging

Finally, we turn our attention to the problem of averaging the volatility of variance η in (9.1). More precisely, suppose we have a stochastic variance process with time-dependent volatility of variance (9.34). We would like to find a constant parameter $\overline{\eta}$ such that the model (9.30), (9.34) is approximated by the model (9.30), (9.31) with $\eta = \overline{\eta}$.

Before discussing our proposed solution method, we note that usage of time-dependent volatility of variance $\eta(t)$ for model calibration purposes may not be quite as necessary as for other parameters. Fundamentally, a time-dependent η will allow us to control the term structure of volatility smile convexity in the maturity direction. On the other hand, we already have control over the curvatures of volatility smiles at different times T via θ, the mean reversion of variance parameter: higher values of θ make implied volatility smiles flatten faster as option expiries increase, while lower values make them flatten slower, see Sections 8.2 and 8.7. Even though the level of control granted through θ is rather crude, it is often sufficient in practice, all the more so since the volatility smile curvatures are typically not observable to a high degree of precision.

The curvature of the volatility smile is related to the kurtosis of the distribution of $S(T)$ which, in stochastic volatility models, is controlled by the variance of the quantity

$$\overline{z\lambda^2}(T) = \int_0^T \lambda(t)^2 z(t)\, dt,$$

i.e. the integrated stochastic variance to expiry time T. Since the curvature of the smile is the main effect of the volatility of variance parameter η, a representative constant volatility of variance $\overline{\eta}$ should intuitively be chosen as the solution to

$$\mathrm{E}\left(\left(\int_0^T \lambda(t)^2 \widehat{z}(t)\, dt\right)^2\right) = \mathrm{E}\left(\left(\int_0^T \lambda(t)^2 z(t)\, dt\right)^2\right), \tag{9.38}$$

where $z(t)$ follows (9.34) and $\widehat{z}(t)$ follows (9.31).

Theorem 9.3.6. *For (9.34), the effective volatility of variance to maturity T, derived from the condition (9.38), is given by*

$$\overline{\eta}^2 = \frac{\int_0^T \eta(t)^2 \rho_T(t)\, dt}{\int_0^T \rho_T(t)\, dt},$$

where the weight function $\rho_T(t)$ is given by

$$\rho_T(r) = \int_r^T ds \int_s^T dt\, \lambda(t)^2 \lambda(s)^2 e^{-\theta(t-s)} e^{-2\theta(s-r)}.$$

Proof. While the proof is straightforward, we here provide full details in order to demonstrate some generally useful manipulations for the computations of moments in stochastic volatility models. First, we have

$$\begin{aligned}
&\mathrm{E}\left(\left(\int_0^T \lambda(t)^2 z(t)\,dt\right)^2\right)\\
&= 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 \mathrm{E}(z(t)z(s))\\
&= 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\mathrm{E}(z(s)^2)\\
&\quad + 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2(1-e^{-\theta(t-s)})z_0\mathrm{E}(z(s)).
\end{aligned}$$

Using (9.37) for $\mathrm{E}(z(s)^2)$ we get

$$\begin{aligned}
\mathrm{E}\left(\left(\int_0^T \lambda(t)^2 z(t)\,dt\right)^2\right) &= 2z_0^2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\\
&\quad + 2z_0\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\int_0^s \eta(r)^2 e^{-2\theta(s-r)}dr\\
&\quad + 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2\left(1-e^{-\theta(t-s)}\right)z_0\mathrm{E}\left(z(s)\right).
\end{aligned}$$

Changing the order of integration for the second term, we obtain

$$\begin{aligned}
\mathrm{E}\left(\left(\int_0^T \lambda(t)^2 z(t)\,dt\right)^2\right) &= 2z_0^2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\\
&\quad + 2z_0\int_0^T dr\,\eta(r)^2 \int_r^T ds \int_s^T dt\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}e^{-2\theta(s-r)}\\
&\quad + 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2\left(1-e^{-\theta(t-s)}\right)z_0\mathrm{E}\left(z(s)\right).
\end{aligned}$$

If we define

$$\rho_T(r) = \int_r^T ds \int_s^T dt\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}e^{-2\theta(s-r)},$$

the equation (9.38) can be rewritten in the form

$$\int_0^T \overline{\eta}^2\rho_T(t)\,dt = \int_0^T \eta(t)^2\rho_T(t)\,dt.$$

The theorem is proved. □

Remark 9.3.7. While we used zero correlation between the underlying and its stochastic variance both in motivating our results and in deriving them,

the same approach can be applied in the non-zero correlation case. Some results, in particular Proposition 9.3.4 and Theorem 9.3.6, remain unchanged. On the other hand, the effective volatility formula in Theorem 9.3.1 is based on the representation (9.22) which, clearly, does not hold with non-zero correlation; despite that, the formula can still be used with good accuracy.

9.3.4 Calibration by Parameter Averaging

The main application of the averaging formulas developed above is in creating efficient model calibration algorithms. In this section, we discuss in some detail how such an algorithm could proceed; the principles that we outline here shall be used repeatedly later in this book. Now, suppose a collection of expiries

$$0 = T_0 < T_1 < T_2 < \ldots < T_N$$

is given, as well as a collection of strikes $K_1, \ldots, K_M$. Let the market values of European call options with expiries T_n and strikes K_m be denoted by

$$\{\widehat{c}_{n,m}, \quad n = 1, \ldots, N, \quad m = 1, \ldots, M\}.$$

Our objective is to find time-dependent model parameters $\lambda(t)$, $b(t)$, and $\eta(t)$ such that the model

$$dS(t) = \lambda(t)\left(b(t)S(t) + (1 - b(t))\, L\right)\sqrt{z(t)}\, dW(t), \tag{9.39}$$

$$dz(t) = \theta\left(z_0 - z(t)\right) dt + \eta(t)\sqrt{z(t)}\, dZ(t), \tag{9.40}$$

values European options with expiries T_n, $n = 1, \ldots, N$, and strikes K_m, $m = 1, \ldots, M$, as closely as possible to their market values[3] $\{\widehat{c}_{n,m}\}$.

Let us denote the prices of options in the model (9.39)–(9.40) by

$$c_{n,m} = c_{n,m}\left(\mathcal{X}\right),$$

where by $\mathcal{X}$ we denote the state of the model,

$$\mathcal{X} = \{\lambda(\cdot), b(\cdot), \eta(\cdot)\}.$$

Typically, calibration would be performed by solving the following non-linear optimization problem

$$\{\lambda(\cdot), b(\cdot), \eta(\cdot)\} = \operatorname{argmin} \sum_{n,m} \left(c_{n,m}\left(\mathcal{X}\right) - \widehat{c}_{n,m}\right)^2, \tag{9.41}$$

[3] In interest rate markets, the underlyings for options of different expiries are often different, in the sense that they represent swap rates of different tenors and fixing dates. We will deal with such complications in due time.

where[4] $c_{n,m}(\mathcal{X})$'s are obtained in some sort of numerical procedure. With the averaging formulas, an appealing alternative is available. To describe it, let us denote triples of SV "market" parameter values by $\{\widehat{\lambda}_n, \widehat{b}_n, \widehat{\eta}_n\}$, $n = 1, \ldots, N$, determined such that the market prices of European options expiring at time T_n, i.e. $\{\widehat{c}_{n,m},\ m = 1, \ldots, M\}$, match prices obtained in the model

$$dS(t) = \widehat{\lambda}_n \left(\widehat{b}_n S(t) + \left(1 - \widehat{b}_n\right) L\right) \sqrt{z(t)}\, dW(t), \tag{9.42}$$

$$dz(t) = \theta\,(z_0 - z(t))\ dt + \widehat{\eta}_n \sqrt{z(t)}\, dZ(t). \tag{9.43}$$

Sets of market parameters are routinely maintained and updated by trading desks, and instead of considering $\{\widehat{c}_{n,m}\}$ to be fundamental market inputs, we can think of $\{\widehat{\lambda}_n, \widehat{b}_n, \widehat{\eta}_n\}$, $n = 1, \ldots, N$, as such. We often refer to them as "term" parameters to highlight the fact that they are constant for the whole "term", or life, of the relevant options.

Critically, the averaging formulas link time-dependent parameters $\{\lambda(t),\ b(t), \eta(t)\}$ to constant parameters $\{\widehat{\lambda}_n, \widehat{b}_n, \widehat{\eta}_n\}$, $n = 1, \ldots, N$, directly without referencing option values. To take advantage of this, let us denote by

$$\left\{\overline{\lambda}_n(\mathcal{X}), \overline{b}_n(\mathcal{X}), \overline{\eta}_n(\mathcal{X})\right\}$$

the averaged parameters (to time T_n) for the model (9.39)–(9.40). Then the optimization problem (9.41) can be replaced by a more convenient one,

$$\begin{aligned}\{\lambda(\cdot),\ b(\cdot), \eta(\cdot)\} = \operatorname{argmin} \Bigg(& W_\lambda \sum_n \left(\overline{\lambda}_n(\mathcal{X}) - \widehat{\lambda}_n\right)^2 \\ & + W_b \sum_n \left(\overline{b}_n(\mathcal{X}) - \widehat{b}_n\right)^2 + W_\eta \sum_n \left(\overline{\eta}_n(\mathcal{X}) - \widehat{\eta}_n\right)^2 \Bigg), \end{aligned} \tag{9.44}$$

where W_λ, W_b, and W_η are weights linked to relative importance of matching particular parameters. Compared to (9.41), this norm formulation is both more intuitive to traders — who often tend to think about the state of the market in terms of model parameters, rather than in terms of absolute option prices — and computationally advantageous, insofar as the norm requires no outright computation of option values.

In practice, the calibration (9.44) needs not be performed by brute-force optimization. By carefully choosing the order of calculations, calibration can be split into independent sub-calibrations: one for volatility of variance (η); one for skewness (b); and one for volatility (λ). Skew and volatility of variance calibrations can be performed by matrix manipulations, and the volatility calibration can be split into a sequence of numerically solved one-dimensional equations. To describe this calibration idea in more detail,

[4] Often different terms are weighted differently.

let us first collect all relevant averaging results for easy reference. For the volatility of variance, we have from Theorem 9.3.6,

$$\overline{\eta}_n(\mathcal{X})^2 = \frac{\int_0^{T_n} \eta(t)^2 \rho_{T_n}(t;\lambda(\cdot))\,dt}{\int_0^{T_n} \rho_{T_n}(t;\lambda(\cdot))\,dt}, \quad n = 1,\ldots,N, \tag{9.45}$$

where we have now explicitly indicated the dependence of weights $\rho_T(t;\lambda(\cdot))$ on the volatility function $\lambda(t)$. For the skews, we have from Corollary 9.3.5,

$$\overline{b}_n(\mathcal{X}) = \int_0^{T_n} b(t) w_{T_n}(t;\lambda(\cdot),\eta(\cdot))\,dt, \quad n = 1,\ldots,N, \tag{9.46}$$

where again the dependence of weights $w_T(t;\lambda(\cdot),\eta(\cdot))$ on model parameters is highlighted. Finally, the equations for volatilities from Theorem 9.3.1 are

$$\overline{\lambda}_n(\mathcal{X}) = F\left(\lambda(\cdot);\overline{b}_n(\mathcal{X}),\overline{\eta}_n(\mathcal{X})\right), \quad n = 1,\ldots,N, \tag{9.47}$$

where, in the notation of Theorem 9.3.1,

$$F(\lambda(\cdot);\overline{b}_n(\mathcal{X}),\overline{\eta}_n(\mathcal{X})) = \sqrt{\frac{h'(\zeta_{T_n})}{h''(\zeta_{T_n})} \times \Psi_{\overline{z}}^{-1}\left(\Psi_{\overline{z\lambda^2}}\left(\frac{h''(\zeta_{T_n})}{h'(\zeta_{T_n})},0;T\right),0;T\right)},$$

$$\zeta_{T_n} = z_0 \int_0^{T_n} \lambda(t)^2\,dt.$$

Note that the function F depends on $\overline{b}_n$ through h, and on $\overline{\eta}_n$ through $\Psi_{\overline{z\lambda^2}}$ and $\Psi_{\overline{z}}$.

Equations (9.45)–(9.47) can be discretized if the model parameters are constant between option expiry dates $\{T_n\}_{n=1}^N$, a common assumption in practice. In this case, we can define λ_i, b_i and η_i by

$$\lambda(t) = \sum_{i=1}^N \lambda_i 1_{\{t\in(T_{i-1},T_i]\}},$$

$$b(t) = \sum_{i=1}^N b_i 1_{\{t\in(T_{i-1},T_i]\}},$$

$$\eta(t) = \sum_{i=1}^N \eta_i 1_{\{t\in(T_{i-1},T_i]\}}.$$

In addition, we discretize the weights and define $\rho_{n,i}(\lambda(\cdot))$ and $w_{n,i}(\lambda(\cdot),\eta(\cdot))$ by

$$\rho_{T_n}(t;\lambda(\cdot)) = \sum_{i=1}^n \rho_{n,i}(\lambda(\cdot))\,1_{\{t\in(T_{i-1},T_i]\}},$$

$$w_{T_n}(t;\lambda(\cdot),\eta(\cdot)) = \sum_{i=1}^n w_{n,i}(\lambda(\cdot),\eta(\cdot))\,1_{\{t\in(T_{i-1},T_i]\}}.$$

Denote

$$\overline{\rho}_{n,i}\left(\lambda(\cdot)\right) = \frac{\rho_{n,i}\left(\lambda(\cdot)\right)}{\int_0^{T_n} \rho_{T_n}\left(t;\lambda(\cdot)\right)\,dt}.$$

Our goal is to solve three systems of equations:

$$\sum_{i=1}^{n} \overline{\rho}_{n,i}\left(\lambda(\cdot)\right)\left(T_i - T_{i-1}\right)\eta_i^2 = \left(\widehat{\eta}_n\right)^2, \tag{9.48}$$

$$\sum_{i=1}^{n} w_{n,i}\left(\lambda(\cdot),\eta(\cdot)\right)\left(T_i - T_{i-1}\right)b_i = \widehat{b}_n, \tag{9.49}$$

$$F\left(\lambda(\cdot);\overline{b}_n\left(\mathcal{X}\right),\overline{\eta}_n\left(\mathcal{X}\right)\right) = \widehat{\lambda}_n, \tag{9.50}$$

for $n = 1,\ldots,N$. At first glance this does not seem entirely straightforward. For example, the system (9.48) appears to be a linear system of equations in $\eta_1^2,\ldots,\eta_N^2$, but the coefficients $\overline{\rho}_{n,i}(\lambda(\cdot))$ depend on $\lambda(t)$, another unknown model parameter. However, by iteratively solving these equations in the right order, we can design a very efficient algorithm, which we now proceed to describe in detail.

First, we note that the equations on volatilities (9.50) do not depend on any other model parameters. They do depend on term parameters $\overline{b}_n(\mathcal{X})$, $\overline{\eta}_n(\mathcal{X})$, which we just replace with their market values, thus solving

$$F\left(\lambda(\cdot);\widehat{b}_n,\widehat{\eta}_n\right) = \widehat{\lambda}_n, \quad n = 1,\ldots,N.$$

The n-th equation in this series only involves λ_i's for $i = 1,\ldots,n$, so the n-th equation can be rewritten as

$$F\left(\lambda_1,\ldots,\lambda_n;\widehat{b}_n,\widehat{\eta}_n\right) = \widehat{\lambda}_n.$$

The case $n = 1$ has the trivial solution

$$\lambda_1^* = \widehat{\lambda}_1.$$

Proceeding iteratively in n, the n-th equation is reduced to

$$F\left(\lambda_1^*,\ldots,\lambda_{n-1}^*,\lambda_n;\widehat{b}_n,\widehat{\eta}_n\right) = \widehat{\lambda}_n, \tag{9.51}$$

where the λ_i^*, $i = 1,\ldots,n-1$, are the model parameters already solved for. Thus, the first step of calibration consists of solving the system of equations (9.50) as N decoupled one-dimensional equations (9.51).

On the second step, we solve the linear system (9.48) for η_i^2, $i = 1,\ldots,N$. The coefficients of the system depend on λ_i's which have already been computed, and we solve

$$\sum_{i=1}^{n} \overline{\rho}_{n,i}\left(\lambda^*(\cdot)\right)\left(T_i - T_{i-1}\right)\eta_i^2 = \left(\widehat{\eta}_n\right)^2, \quad n = 1,\ldots,N.$$

The solution η_i^*, $i = 1, \ldots, N$, to this system can either be found by matrix methods, or by simple sequential substitution since the n-th equation involves η_i^2 for $i = 1, \ldots, n$ only.

Finally, on the third step, we solve the linear system

$$\sum_{i=1}^{n} w_{n,i}\left(\lambda^*(\cdot), \eta^*(\cdot)\right)(T_i - T_{i-1})b_i = \widehat{b}_n, \quad n = 1, \ldots, N, \tag{9.52}$$

for b_i, $i = 1, \ldots, N$. This system is obtained from (9.49) by substituting $\lambda(\cdot)$, $\eta(\cdot)$ with their solved-for values $\lambda^*(\cdot)$, $\eta^*(\cdot)$. Again, the system can be solved sequentially.

To prevent overfitting, it is often useful to regularize the optimization problem through introduction of smoothing terms in the objective function. This can help to, for example, dampen the noise that could be present in market-observed parameters. Taking (9.52) as an example and fixing a smoothing weight $W > 0$, we can replace (9.52) with the minimization problem

$$\sum_{n=1}^{N}\left(\sum_{i=1}^{n} w_{n,i}\left(\lambda^*(\cdot), \eta^*(\cdot)\right)(T_i - T_{i-1})b_i - \widehat{b}_n\right)^2 + W\sum_{i=2}^{N}(b_i - b_{i-1})^2 \to \min. \tag{9.53}$$

This is a simple quadratic minimization problem with no constraints and is easily solved by linear algebra methods, see Golub and van Loan [1989]. The same regularization idea could be applied to the problem of finding $\lambda(t)$ and $b(t)$.

If the regularization weight W in (9.53) is too high then the averaged skew calculated by the model can be significantly different from the market skew, $\bar{b}_n(\mathcal{X}^*) \neq \widehat{b}_n$, $n = 1, \ldots, N$. By itself this may not be such a bad thing as one may prefer a smoother model skew over the exact fit to market skews. However, this poses problems to the *volatility* calibration, as the equation for model volatility (9.51) used the "wrong" skew (and volatility of variance as well, were we to apply regularization to that). The exact fit to market volatilities is often much more important than the exact fit to skews or volatilities of variance. Fortunately, this problem is easy to rectify by solving the system (9.51) again, this time using the true model averaged skews $\bar{b}_n(\mathcal{X}^*)$ (and volatilities of variance) on the left-hand side of (9.51) which are available at this stage of the algorithm.

9.4 PDE Method

In the previous three sections, we discussed the development of methods for efficient model calibration and for the pricing of simple European options. In

the remainder of this chapter, we turn our attention to numerical techniques that allow a calibrated model to be used for pricing of general fixed income derivatives. We start out with the application of the PDE methods from Chapter 2.

9.4.1 PDE Formulation

The flexibility of the PDE method makes it applicable to a generalization of the specification (8.1)–(8.2) with a fully general time-dependent volatility function $\varphi(t, S)$. Let us therefore consider the following vector SDE

$$dS(t) = \varphi\left(t, S(t)\right)\sqrt{z(t)}\, dW(t), \tag{9.54}$$

$$dz(t) = \theta\left(z_0 - z(t)\right)\, dt + \eta(t)\psi\left(z(t)\right)\, dZ(t), \tag{9.55}$$

where $\langle dZ(t), dW(t)\rangle = \rho\, dt$ and $z(0) = z_0$. Let $V(T)$ be an $\mathcal{F}_T$-measurable payoff and let $V(t, z, S)$ denote the numeraire-deflated value at time t, given $S(t) = S$ and $z(t) = z$, of a derivative that pays $V(T)$ at time T, $t \leq T$. By the usual arguments, $V(t, z, S)$ satisfies the following partial differential equation

$$\begin{aligned} 0 = {} & \frac{\partial}{\partial t}V\left(t, z, S\right) + \theta\left(z_0 - z\right)\frac{\partial}{\partial z}V\left(t, z, S\right) + \frac{\eta(t)^2}{2}\psi(z)^2\frac{\partial^2}{\partial z^2}V\left(t, z, S\right) \\ & + \frac{z}{2}\varphi\left(t, S\right)^2\frac{\partial^2}{\partial S^2}V\left(t, z, S\right) + \rho\eta(t)\psi(z)\sqrt{z}\varphi\left(t, S\right)\frac{\partial^2}{\partial z \partial S}V\left(t, z, S\right). \end{aligned} \tag{9.56}$$

This PDE holds for $t \in [0, T]$ and $(S, z) \in \mathbb{R} \times \mathbb{R}^+$.

Fundamentally, (9.56) can be solved numerically by an application of the two-dimensional ADI scheme with a predictor-corrector step, as developed in Section 2.11.2. In an actual implementation of the ADI method, however, several issues in grid design and choice of boundary conditions must be addressed, a task to which we now turn.

9.4.2 Range for Stochastic Variance

Fixing a small probability $q_z > 0$, the range $[z_{\min}, z_{\max}]$ for z in the ADI grid can be set to cover the fraction $(1 - q_z)$ of the range of $z(T)$ in probability, i.e. from the conditions

$$\mathrm{P}\left(z(T) < z_{\min}\right) = \mathrm{P}\left(z(T) > z_{\max}\right) = q_z/2.$$

These probabilities are not known in closed form for $z(T)$ satisfying (9.55), so we will often have to resort to approximations. For instance, if ψ is not too different from a square root, we can replace

$$\psi(z) \to \frac{\psi\left(z_0\right)}{\sqrt{z_0}}\sqrt{z}, \tag{9.57}$$

to obtain a process of the square root type with time-dependent $\eta(t)$. From this representation, we can find an effective $\overline{\eta}$ to time horizon T by Theorem 9.3.6 and then apply the exact distribution of $z(T)$ with time-constant parameters from Proposition 8.3.2. Of course an even simpler, Gaussian, approximation is available if ψ is not too different from a constant.

A bit more crudely, but with less effort, we can also attempt to find the range for z from the stationary distribution of $z(t)$. When available, stationary distributions are a good source of approximations for tail probabilities — which is what we are interested in here — as we can often substitute large-z behavior with long-time behavior. The moments $\mathrm{E}(z(T))$, $\mathrm{Var}(z(T))$ of $z(T)$ that follows (9.55) are given by

$$\mathrm{E}\left(z(T)\right) = z_0, \quad \mathrm{Var}\left(z(T)\right) \approx \psi\left(z_0\right)^2 \int_0^T \eta(t)^2 e^{-2\theta(T-t)}\, dt,$$

where we have applied the approximation (9.57). Assuming that (9.57) is reasonable, the stationary distribution of $z(t)$ can be approximated with the Gamma distribution of Proposition 8.3.4; we choose the parameters of the Gamma distribution to match the mean and variance of $z(T)$,

$$\beta = \frac{\mathrm{E}\left(z(T)\right)}{\mathrm{Var}\left(z(T)\right)}, \quad \alpha = \beta \mathrm{E}\left(z(T)\right).$$

The range of z in the ADI scheme can then be established by

$$z_{\min} = F^{-1}\left(q_z/2; \alpha, \beta\right), \quad z_{\max} = F^{-1}\left(1 - q_z/2; \alpha, \beta\right),$$

where $F(q; \alpha, \beta)$ is the Gamma CDF. Finally, we note that we can just use

$$z_{\min} = 0,$$

as long as we use one-sided discretization for boundary conditions at that point, as explained in Section 9.4.4 below.

9.4.3 Discretizing Stochastic Variance

Uniform discretization of z in the PDE (9.56) is rarely the best choice. If we look at the important case of $\psi(z) = \sqrt{z}$, assuming $z(0) = z_0 = 1$, the interval $[z_{\min}, z_{\max}]$ would be something like $[0, 10]$, with the mean of $z(t)$ being 1. Uniformly discretizing the range $[0, 10]$ would tend to put too few points in the interval $[0, 1]$, resulting in poor resolution in an important part of the range (see also Figure 9.2 in Section 9.5.3.1). To provide a remedy, we may recall the discussion in Section 7.4, which considered the transform

$$u(t) = \Psi\left(z(t)\right), \quad \Psi(z) = \int_{z_0}^{z} \frac{dy}{\psi(y)}. \tag{9.58}$$

Applying Ito's lemma, we get

$$du(t) = \frac{1}{\psi\left(\Psi^{-1}\left(u(t)\right)\right)} \times \left(\theta\left(z_0 - \Psi^{-1}\left(u(t)\right)\right) - \frac{1}{2}\frac{\psi'\left(\Psi^{-1}\left(u(t)\right)\right)}{\psi\left(\Psi^{-1}\left(u(t)\right)\right)}\eta(t)^2\right) dt + \eta(t)\, dZ(t). \quad (9.59)$$

Noticing that the diffusion coefficient of $u(t)$ is not state-dependent, it appears reasonable to construct the grid in z-space from a uniform discretization in u. For this, suppose $N_z + 1$ points are used for the z-domain. We then define the grid $\{\zeta_n\}_{n=0}^{N_z}$ for z by the condition that $u_n \triangleq \Psi(\zeta_n)$ are spaced uniformly over $[\Psi(z_{\min}), \Psi(z_{\max})]$, so that

$$\begin{aligned} u_n &= \Psi\left(z_{\min}\right) + \frac{n}{N_z}\left(\Psi\left(z_{\max}\right) - \Psi\left(z_{\min}\right)\right), \\ \zeta_n &= \Psi^{-1}\left(u_n\right) \\ &= \Psi^{-1}\left(\Psi\left(z_{\min}\right) + \frac{n}{N_z}\left(\Psi\left(z_{\max}\right) - \Psi\left(z_{\min}\right)\right)\right), \quad n = 0, \ldots, N_z. \end{aligned}$$

To give an example, consider the square root case $\psi(z) = \sqrt{z}$ where we have

$$\Psi(z) = \int_{z_0}^{z} \frac{dy}{\sqrt{y}} = 2\left(\sqrt{z} - \sqrt{z_0}\right), \quad \Psi^{-1}(u) = \left(\frac{u}{2} + \sqrt{z_0}\right)^2,$$

such that

$$\zeta_n = \left(\sqrt{z_{\min}} + \frac{n}{N_z}\left(\sqrt{z_{\max}} - \sqrt{z_{\min}}\right)\right)^2, \quad n = 0, \ldots, N_z. \quad (9.60)$$

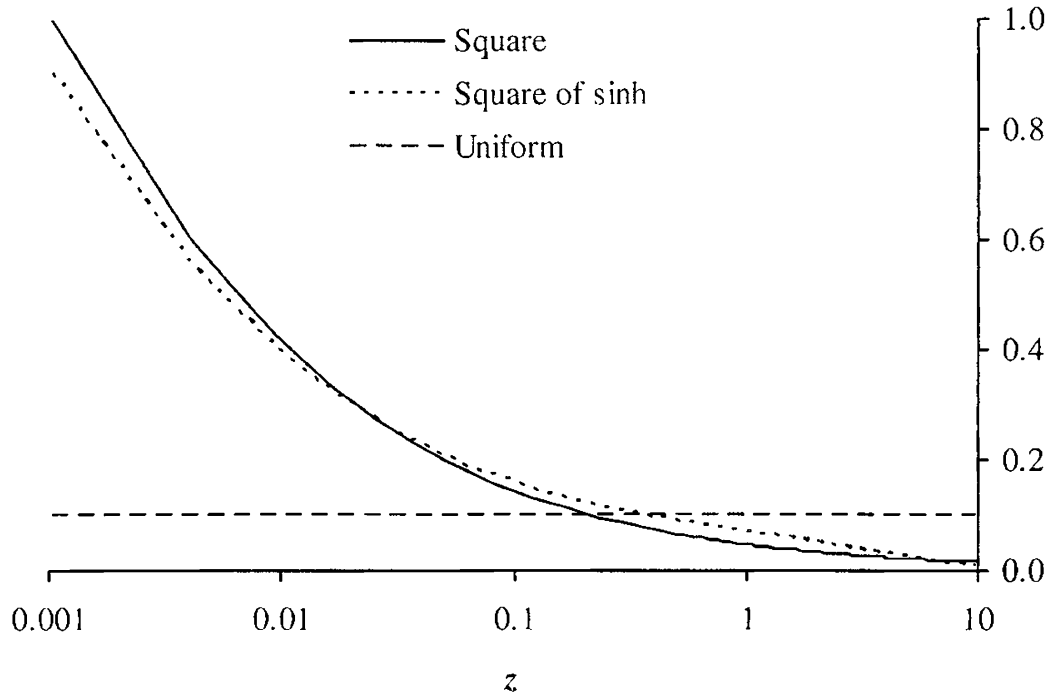
Empirically, it appears that concentrating points around the mean $z = z_0$ further improves numerical properties. We can achieve this effect by applying the sinh transform, see p. 167 of Tavella and Randall [2000], and then using (9.60):

$$\zeta_n = \left(z_0 + \sinh\left(\alpha_{\min} + \frac{n}{N_z}\left(\alpha_{\max} - \alpha_{\min}\right)\right)\right)^2, \quad (9.61)$$

$$\alpha_{\min,\max} = \sinh^{-1}\left(\sqrt{z_{\min,\max}} - z_0\right).$$

To illustrate the discretization strategies above, Figure 9.1 shows the density of grid points over $[z_{\min}, z_{\max}]$ using uniform discretization, quadratic discretization (9.60), and the sinh-quadratic discretization (9.61). As discussed, the quadratic and sinh-quadratic discretizations both increase the density of points in $(0, z_0]$, relative to a uniform discretization. In addition,

Fig. 9.1. Grid Density



Notes: Density of grid points (number of grid points per unit length) as a function of z for three different discretization schemes for z-domain: uniform, quadratic (9.60), and sinh-quadratic (9.61). We assume $z_{\min} = 0$, $z_0 = 1$, $z_{\max} = 10$. The abscissa axis is in logarithmic scale.

the sinh-quadratic scheme places more points around z_0 than does the quadratic scheme.

Let us finally note that instead of drawing on (9.58) as an inspiration for grid discretization in z, we could in principle use the variable u directly in the ADI scheme. Indeed, all that would be required is to rewrite (9.56) in terms of u, S and apply a uniform discretization to u. However, the drift of $u(t)$ is rather complicated and, importantly, grows to infinity as $u \to 0$ in the special case of $\psi(z) = \sqrt{z}$, see (9.59). A scheme that can handle large values of the drift robustly, such as the upwinding scheme from Section 2.6.1, would therefore be a necessity.

9.4.4 Boundary Conditions for Stochastic Variance

Practical experience shows that numerical schemes for solving the PDE (9.56) are quite robust with respect to the specifications of boundary conditions for z. Any reasonable choice from Chapter 2 appears to work well, including the standard $\partial^2 V/\partial z^2 = 0$ for $z = z_{\min}$, $z = z_{\max}$. In the case of $\psi(z) = \sqrt{z}$, if $z_{\min} = 0$, i.e. if we use $z = 0$ as the lower bound on the grid, for best results we should derive the boundary conditions for $z_{\min}$ from the PDE itself, see Section 2.2.2. Setting $z = 0$ in (9.56) we obtain

$$0 = \frac{\partial}{\partial t} V(t, 0, S) + \theta z_0 \frac{\partial}{\partial z} V(t, 0, S), \tag{9.62}$$

a boundary condition of Neumann type. The validity of this boundary condition is intuitively justified by the fact that the solution to the SDE for $z(t)$ is unique, i.e. the behavior of $z(t)$ at the boundary $z = 0$ is determined by the SDE itself — and hence the boundary condition is determined by setting $z = 0$ in the PDE[5]. Incorporation of (9.62) into the finite difference solver would generally require one to discretize the z-derivative by one-sided differences; see Section 10.1.5.2 for details in a slightly more general setting.

Another, also reasonable, specification for the boundary $z = 0$ is obtained from the fact that the square-root process for $z(t)$ is strongly reflecting at $z = 0$, see Proposition 8.3.1. A reflection at the boundary translates into the boundary condition

$$\frac{\partial}{\partial z} V(t, 0, S) = 0$$

(see Karatzas and Shreve [1997]), which is quite similar to (9.62) and is another reasonable choice.

Interestingly, using the correct boundary conditions for the *forward* PDE, i.e. the forward Kolmogorov equation that the density of the process satisfies, is crucial, especially when the Feller condition (Proposition 8.3.1) is violated. As we have no use for forward PDEs for stochastic volatility processes in this book, we refer the reader to Lucic [2008] for the details.

9.4.5 Range for Underlying

To obtain the range

$$[S_{\min}, S_{\max}]$$

for the underlying S, we need to compute the approximate distribution of $S(T)$. Replacing the stochastic variance process with its expected value $\mathrm{E}(z(t)) = z_0$, we obtain

$$dS(t) \approx \varphi(t, S(t)) \sqrt{z_0}\, dW(t).$$

To proceed, we can for example use the connection between option prices and the probability density, and apply various asymptotic results for local volatility models from Section 7.5. In the important special case of a time-dependent linear local volatility

$$\varphi(t, S) = \lambda(t) \left(b(t) S + (1 - b(t)) L \right), \tag{9.63}$$

a reasonable approach is to replace time-dependent $b(t)$ with the effective time-independent skew $\overline{b}$ via Proposition 9.3.4, and then apply a Gaussian approximation:

[5] A formal proof that (9.62) is theoretically correct, at least for payoffs that depend on z only (and not on S), is given in Ekström and Tysk [2008].

$$S(T) \approx \left[\left(\bar{b}S(0) + \left(1-\bar{b}\right)L\right)e^{\xi} - \left(1-\bar{b}\right)L\right]/\bar{b}, \tag{9.64}$$

$$\xi \sim \mathcal{N}\left(-\frac{z_0\bar{b}^2}{2}\int_0^T \lambda(t)^2 dt, z_0\bar{b}^2\int_0^T \lambda(t)^2 dt\right).$$

As ξ is Gaussian, it is easy to find $[\xi_{\min}, \xi_{\max}]$ so that

$$\mathrm{P}\left(\xi < \xi_{\min}\right) = \mathrm{P}\left(\xi > \xi_{\max}\right) = q_S/2$$

for a given small probability $q_S > 0$. This trivially translates into the range for $S(T)$.

9.4.6 Discretizing the Underlying

The representation (9.64) proves useful for discretizing S as well. One approach is to discretize S so that the grid is uniform in ξ,

$$\begin{aligned} S_n &= \left[\left(\bar{b}S(0) + \left(1-\bar{b}\right)L\right)e^{\xi_n} - \left(1-\bar{b}\right)L\right]/\bar{b}, \\ \xi_n &= \xi_{\min} + \frac{n}{N_S}\left(\xi_{\max} - \xi_{\min}\right), \\ n &= 0, \ldots, N_S, \end{aligned}$$

where N_S is the grid size. Alternatively, we can apply a transformation

$$y(S) = \ln\left(\frac{\bar{b}S + \left(1-\bar{b}\right)L}{\bar{b}S(0) + \left(1-\bar{b}\right)L}\right),$$

rewrite the PDE (9.56) in y instead of S, and discretize in y uniformly.

To conclude we note that even if $\varphi(t,S)$ is not of the form (9.63), we can always approximate it as such in order to compute the effective $\bar{b}$ that is then used in discretization for S or in the mapping $S \to y$. Alternatively, we can always employ the same strategy (integral variable transform) that was advocated in Section 9.4.3 for z — which is what we used in Section 7.4 for discretizing local volatility models as well.

9.5 Monte Carlo Method

For generic stochastic volatility models such as (9.54)–(9.55), little can be said about Monte Carlo simulation that has not already been covered in Chapter 3. For any particular model parameterization, however, special-purpose discretization schemes can be constructed that have significant computational advantages over, say, the general-purpose Ito-Taylor schemes in Section 3.2.6. To demonstrate, we shall here specialize to the standard SV model, i.e. we consider the system

$$dS(t) = \lambda\left(bS(t) + (1-b)\,L\right)\sqrt{z(t)}\,dW(t), \tag{9.65}$$

$$dz(t) = \theta\left(z_0 - z(t)\right)\,dt + \eta\sqrt{z(t)}\,dZ(t), \tag{9.66}$$

with $\langle dZ(t), dW(t)\rangle = \rho\,dt$ and $z(0) = z_0$. Our primary objective is to establish a scheme that allows us to time-discretize the SV model dynamics in an efficient manner; as it turns out, this is a surprisingly challenging, particularly for the z-process. We shall consequently deal with the Monte Carlo simulation of the SV model in a fairly careful manner, listing a number of schemes with different efficiency/bias trade-offs.

Remark 9.5.1. While we have assumed that parameters in the SV process are constants, all that is ultimately required is that parameters are piecewise constant on the simulation time line. As such, the schemes we suggest will also apply to time-dependent dynamics.

9.5.1 Exact Simulation of Variance Process

According to Proposition 8.3.2, the distribution of $z(t+\Delta)$ given $z(t)$ is known in closed form, and generation of a random sample of $z(t+\Delta)$ given $z(t)$ can be done entirely bias-free by sampling from a non-central chi-square distribution. Using the fact that a non-central chi-square distribution can be seen as a regular chi-square distribution with Poisson-distributed degrees of freedom (see Section 3.1.1.3), the following algorithm can be used.

1. Draw a Poisson random variable N, with mean $\frac{1}{2}z(t)n(t,t+\Delta)$ (here $n(t,T)$ is defined in (8.6)).
2. Given N, draw a regular chi-square random variable χ^2_v, with $v = d+2N$ degrees of freedom (d is defined in (8.6)).
3. Set $z(t+\Delta) = \chi^2_v \cdot \exp(-\theta\Delta)/n(t,t+\Delta)$.

Steps 1 and 3 of this algorithm are straightforward, and Step 2 can be accomplished using the acceptance-rejection technique discussed in Section 3.1.1.2.

As mentioned in Section 3.1.1.3, if $d > 1$ it may be numerically advantageous to use a different algorithm, based on the relation

$$\chi'^2_d(\gamma) \stackrel{d}{=} \left(Z + \sqrt{\gamma}\right)^2 + \chi^2_{d-1}, \quad d > 1, \tag{9.67}$$

where $\stackrel{d}{=}$ denotes equality in distribution, $\chi'^2_d(\gamma)$ is a non-central chi-square variable with d degrees of freedom and non-centrality parameter γ, and Z is an ordinary $\mathcal{N}(0,1)$ Gaussian variable. We trust that the reader can complete the details on application of (9.67) in a simulation algorithm for $z(t+\Delta)$.

One might think that the existence of an exact simulation scheme for $z(t+\Delta)$ would settle once and for all the question of how to generate paths

of the square-root process. In practice, however, several complications may arise with the application of the algorithm above. Indeed, the scheme is quite complex compared with many standard SDE discretization schemes and may not fit smoothly into existing software architecture for SDE simulation routines. Also, computational speed may be an issue, and the application of acceptance-rejection sampling will potentially cause a "scrambling effect" when process parameters are perturbed[6], resulting in poor convergence of numerically computed sensitivities, see Section 3.3. While caching techniques can be designed to overcome some of these issues, storage, look-up, and interpolation of such a cache pose their own challenges. Further, the basic scheme above provides no explicit link between the paths of the Brownian motion $Z(t)$ and that of $z(t)$, complicating applications in which, say, multiple correlated Brownian motions need to be advanced through time.

In light of the discussion above, it seems reasonable to also investigate the application of simpler simulation algorithms. These will typically exhibit a bias — in the sense discussed in Section 3.2.8 — for finite values of Δ, but convenience and speed may more than compensate for this, especially if the bias is small and easy to control by reduction of step size. We proceed to discuss several classes of such schemes.

9.5.2 Biased Taylor-Type Schemes for Variance Process

9.5.2.1 Euler Schemes

Going forward, let us use $\widehat{z}$ to denote a discrete-time (biased) approximation to z. A classical approach to simulating a path $\widehat{z}$ involves the application of Ito-Taylor expansions, suitably truncated, see Sections 3.2.3 and 3.2.6 for details. The simplest such scheme is the Euler scheme, a direct application of which would here give

$$\widehat{z}(t+\Delta) = \widehat{z}(t) + \theta(z_0 - \widehat{z}(t))\Delta + \eta\sqrt{\widehat{z}(t)}Z\sqrt{\Delta}, \tag{9.68}$$

where Z is a $\mathcal{N}(0,1)$ Gaussian variable. One immediate (and fatal) problem with (9.68) is that the discrete process $\widehat{z}$ can become negative with non-zero probability. The first time this happens on a path, computation of $\sqrt{\widehat{z}(t)}$ will be impossible and the time-stepping scheme will fail. To get around this problem, several remedies have been proposed in the literature, starting with the suggestion in Kloeden and Platen [2000] that one simply replace $\sqrt{\widehat{z}(t)}$ in (9.68) with $\sqrt{|\widehat{z}(t)|}$. Lord et al. [2006] review a number of similar "fixes" and conclude that the following works best:

$$\widehat{z}(t+\Delta) = \widehat{z}(t) + \theta(z_0 - \widehat{z}(t)^+)\Delta + \eta\sqrt{\widehat{z}(t)^+}Z\sqrt{\Delta}. \tag{9.69}$$

[6] After a perturbation of parameters, the number of rejected samples in the Monte Carlo trial will likely change.

In Lord et al. [2006] this scheme is denoted "full truncation"; its main characteristic is that the process for $\widehat{z}$ is allowed to go below zero, at which point $\widehat{z}$ becomes deterministic with an upward drift of θz_0.

9.5.2.2 Higher-Order Schemes

The scheme (9.69) has first-order weak convergence, i.e. expectations of functions of $\widehat{z}$ will approach their true values as $O(\Delta)$. To improve convergence, it is tempting to apply a Milstein scheme (see Section 3.2.6.3), the most basic of which is

$$\widehat{z}(t+\Delta) = \widehat{z}(t) + \theta(z_0 - \widehat{z}(t))\Delta + \eta\sqrt{\widehat{z}(t)}\, Z\sqrt{\Delta} + \frac{1}{4}\eta^2\Delta\left(Z^2 - 1\right).$$

As was the case for (9.68), this scheme has a positive probability of generating negative values of $\widehat{z}$ and therefore cannot be used without suitable modifications. Kahl and Jäckel [2006] list several other Milstein-type schemes, some of which allow for a certain degree of control over the likelihood of generating negative values. One interesting variation is the *implicit Milstein scheme*, defined as

$$\widehat{z}(t+\Delta) = \frac{\widehat{z}(t) + \theta z_0\Delta + \eta\sqrt{\widehat{z}(t)}Z\sqrt{\Delta} + \frac{1}{4}\eta^2\Delta\left(Z^2 - 1\right)}{1 + \theta\Delta}. \tag{9.70}$$

It is easy to verify that this discretization scheme will result in strictly positive paths for the z process if $4\theta z_0 > \eta^2$. For cases where this bound does not hold, it will be necessary to modify (9.70) to prevent problems with the computation of $\sqrt{\widehat{z}(t)}$. For instance, whenever $\widehat{z}(t)$ drops below zero, we could use (9.69) rather than (9.70).

Under certain sufficient regularity conditions, we have seen in Chapter 3 that Milstein schemes have second-order weak convergence. Due to the presence of a square root in (9.66), these sufficient conditions are violated here, and one should not expect (9.70) to have second-order convergence for all parameter values, even the ones that satisfy $4\theta z_0 > \eta^2$. Numerical tests of Milstein schemes for square-root processes can be found in Kahl and Jäckel [2006] and Glasserman [2004]; overall these schemes perform fairly well in benign parameter regimes, but are typically less robust than the Euler scheme.

9.5.3 Moment Matching Schemes for Variance Process

9.5.3.1 Log-normal Approximation

The simulation schemes introduced in Section 9.5.2 all suffer to various degrees from an inability to keep the path of z non-negative. One, rather obvious, way around this is to draw $\widehat{z}(t+\Delta)$ from a user-selected probability

distribution that i) is reasonably close to the true distribution of $z(t+\Delta)$; and ii) is certain not to produce negative values[7]. To ensure that i) is satisfied, it is natural to select the parameters of the chosen distribution to match one or more of the true moments for $z(t+\Delta)$, conditional upon $z(t) = \widehat{z}(t)$. For instance, if we assume that the true distribution of $z(t+\Delta)$ is well approximated by a log-normal distribution with parameters μ and σ^2, we write (see Andersen and Brotherton-Ratcliffe [2005])

$$\widehat{z}(t+\Delta) = e^{\mu+\sigma Z}, \tag{9.71}$$

where Z is a standard Gaussian random variable, and μ, σ are chosen to satisfy

$$e^{\mu+\frac{1}{2}\sigma^2} = \mathrm{E}\left(z(t+\Delta)|z(t) = \widehat{z}(t)\right), \tag{9.72}$$

$$e^{2\left(\mu+\frac{1}{2}\sigma^2\right)}\left(e^{\sigma^2}-1\right) = \mathrm{Var}\left(z(t+\Delta)|z(t) = \widehat{z}(t)\right). \tag{9.73}$$

The results in Corollary 8.3.3 can be used to compute the right-hand sides of this system of equations, which can then easily be solved analytically for μ and σ.

As is the case for many other schemes, (9.71) works best if the Feller condition, as defined in Proposition 8.3.1, is satisfied. If not, the lower tail of the log-normal distribution is often too thin to capture the true distribution shape of $\widehat{z}(t+\Delta)$ — see Figure 9.2 for an example.

9.5.3.2 Truncated Gaussian

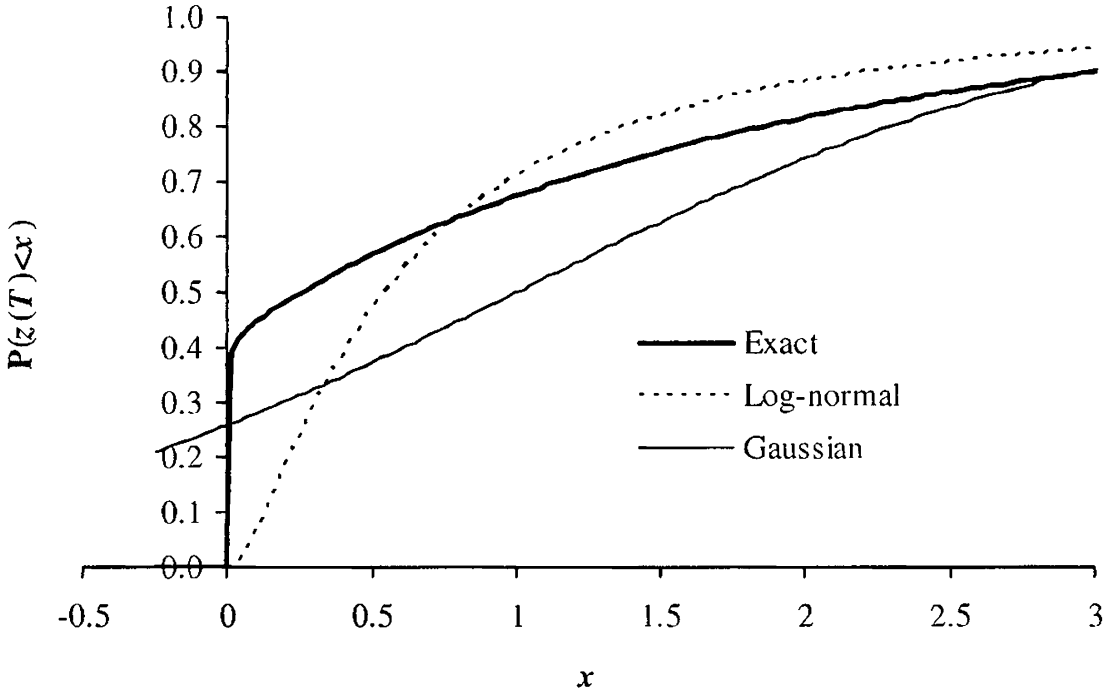
Figure 9.2 demonstrates that the density of $z(t+\Delta)|z(t)$ may sometimes be nearly singular at the origin. To accommodate this, one could contemplate inserting an actual singularity through outright truncation at the origin of a distribution that may otherwise go negative. Using a Gaussian distribution for this, say, one could write

$$\widehat{z}(t+\Delta) = (\mu+\sigma Z)^+, \tag{9.74}$$

where μ and σ are determined by moment-matching, along the same lines as in Section 9.5.3.1 above. While this moment-matching exercise cannot be done in entirely analytical fashion, a number of caching tricks outlined in Andersen [2008] can be used to make the determination of μ and σ essentially instantaneous. As documented in Andersen [2008], the scheme

[7] As pointed out in Section 3.2.2, weak consistency — convergence of the first and second moments in the discretization scheme to those of the original SDE — is sufficient (together with some regularity conditions) for weak convergence. Hence, the actual distribution used for time-stepping can be chosen almost arbitrarily. Of course, matching other characteristics of the actual distribution may substantially improve the performance of the scheme.

Fig. 9.2. Cumulative Distribution of z



Notes: The figure shows the cumulative distribution function for $z(T)$ given $z(0)$, with $T = 0.1$. Model parameters were $z(0) = z_0 = 1$, $\theta = 50\%$, and $\eta = 100\%$. The log-normal and Gaussian distributions in the graph were parameterized by matching mean and variances to the exact distribution of $z(T)$.

(9.74) is robust and generally has attractive convergence properties when applied to standard option pricing problems. Being fundamentally Gaussian when $\widehat{z}(t)$ is far from the origin, (9.74) is qualitatively similar to the Euler scheme (9.69), although performance of (9.74) is typically somewhat better than (9.69). Unlike (9.69), the truncated Gaussian scheme (9.74) also ensures, by construction, that negative values of $\widehat{z}(t + \Delta)$ cannot be attained.

9.5.3.3 Quadratic-Exponential

We finish our discussion of biased schemes for (9.66) with a more elaborate moment-matched scheme, based on a combination of a squared Gaussian and an exponential distribution. In this scheme, for large values of $\widehat{z}(t)$, we write

$$\widehat{z}(t + \Delta) = a\,(b + Z)^2\,, \tag{9.75}$$

where Z is a standard Gaussian random variable, and a and b are certain constants, to be determined by moment-matching. The constants a and b will depend on the time step Δ and $\widehat{z}(t)$, as well as the parameters of the SDE for $z(t)$. While based on well-established asymptotics for the non-central chi-square distribution (see Andersen [2008]), formula (9.75) does not work well for low values of $\widehat{z}(t)$ — in fact, the moment-matching exercise fails to

work — so we supplement it with a scheme to be used when $\widehat{z}(t)$ is small. Examination of the true conditional density for $z(t+\Delta)|z(t)$ shows that the upper density tail decays exponentially, so a good choice is to approximate the distribution of $\widehat{z}(t+\Delta)$ with

$$\mathrm{P}\left(\widehat{z}(t+\Delta) \in [x, x+dx]\right) = \left(p\delta(x) + \beta(1-p)e^{-\beta x}\right)\, dx, \quad x \geq 0, \quad (9.76)$$

where δ is the Dirac delta function, and p and β are non-negative constants to be determined. As in the scheme in Section 9.5.3.2, we have a probability mass at the origin, but now the strength of this mass (p) is explicitly specified, rather than implied from other parameters. It can be verified that if $p \in [0,1]$ and $\beta \geq 0$, then (9.76) constitutes a valid density function.

Assuming that we have determined a and b, Monte Carlo sampling from (9.75) is trivial. To draw samples in accordance with (9.76), we can generate a cumulative distribution function

$$\Psi(x) = \mathrm{P}\left(\widehat{z}(t+\Delta) \leq x\right) = p + (1-p)\left(1 - e^{-\beta x}\right), \quad x \geq 0. \quad (9.77)$$

Here, the inverse of Ψ is readily computable:

$$\Psi^{-1}(u) = \Psi^{-1}(u; p, \beta) = \begin{cases} 0, & 0 \leq u \leq p, \\ \beta^{-1} \ln\left(\frac{1-p}{1-u}\right), & p < u < 1. \end{cases} \quad (9.78)$$

By the standard inverse distribution function method from Section 3.1.1.1, we thus get the simple sampling scheme

$$\widehat{z}(t+\Delta) = \Psi^{-1}\left(U_z; p, \beta\right) \quad (9.79)$$

where U_z is a draw from a uniform distribution. Note that this scheme is extremely fast to execute.

Equations (9.75) and (9.79) together define the QE (for *Q*uadratic-*E*xponential) discretization scheme. What remains is the determination of the constants a, b, p, and β, as well as a rule for when to switch from (9.75) to (9.79). The first problem is easily settled by moment-matching techniques, as shown in the following two propositions. We omit their straightforward proofs, which can be found in Andersen [2008].

Proposition 9.5.2. *Let*

$$m \triangleq \mathrm{E}\left(z(t+\Delta)|z(t) = \widehat{z}(t)\right), \quad s^2 \triangleq \mathrm{Var}\left(z(t+\Delta)|z(t) = \widehat{z}(t)\right),$$

and set $\psi = s^2/m^2$. *Provided that* $\psi \leq 2$, *set*

$$b^2 = 2\psi^{-1} - 1 + \sqrt{2\psi^{-1}}\sqrt{2\psi^{-1} - 1} \geq 0 \quad (9.80)$$

and

$$a = \frac{m}{1+b^2}. \quad (9.81)$$

Let $\widehat{z}(t+\Delta)$ *be as defined in (9.75); then* $\mathrm{E}(\widehat{z}(t+\Delta)) = m$ *and* $\mathrm{Var}(\widehat{z}(t+\Delta)) = s^2$.

Proposition 9.5.3. *Let m, s, and ψ be as defined in Proposition 9.5.2. Assume that $\psi \geq 1$ and set*

$$p = \frac{\psi - 1}{\psi + 1} \in [0, 1), \tag{9.82}$$

and

$$\beta = \frac{1 - p}{m} = \frac{2}{m(\psi + 1)} > 0. \tag{9.83}$$

Let $\widehat{z}(t + \Delta)$ be sampled from (9.79); then $\mathrm{E}(\widehat{z}(t + \Delta)) = m$ and $\mathrm{Var}(\widehat{z}(t + \Delta)) = s^2$.

The terms m, s, ψ defined in the two propositions above are explicitly computable from the result in Corollary 8.3.3. For any ψ_c in $[1, 2]$, a valid *switching rule* is to use (9.75) if $\psi \leq \psi_c$ and to sample (9.77) otherwise. The exact value selected for ψ_c is non-critical; $\psi_c = 1.5$ is a natural choice.

9.5.3.4 Summary of QE Algorithm

As the QE algorithm is fairly complex, let us for convenience summarize the entire sampling algorithm step-by-step.

Assume that some arbitrary level $\psi_c \in [1, 2]$ has been selected. The detailed algorithm for the QE simulation step from $\widehat{z}(t)$ to $\widehat{z}(t + \Delta)$ is then:

1. Given $z(t) = \widehat{z}(t)$, compute $m = \mathrm{E}(z(t + \Delta)|z(t) = \widehat{z}(t))$ and $s^2 = \mathrm{Var}(z(t + \Delta)|z(t)\widehat{z}(t))$ from Corollary 8.3.3.
2. Compute $\psi = s^2/m^2$.
3. Draw a uniform random number U_z.
4. **If** $\psi \leq \psi_c$:
   a) Compute a and b from equations (9.81) and (9.80).
   b) Compute $Z = \Phi^{-1}(U_z)$.
   c) Use (9.75), i.e. set $\widehat{z}(t + \Delta) = a(b + Z)^2$.
5. **Otherwise**, if $\psi > \psi_c$:
   a) Compute p and β according to equations (9.82) and (9.83).
   b) Use (9.79), i.e. set $\widehat{z}(t + \Delta) = \Psi^{-1}(U_z; p, \beta)$, where Ψ^{-1} is given in (9.78).

For efficiency, exponentials used in computation of m and s^2 should be pre-cached. The inversion of the Gaussian CDF in Step 4 can be done using the techniques described in Section 3.1.1.1.

The quadratic-exponential (QE) scheme outlined above is typically the most accurate of the biased schemes discussed here. Indeed, in most practical application the bias introduced by the scheme is statistically undetectable at the levels of Monte Carlo noise typically encountered in practical applications; see Andersen [2008] for numerical tests under a range of challenging conditions. Variations on the QE scheme without an explicit singularity in zero can also be found in Andersen [2008].

9.5.4 Broadie-Kaya Scheme for the Underlying

At this point, we are done discussing simulation schemes for the z-process, and now turn to the underlying process (9.65) itself.

For numerical work, it is useful to work with a logarithmic transformation of $S(t)$, rather than $S(t)$ itself. Specifically, we set

$$X(t) = \frac{bS(t) + (1-b)\,L}{bS(0) + (1-b)\,L},$$

the logarithm of which, from Proposition 8.3.6, satisfies the SDE

$$d\ln X(t) = -\frac{1}{2}\lambda^2 b^2 z(t)\,dt + \lambda b\sqrt{z(t)}\,dW(t). \tag{9.84}$$

As demonstrated in Broadie and Kaya [2006], it is possible to simulate (9.84) bias-free. To show this, first integrate the SDE for $z(t)$ in (9.66) and rearrange:

$$\int_t^{t+\Delta} \sqrt{z(u)}\,dZ(u) = \frac{1}{\eta}\left(z(t+\Delta) - z(t) - \theta z_0 \Delta + \theta \int_t^{t+\Delta} z(u)\,du\right). \tag{9.85}$$

Performing a Cholesky decomposition we can also write

$$d\ln X(t) = -\frac{1}{2}\lambda^2 b^2 z(t)\,dt + \lambda b\left(\rho\sqrt{z(t)}\,dZ(t) + \sqrt{1-\rho^2}\sqrt{z(t)}\,dB(t)\right),$$

where B is a Brownian motion independent of Z. An integration then yields

$$\begin{aligned}\ln X(t+\Delta) = \ln X(t) &+ \frac{\rho\lambda b}{\eta}\left(z(t+\Delta) - z(t) - \theta z_0 \Delta\right) \\ &+ \left(\frac{\theta\rho\lambda b}{\eta} - \frac{\lambda^2 b^2}{2}\right)\int_t^{t+\Delta} z(u)\,du + \lambda b\sqrt{1-\rho^2}\int_t^{t+\Delta}\sqrt{z(u)}\,dB(u),\end{aligned} \tag{9.86}$$

where we have used (9.85). Conditional on $z(t+\Delta)$ and $\int_t^{t+\Delta} z(u)\,du$, it is clear that the distribution of $\ln X(t+\Delta)$ is Gaussian with easily computable moments. After first sampling $z(t+\Delta)$ bias-free from the non-central chi-square distribution (as described in Section 9.5.1), one then performs the following steps:

1. Conditional on $z(t+\Delta)$ (and $z(t)$) draw a bias-free sample of $I = \int_t^{t+\Delta} z(u)\,du$.
2. Conditional on $z(t+\Delta)$ and I, use (9.86) to draw a sample of $\ln X(t+\Delta)$ from a Gaussian distribution.

While execution of the second step is straightforward, the first one is decidedly not, as the conditional distribution of the integral I is not known in closed form. In Broadie and Kaya [2006], the authors instead derive a characteristic function, which they numerically Fourier-invert to generate the cumulative distribution function for I, given $z(t+\Delta)$ and $z(t)$. Numerical inversion of this distribution function over a uniform random variable finally allows for generation of a sample of I. The total algorithm requires great care in numerical discretization to prevent introduction of noticeable biases and is further complicated by the fact that the characteristic function for I contains two modified Bessel functions.

The Broadie-Kaya algorithm is bias-free by construction, but its complexity and lack of speed is problematic in many applications. Smith [2007] and Glasserman and Kim [2008] discuss various techniques to improve computational efficiency of the basic algorithm, but even with such improvements it is safe to say that the method is competitive only for applications that involve long time steps and require very high accuracy (and neither are the norm for fixed income applications).

9.5.5 Other Schemes for the Underlying

9.5.5.1 Taylor-Type Schemes

In their examination of "fixed" Euler-schemes, Lord et al. [2006] suggest simulation of the Heston model by combining (9.69) with the following scheme for $\ln X$:

$$\ln \widehat{X}(t+\Delta) = \ln \widehat{X}(t) - \frac{1}{2}\lambda^2 b^2 \widehat{z}(t)^+ \Delta + \lambda b \sqrt{\widehat{z}(t)^+}\, W \sqrt{\Delta}\,, \qquad (9.87)$$

where W is a Gaussian $\mathcal{N}(0,1)$ draw, correlated to Z in (9.69) with correlation coefficient ρ. For the periods where $\widehat{z}$ drops below zero in (9.69), the process for $\widehat{X}$ comes to a standstill.

Kahl and Jäckel [2006] examine the usage of Ito-Taylor expansions for joint simulation of $X(t)$ and $z(t)$, proposing several concrete schemes. As these schemes are rather complex, we simply refer the reader to Kahl and Jäckel [2006] for the details. Andersen [2008] tests the most prominent of the schemes in Kahl and Jäckel [2006] (the "IJK" scheme) and concludes that the scheme works well in benign parameter ranges, but has a tendency to deteriorate when parameters are made more extreme.

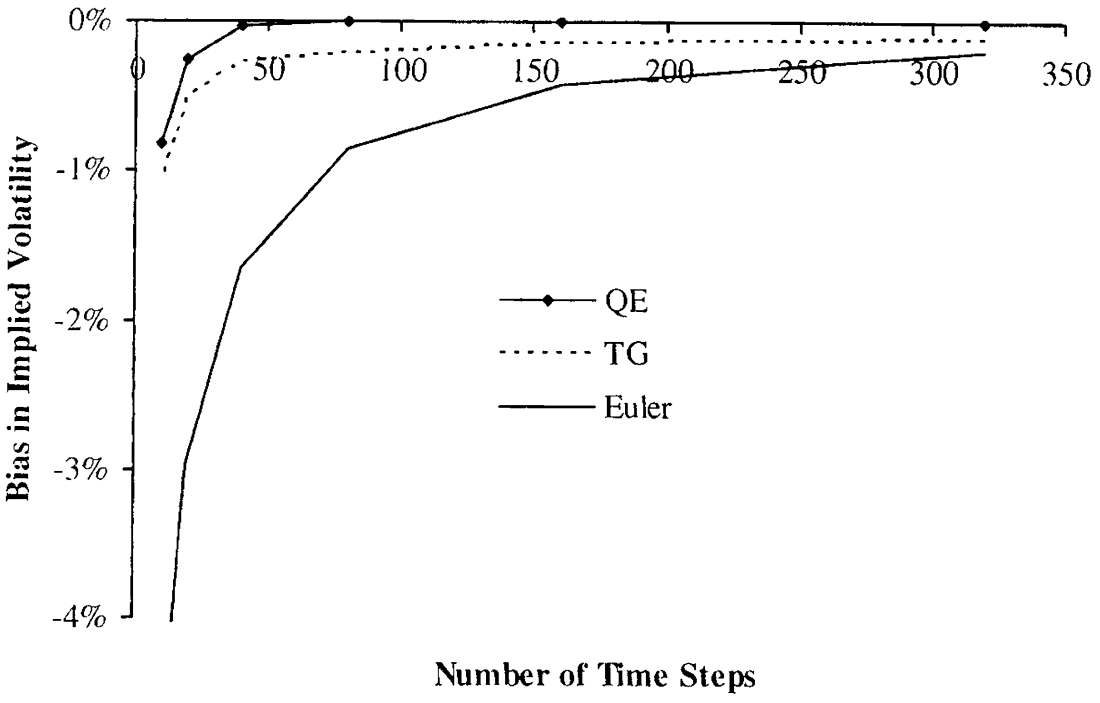
9.5.5.2 Simplified Broadie-Kaya

We recall from the discussion earlier that the complicated part of the Broadie-Kaya algorithm was the computation of $\int_t^{t+\Delta} z(u)\,du$, conditional on $z(t)$ and $z(t+\Delta)$. Andersen [2008] suggests a naive, but effective, approximation, based on the idea that

$$\int_t^{t+\Delta} z(u)\,du \approx \Delta\left[\gamma_1 z(t) + \gamma_2 z(t+\Delta)\right] , \tag{9.88}$$

for certain constants γ_1 and γ_2. The constants γ_1 and γ_2 can be found by moment-matching techniques (using calculations similar to those from the proof of Theorem 9.3.6, or results from Dufresne [2001], p. 16), but Andersen [2008] presents evidence that it will often be sufficient to use either an Euler-like setting ($\gamma_1 = 1$, $\gamma_2 = 0$) or a central discretization ($\gamma_1 = \gamma_2 = \frac{1}{2}$). In any case, (9.88) combined with (9.86) gives rise to a scheme for Y-simulation that can be combined with any basic algorithm that can produce $\widehat{z}(t)$ and $\widehat{z}(t+\Delta)$. Andersen [2008] contains numerical results for the case where $\widehat{z}(t)$ and $\widehat{z}(t+\Delta)$ are simulated by the algorithms in Sections 9.5.3.2 and 9.5.3.3; results are excellent, particularly when the QE algorithm in Section 9.5.3.3 is used to sample $\widehat{z}$. Figure 9.3 reproduces some sample convergence results from Andersen [2008].

Fig. 9.3. Convergence of Bias



Notes: The figure shows the convergence of the call option price bias in implied volatility terms, as a function of the number of time steps per path ($=T/\Delta$). The Euler scheme graph was computed using the full truncation scheme in (9.69), and the QE scheme used $\gamma_1 = \gamma_2 = 0.5$ and $\psi_c = 1.5$. Model parameters: $S(0) = L = 100$, $b = 1$, $z(0) = z_0 = 1$, $\theta = 0.5$, $\rho = -0.9$, $\eta = 1$, $\lambda = 20\%$. The option maturity is $T = 10$ and the strike is $K = 100$. The bias was estimated from 1,000,000 simulation paths, using the Fourier technique to establish exact prices.

9.5.5.3 Martingale Correction

Finally, let us note that some of the schemes outlined above, including the one in Section 9.5:5.2, will generally not lead to martingale behavior of $\widehat{X}$; that is, $\mathrm{E}(\widehat{X}(t+\Delta)|\widehat{X}(t)) \neq \widehat{X}(t)$. For the cases where the error $e = \mathrm{E}(\widehat{X}(t+\Delta)|\widehat{X}(t)) - \widehat{X}(t)$ is analytically computable, it is, however, straightforward to remove the bias by simply adding $-e$ to the sample value for $\widehat{X}(t+\Delta)$. Andersen [2008] gives several examples of this idea and shows that, for the QE scheme at least, the improvements from martingale correction are minor.

9.A Appendix: Proof of Proposition 9.3.4

Let us fix a time horizon $T > 0$. Let $f(t,x)$ be a local volatility function,

$$f(t,x) \in C^1([0,T] \times \mathbb{R}),$$

satisfying the usual growth requirements. Let $\lambda(t)$, $t \in [0,T]$, be a function of time only. Fix $x_0 \in \mathbb{R}$. For any $\epsilon \geq 0$, define a rescaled local volatility function

$$f_\epsilon(t,x) = f\left(t\epsilon^2, x_0 + (x - x_0)\,\epsilon\right). \tag{9.89}$$

Without loss of generality we can assume that

$$f(t,x_0) \equiv 1, \quad t \in [0,T],$$

which implies

$$f_\epsilon(t,x_0) \equiv 1, \quad t \in [0,T]. \tag{9.90}$$

Let $w(t)$, $t \in [0,T]$, be a weight function such that

$$\int_0^T w(t)\,dt = 1, \tag{9.91}$$

and let us define an averaged local volatility function

$$\overline{f}_\epsilon(x)^2 = \int_0^T f_\epsilon(t,x)^2\, w(t)\,dt. \tag{9.92}$$

Define two families of diffusions indexed by ϵ,

$$\begin{aligned} dX_\epsilon(t) &= f_\epsilon(t, X_\epsilon(t))\sqrt{z(t)}\lambda(t)\,dW(t), \quad X_\epsilon(0) = x_0, \\ dY_\epsilon(t) &= \overline{f}_\epsilon(Y_\epsilon(t))\sqrt{z(t)}\lambda(t)\,dW(t), \quad Y_\epsilon(0) = x_0, \end{aligned}$$

for $t \in [0,T]$, where $z(t)$ is defined by (9.31). The following theorem can be found in Piterbarg [2005b].

Theorem 9.A.1. *If the weight function $w(t)$ is set to equal $w_T(t)$, where*

$$w_T(t) \triangleq \frac{v(t)^2\lambda(t)^2}{\int_0^T v(t)^2\lambda(t)^2\,dt}, \tag{9.93}$$
$$v(t)^2 = \mathrm{E}\left(z(t)\left(X_0(t)-x_0\right)^2\right),$$

then, as $\epsilon \to 0$,

$$\mathrm{E}\left(\left(X_\epsilon(T)-x_0\right)^2\right) - \mathrm{E}\left(\left(Y_\epsilon(T)-x_0\right)^2\right) = o\left(\epsilon^2\right), \tag{9.94}$$
$$\mathrm{E}\left(\left(X_\epsilon(T)-x_0\right)^3\right) - \mathrm{E}\left(\left(Y_\epsilon(T)-x_0\right)^3\right) = o\left(\epsilon^2\right). \tag{9.95}$$

Proof. The stochastic variance process $z(t)$ is Markovian. We denote its infinitesimal generator by L^z,

$$L^z : \phi \longmapsto \theta\left(z_0 - z\right)\frac{\partial\phi}{\partial z} + \frac{1}{2}\eta^2 z\frac{\partial^2\phi}{\partial z^2}.$$

We note that the process $X_0(t)$ $(\equiv Y_0(t))$ satisfies the following SDE,

$$dX_0(t) = \sqrt{z(t)}\lambda(t)\,dW(t), \quad X_0(0) = x_0.$$

Let us denote the Markov semi-group of operators that corresponds to the process $(X_0(t), z(t))$ by $P_0\left(s,t\right)$, and the time-dependent infinitesimal generator by $L_0(t)$,

$$\left[P_0\left(s,t\right)\phi\right](x,z) = \mathrm{E}_s\left(\phi\left(X_0(t),z(t)\right)\middle|\, X_0(s)=x, z(s)=z\right),$$
$$L_0(t) : \phi \longmapsto \frac{1}{2}\lambda(t)^2 z\frac{\partial^2\phi}{\partial x^2} + L^z.$$

Let us denote the same for $(X_\epsilon(t), z(t))$ and for $(Y_\epsilon(t), z(t))$ by $P_\epsilon^X(s,t)$, $L_\epsilon^X(t)$ and $P_\epsilon^Y(s,t)$, $L_\epsilon^Y(t)$, respectively.

From the general operator semigroup theory (see Ethier and Kurtz [1986]) it follows that

$$P_\epsilon^Y\left(0,T\right) = P_\epsilon^X\left(0,T\right) + \int_0^T P_\epsilon^Y\left(0,t\right)\left(L_\epsilon^Y(t) - L_\epsilon^X(t)\right)P_\epsilon^X\left(t,T\right)\,dt. \tag{9.96}$$

By Proposition 8.4.13 applied to $f(x) = (x-x_0)^2/2$ and $f(x) = (x-x_0)^3/6$,

$$\frac{1}{2}\mathrm{E}\left(X_\epsilon(T)-x_0\right)^2 = \int_{-\infty}^{x_0}\mathrm{E}\left(K - X_\epsilon(T)\right)^+ dK + \int_{x_0}^{\infty}\mathrm{E}\left(X_\epsilon(T)-K\right)^+ dK,$$
$$\frac{1}{6}\mathrm{E}\left(X_\epsilon(T)-x_0\right)^3 = \int_{-\infty}^{x_0}\left(K-x_0\right)\mathrm{E}\left(K - X_\epsilon(T)\right)^+ dK$$
$$+ \int_{x_0}^{\infty}\left(K-x_0\right)\mathrm{E}\left(X_\epsilon(T)-K\right)^+ dK,$$

and the same for Y_ϵ. Expressed in terms of the Markovian semigroup,

$$\frac{1}{(i+2)!}\mathrm{E}\left(X_\epsilon(T)-x_0\right)^{i+2}=\int_{-\infty}^{\infty}\left\langle \delta_{x_0,z_0},(K-x_0)^i P_\epsilon^X\left(0,T\right)\pi_K\right\rangle\, dK,$$

(and the same for Y_ϵ) for $i=0,1$, where we have defined the payoff π_K by

$$\pi_K\left(x,z\right)=\begin{cases}\left(x-K\right)^+, & K\geq x_0,\\ \left(K-x\right)^+, & K<x_0.\end{cases}$$

Let us denote

$$\Delta(i)=\frac{1}{(i+2)!}\left(\mathrm{E}\left(Y_\epsilon(T)-x_0\right)^{i+2}-\mathrm{E}\left(X_\epsilon(T)-x_0\right)^{i+2}\right),\quad i=0,1.$$

To prove the theorem, we need to show that with the appropriate choice of weights $w_T(t)$,

$$\Delta(i)=o\left(\epsilon^2\right),\quad \epsilon\to 0,\quad i=0,1. \tag{9.97}$$

Clearly,

$$\Delta(i)=\int_{-\infty}^{\infty}(K-x_0)^i\left\langle \delta_{x_0,z_0},\left(P_\epsilon^Y\left(0,T\right)-P_\epsilon^X\left(0,T\right)\right)\pi_K\right\rangle\, dK.$$

By (9.96) we have,

$$\Delta(i)=\int_{-\infty}^{\infty}(K-x_0)^i \times\left(\int_0^T\left\langle \delta_{x_0,z_0},P_\epsilon^Y\left(0,t\right)\left(L_\epsilon^Y(t)-L_\epsilon^X(t)\right)P_\epsilon^X\left(t,T\right)\pi_K\right\rangle dt\right) dK.$$

After a series of manipulations (see Piterbarg [2005b] for details) we obtain, to order $o(\epsilon^2)$,

$$\Delta(i)=\frac{1}{2}\int_0^T\int\widehat{p}(x,z)\left(x-x_0\right)^i\left(\overline{f}_\epsilon(x)^2-f_\epsilon\left(t,x\right)^2\right)\lambda(t)^2\,dx\,dt, \tag{9.98}$$

$\widehat{p}(t,x)\triangleq\mathrm{E}\left(z(t)\delta(X_0(t)-x_0)\right).$

Expanding f, $\overline{f}$ to the first order around (s,x_0), we obtain

$$\begin{aligned}\delta(t;i) &\triangleq \int \widehat{p}(t,x)(x-x_0)^i\left(\overline{f}_\epsilon(x)^2 - f_\epsilon(t,x)^2\right)dx \\ &= 2\epsilon\left(\frac{\partial f(s\epsilon^2,x_0)}{\partial x} - \int_0^T \frac{\partial f(s\epsilon^2,x_0)}{\partial x} w(s)\,ds\right) \\ &\quad \times \int \widehat{p}(t,x)(x-x_0)^{i+1}\,dx \\ &\quad + \epsilon^2\left(\left[\frac{\partial f(s\epsilon^2,x_0)}{\partial x}\right]^2 - \int_0^T \left[\frac{\partial f(s\epsilon^2,x_0)}{\partial x}\right]^2 w(s)\,ds\right) \\ &\quad \times \int \widehat{p}(t,x)(x-x_0)^{i+2}\,dx \\ &\quad + o(\epsilon^2).\end{aligned}$$

Calculating the integrals, we obtain to order $o(\epsilon^2)$,

$$\delta(t;i) = 2\epsilon v(t)^2\left(\frac{\partial f(s\epsilon^2,x_0)}{\partial x} - \int_0^T \frac{\partial f(s\epsilon^2,x_0)}{\partial x} w(s)\,ds\right),$$

$$\Delta(i) = \frac{1}{2}\int_0^T \delta(t;i)\,\lambda(t)^2\,dt.$$

For $w(t) = w_T(t)$, we obtain $\Delta(i) = 0$, $i = 0, 1$, and the theorem follows. □

Proposition 9.3.4 is proved by applying Theorem 9.A.1 to the equation (9.30). To compute $v(t)^2$, conditioning on $z(t)$ and using conditional independence of $X_0(t)$ and $z(t)$ we obtain,

$$\begin{aligned}\mathrm{E}\left((X_0(t)-x_0)^2 z(t)\right) &= \mathrm{E}\left(z(t)\mathrm{E}\left((X_0(t)-x_0)^2 \middle| z(\cdot)\right)\right) \\ &= \mathrm{E}\left(z(t)\int_0^t z(s)\lambda(s)^2\,ds\right) \\ &= \int_0^t \lambda(s)^2 \mathrm{E}(z(t)z(s))\,ds.\end{aligned} \tag{9.99}$$

Clearly

$$z(t) - z_0 = e^{-\theta(t-s)}(z(s) - z_0) + O(dW),$$

so that

$$\begin{aligned}\mathrm{E}(z(t)z(s)) &= z_0^2 + \mathrm{E}\left(e^{-\theta(t-s)}(z(s)-z_0)z(s)\right) \\ &= e^{-\theta(t-s)}\mathrm{E}\left(z(s)^2\right) + \left(1 - e^{-\theta(t-s)}\right)z_0^2.\end{aligned}$$

We also have that

$$\mathrm{E}\left(z(s)^2\right) = z_0^2 + z_0\eta^2\frac{1-e^{-2\theta s}}{2\theta}. \tag{9.100}$$

Substituting into (9.99) yields

$$v(t)^2 = \mathrm{E}\left((S(t)-x_0)^2 z(t)\right) \tag{9.101}$$

$$\begin{aligned}
&= \int_0^t \lambda(s)^2 \left(e^{-\theta(t-s)}\mathrm{E}\left(z(s)^2\right) + \left(1-e^{-\theta(t-s)}\right) z_0^2\right) ds \\
&= \int_0^t \lambda(s)^2 \Bigg(e^{-\theta(t-s)} z_0^2 + e^{-\theta(t-s)} z_0 \eta^2 \frac{1-e^{-2\theta s}}{2\theta} \\
&\quad + z_0^2\left(1-e^{-\theta(t-s)}\right)\Bigg) ds \qquad (9.102) \\
&= \int_0^t \lambda(s)^2 \left(z_0^2 + z_0\eta^2 e^{-\theta(t-s)} \frac{1-e^{-2\theta s}}{2\theta}\right) ds \\
&= z_0^2 \int_0^t \lambda(s)^2\, ds + z_0\eta^2 e^{-\theta t} \int_0^t \lambda(s)^2 \frac{e^{\theta s} - e^{-\theta s}}{2\theta}\, ds.
\end{aligned}$$

9.B Appendix: Coefficients for Asymptotic Expansion

Set $\Omega = \Omega_0 \overline{v}^{1/2}\tau^{1/2} + \Omega_1 \overline{v}^{3/2}\tau^{3/2}$ where $\overline{v} = \mu_{\overline{z\lambda^2}}(0, z_0)/T$. Also define the easily computed quantities

$$\Omega_{mn} = \frac{\partial^m \Omega / \partial \overline{v}^m}{\partial^n \Omega / \partial \overline{v}^n}.$$

Then the expansion coefficients in Proposition 9.2.4 are given by

$$\alpha_0 = \tau^{-2} l_{1,2}\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right), \quad \alpha_1 = \tau^{-2} l_{1,2}\Omega^{-2}\Omega_{10},$$

and

$$\begin{aligned}
\beta_0 =& \tau^{-2} l_{2,2}\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right) \\
&- \tau^{-3} l_{2,3}\left(\Omega_{31} - \Omega_{21}^2 - \frac{1}{4}\Omega^2\left(\Omega_{20} + \Omega_{10}^2\right) + \left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)^2\right) \\
&+ \frac{1}{2}\tau^{-4} l_{1,2}^2\left(\Omega_{41} - 3\Omega_{31}\Omega_{21} + 2\Omega_{21}^2 - \frac{1}{4}\Omega^2\Omega_{30} - \frac{3}{4}\Omega^2\Omega_{10}\Omega_{20}\right) \\
&+ \frac{3}{2}\tau^{-4} l_{1,2}^2\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)\left(\Omega_{31} - \Omega_{21}^2 - \frac{1}{4}\Omega^2\left(\Omega_{20} + \Omega_{10}^2\right)\right),
\end{aligned}$$

$$\begin{aligned}\beta_1 &= \Omega^{-2}\tau^{-2}l_{2,2}\Omega_{10} \\ &\quad - \Omega^{-2}\tau^{-3}l_{2,3}\left(\Omega_{20} - 3\Omega_{10}^2 + 2\Omega_{10}\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)^2\right) \\ &\quad + \Omega^{-2}\frac{1}{2}\tau^{-4}l_{1,2}^2\Big(\Omega_{30} - 9\Omega_{10}\Omega_{20} + 12\Omega_{10}^3 \\ &\qquad + 3\Omega_{10}\left(\Omega_{31} - \Omega_{21}^2 - \frac{1}{4}\Omega^2\left(\Omega_{20} + \Omega_{10}^2\right)\right)\Big) \\ &\quad + \Omega^{-2}\frac{3}{2}\tau^{-4}l_{1,2}^2\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)\left(\Omega_{20} - 3\Omega_{10}^2\right),\end{aligned}$$

$$\beta_2 = -\Omega^{-4}\tau^{-3}l_{2,3}\Omega_{10}^2 + \Omega^{-4}\frac{3}{2}\tau^{-4}l_{1,2}^2\Omega_{10}\left(\Omega_{20} - 3\Omega_{10}^2\right).$$