1

Introduction to Arbitrage Pricing Theory

For reference, this chapter reviews selected results from stochastic calculus and from the modern theory of asset pricing. The material in this chapter is well covered in existing literature, so we keep the chapter brief and the mathematical treatment informal. For a more rigorous treatment we refer to Duffie [2001] or Musiela and Rutkowski [1997]. Most of the necessary mathematical foundation for the theory is available in Karatzas and Shreve [1997], Øksendal [1992], and Protter [2005].

The treatment in this chapter focuses on asset pricing in general; we shall specialize it to interest rate securities in Chapter 4. Chapter 5 introduces fixed income markets in detail.

1.1 The Setup

Unless otherwise noted, in this book we shall always consider an economy with continuous and frictionless trading taking place inside a finite horizon $[0, T]$. We assume the existence of traded dividend-free assets with prices characterized by a p-dimensional vector-valued stochastic process $X(t) = (X_1(t), \ldots X_p(t))^\top$. Uncertainty and information arrival is modeled by a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, with Ω being a sample space with outcome elements ω; $\mathcal{F}$ being a σ-algebra on Ω; and P being a probability measure on the measure space $(\Omega, \mathcal{F})$. Information is revealed over time according to a filtration $\{\mathcal{F}_t,\ t \in [0, T]\}$, a family of sub-$\sigma$-algebras of $\mathcal{F}$ satisfying $\mathcal{F}_s \subseteq \mathcal{F}_t$ whenever $s \leq t$. We can loosely think of $\mathcal{F}_t$ as the information available at time t. We assume that the process $X(t)$ is adapted to $\{\mathcal{F}_t\}$, i.e. that $X(t)$ is fully observable at time t. For technical reasons, we require that the filtration satisfies the "usual conditions"[1]. Let $\mathrm{E}^{\mathrm{P}}(\cdot)$ be the expectation

[1] To satisfy the "usual conditions", $\mathcal{F}_t$ must be right-continuous for all t, and $\mathcal{F}_0$ must contain all the null-sets of $\mathcal{F}$, i.e. all subsets of sets of zero P-probability.

operator for the measure P; when conditioning on information at time t, we will use the notation $\mathrm{E}_t^{\mathrm{P}}(\cdot) = \mathrm{E}^{\mathrm{P}}(\cdot|\mathcal{F}_t)$.

In all of the models in this book, we specialize the abstract setup above to the situation where information is generated by a d-dimensional vector-valued *Brownian motion* (or *Wiener process*) $W(t) = (W_1(t), \ldots, W_d(t))^\top$, where W_i is independent of W_j for $i \neq j$. Brownian motions are treated in detail in Karatzas and Shreve [1997]; here, we just recall that a scalar Brownian motion W_i is a continuous stochastic process starting at 0 (i.e. $W_i(0) = 0$), having independent Gaussian increments: $W_i(t) - W_i(s) \sim \mathcal{N}(0, t-s)$, $t \geq s$. The filtration we consider is normally always the one *generated* by W, $\mathcal{F}_t = \sigma\{W(u), 0 \leq u \leq t\}$, possibly augmented to satisfy the usual conditions. We will generally assume that the price vector $X(t)$ is described by a vector-valued *Ito process*:

$$X(t) = X(0) + \int_0^t \mu(s,\omega)\, ds + \int_0^t \sigma(s,\omega)\, dW(s), \tag{1.1}$$

or, in differential notation,

$$dX(t) = \mu(t,\omega)\, dt + \sigma(t,\omega)\, dW(t), \tag{1.2}$$

where $\mu : \mathbb{R} \times \Omega \to \mathbb{R}^p$ and $\sigma : \mathbb{R} \times \Omega \to \mathbb{R}^{p \times d}$ are processes of dimension p and $p \times d$, respectively. We assume that both μ and σ are adapted to $\{\mathcal{F}_t\}$ and are in L^1 and L^2 respectively, in the sense that for all $t \in [0, T]$,

$$\int_0^t |\mu(s,\omega)|\, ds < \infty, \tag{1.3}$$

$$\int_0^t |\sigma(s,\omega)|^2\, ds < \infty, \tag{1.4}$$

almost surely[2]. In (1.4), we have defined

$$|\sigma(t,\omega)|^2 = \mathrm{tr}\left(\sigma(t,\omega)\sigma(t,\omega)^\top\right). \tag{1.5}$$

We notice that the sample paths of X generated by (1.1) are almost surely continuous, with no jumps in asset prices.

A technical treatment of Ito processes and the Ito integral with respect to Brownian motion can be found in Karatzas and Shreve [1997]. For our needs, it suffices to think of the Ito integral as

$$\int_0^t \sigma(s,\omega)\, dW(s) = \lim_{n\to\infty} \sum_{i=1}^{n} \sigma\left((i-1)\delta, \omega\right) \left[W(i\delta) - W((i-1)\delta)\right], \tag{1.6}$$

[2] An event holds "almost surely" — often abbreviated by "a.s." — if the probability of the event is one.

where $\delta \triangleq t/n$. We note that the integrand σ is here always evaluated at the *left* of each interval $[(i-1)\delta, i\delta]$. Other choices are possible[3], but, as we shall see, the "non-anticipative" structure of the Ito integral gives rise to a number of useful results and makes it particularly useful as a model of trading gains (see Section 1.2).

We list a few relevant definitions and results below.

Definition 1.1.1 (Martingale). *Let $Y(t)$ be an adapted vector-valued process with $\mathrm{E}^{\mathrm{P}}(|Y(t)|) < \infty$ for all $t \in [0,T]$. We say that $Y(t)$ is a* martingale *under measure* P *if for all $s, t \in [0,T]$ with $t \leq s$,*

$$\mathrm{E}_t^{\mathrm{P}}(Y(s)) = Y(t), \quad a.s.$$

If we replace the equality sign in this equation with $\leq$ or $\geq$, $Y(t)$ is said to be a *supermartingale* or a *submartingale*, respectively.

Definition 1.1.2 (Space H^2). *Let $|\sigma(t,\omega)|^2$ be as defined in* (1.5). *We say that σ is in H^2, if for all $t \in [0,T]$ we have*

$$\mathrm{E}^{\mathrm{P}}\left(\int_0^t |\sigma(s,\omega)|^2\, ds\right) < \infty.$$

The importance of Definition 1.1.2 becomes clear from the following result:

Theorem 1.1.3 (Properties of Ito Integral). *Define $I(t) = \int_0^t \sigma(s,\omega)\, dW(s)$ and assume that σ is in H^2. Then*

1. *$I(t)$ is $\mathcal{F}_t$-measurable.*
2. *$I(t)$ is a continuous martingale. In particular, $\mathrm{E}^{\mathrm{P}}(I(t)) = 0$ for all $t \in [0,T]$.*
3. *$\mathrm{E}^{\mathrm{P}}(|I(t)|^2) = \mathrm{E}^{\mathrm{P}}(\int_0^t |\sigma(s,\omega)|^2\, ds) < \infty$.*
4. *$\mathrm{E}^{\mathrm{P}}(I(t)I(s)^\top) = \mathrm{E}^{\mathrm{P}}(\int_0^{\min(t,s)} \sigma(u,\omega)\sigma(u,\omega)^\top\, du)$.*

A proof of Theorem 1.1.3 can be found in, e.g., Karatzas and Shreve [1997]. The equality in the third item of Theorem 1.1.3 is known as the *Ito isometry*. Due to the inequality in the third item, we say that the martingale defined in the process is a *square-integrable martingale*.

While it is common in applied work to simply assume that Ito integrals are martingales, without technical regularity conditions on $\sigma(t,\omega)$ (such as the H^2 restriction in Theorem 1.1.3), we should note that Ito integrals involving general processes in L^2 can, in fact, only be guaranteed to be *local martingales*. A process X is said to be a local martingale if there exists a

[3]The *Stratonovich* stochastic integral evaluates σ at the mid-point of each interval.

sequence of stopping times[4] $\{\tau_n\}_{n=1}^{\infty}$, with $\tau_n \to \infty$ as $n \to \infty$, such that $X(\min(t, \tau_n))$, $t \geq 0$, is a martingale for all n. In other words, all "driftless" Ito processes of the type

$$dY(t) = \sigma(t, \omega)\, dW(t) \tag{1.7}$$

are local martingales, but not necessarily martingales. Interestingly, a converse result holds as well; all local martingales adapted to the filtration generated by the Brownian motion W can be represented as Ito processes of the form (1.7):

Theorem 1.1.4 (Martingale Representation Theorem). *If Y is a local martingale adapted to the filtration generated by a Brownian motion W, then there exists a process σ such that* (1.7) *holds. If Y is a square-integrable martingale, then σ is in H^2.*

The proof of Theorem 1.1.4 can be found in Karatzas and Shreve [1997].

In the manipulation of functionals of Ito processes, the key result is a famous result by K. Ito:

Theorem 1.1.5 (Ito's Lemma). *Let $f(t, x)$, $x = (x_1, \ldots, x_p)^\top$, denote a continuous function, $f : [0, T] \times \mathbb{R}^p \to \mathbb{R}$, with continuous partial derivatives $\partial f/\partial t = f_t$, $\partial f/\partial x_i = f_{x_i}$, $\partial^2 f/\partial x_i \partial x_j = f_{x_i x_j}$. Let $X(t)$ be given by the Ito process* (1.2) *and define a scalar process $Y(t) = f(t, X(t))$. Then $Y(t)$ is an Ito process with stochastic differential*

$$\begin{aligned} dY(t) = {} & f_t\left(t, X(t)\right) dt + f_x\left(t, X(t)\right) \mu(t, \omega)\, dt + f_x\left(t, X(t)\right) \sigma(t, \omega)\, dW(t) \\ & + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} f_{x_i x_j}\left(t, X(t)\right) \left(\sigma(t, \omega)\sigma(t, \omega)^\top\right)_{i,j} dt, \end{aligned}$$

where $f_x = (f_{x_1}, \ldots, f_{x_p})$.

For easy reference, the result below lists Ito's lemma for the special case where $p = d = 1$.

Corollary 1.1.6. *For the case $p = d = 1$, Ito's lemma becomes*

$$\begin{aligned} dY(t) = {} & \left(f_t\left(t, X(t)\right) + f_x\left(t, X(t)\right) \mu(t, \omega) + \frac{1}{2} f_{xx}\left(t, X(t)\right) \sigma(t, \omega)^2 \right) dt \\ & + f_x\left(t, X(t)\right) \sigma(t, \omega)\, dW(t). \end{aligned}$$

Ito's lemma can be motivated heuristically from a Taylor expansion. For instance, for the scalar case in Corollary 1.1.6, we write informally

[4] Recall that a stopping time τ is simply a random time adapted to the given filtration, in the sense that the event $\{\tau \leq t\}$ belongs to $\mathcal{F}_t$.

$$f\left(t+dt, X(t+dt)\right) = f\left(t, X(t)\right) + f_t\, dt + f_x\, dX(t) + \frac{1}{2} f_{xx} \left(dX(t)\right)^2 + \ldots. \tag{1.8}$$

Here, we have

$$\left(dX(t)\right)^2 = \mu(t,\omega)^2 \left(dt\right)^2 + \sigma(t,\omega)^2 \left(dW(t)\right)^2 + 2\mu(t,\omega)\sigma(t,\omega)\, dt\, dW(t).$$

As shown earlier, $(dW(t))^2 = dt$ in quadratic mean, whereas all other terms in the expression for $(dX(t))^2$ are of order $O(dt^{3/2})$ or higher and can be neglected for small dt. In the limit, we therefore have $(dX(t))^2 = \sigma(t,\omega)^2\, dt$ which can be inserted into (1.8). The result in Corollary 1.1.6 then emerges.

Remark 1.1.7. The quantity $(dX(t))^2$ discussed above is the differential of the *quadratic variation* of $X(t)$, often denoted by $\langle X(t), X(t)\rangle$. That is,

$$d\left\langle X(t), X(t)\right\rangle = \left(dX(t)\right)^2 \quad \Rightarrow \quad \left\langle X(t), X(t)\right\rangle = \int_0^t \left(dX(u)\right)^2.$$

For two different (scalar) Ito processes $X(t)$ and $Y(t)$, we may equivalently define the *quadratic covariation* process $\langle X(t), Y(t)\rangle$ by

$$d\left\langle X(t), Y(t)\right\rangle = dX(t)\, dY(t).$$

Sometimes we also write $d\langle X(t), Y(t)\rangle = \langle dX(t), dY(t)\rangle$. If $X(t)$ is a p-dimensional process and $Y(t)$ is a q-dimensional process, the quadratic covariation $\langle X(t), Y(t)^\top\rangle$ is a $(p \times q)$-dimensional matrix process whose (i,j)-th element is $\langle X_i(t), Y_j(t)\rangle$, $i = 1, \ldots, p$, $j = 1, \ldots, q$.

The so-called *Tanaka extension* (see Karatzas and Shreve [1997]) extends Ito's lemma to continuous but non-differentiable functions. At points where the function has a kink, the Tanaka extension (loosely speaking) justifies using the Heaviside (step-) function for the first-order derivative and the Dirac delta function for the second-order derivative. An application of the Tanaka extension can be found in Section 1.9.2 and in Chapter 7, along with further discussion and references.

1.2 Trading Gains and Arbitrage

Working in the setting of Section 1.1 with assets driven by Ito processes, we now consider an investor engaging in a trading strategy involving the p assets $X_1, \ldots, X_p$. Let the trading strategy be characterized by a predictable[5] adapted process $\phi(t,\omega) = (\phi_1(t,\omega), \ldots, \phi_p(t,\omega))^\top$, with $\phi_i(t,\omega)$ denoting

[5] A *predictable* process is one where we, loosely speaking, can "foretell" the value of the process at time t, given all information available up to, but not including, time t. All adapted continuous processes are thus predictable. For a technical definition of predictable processes, see Karatzas and Shreve [1997].

the holdings at time t in the i-th asset X_i. The value $\pi(t)$ of the trading strategy at time t is thus (dropping the dependence on ω in the notation)

$$\pi(t) = \phi(t)^\top X(t). \tag{1.9}$$

The gain from trading over a small time interval $[t, t+\delta]$ is (approximately) $\phi(t)^\top [X(t+\delta) - X(t)]$, suggesting (compare to (1.6)) that the Ito integral

$$\int_0^t \phi(s)^\top \, dX(s) = \int_0^t \phi(s)^\top \mu(s) \, ds + \int_0^t \phi(s)^\top \sigma(s) \, dW(s)$$

is a proper model for trading gains over $[0, t]$. An investment strategy is said to be *self-financing* if, for any $t \in [0, T]$,

$$\pi(t) - \pi(0) = \int_0^t \phi(s)^\top dX(s). \tag{1.10}$$

This relationship simply expresses that changes in portfolio value are solely caused by trading gains or losses, with no funds being added or withdrawn.

Self-financing trading strategies allow investors to turn a certain initial investment $\pi(0)$ into stochastic future wealth $\pi(t)$. Under natural assumptions on possible trading strategies (e.g., that there is finite supply of all assets) we would expect that there should be limitations to the profits that self-financing strategies can create. Most notably, it should be impossible to create "something for nothing", that is, to turn a zero initial investment into future wealth that is certain to be non-negative and may be positive with non-zero probability. To express this formally, we introduce the concept of an *arbitrage opportunity*:

Definition 1.2.1 (Arbitrage). *An arbitrage opportunity is a self-financing strategy ϕ for which $\pi(0) = 0$ and, for some $t \in [0, T]$,*

$$\pi(t) \geq 0 \text{ a.s., and } \mathrm{P}\left(\pi(t) > 0\right) > 0, \tag{1.11}$$

with π given in (1.9).

In economic equilibrium, arbitrage strategies cannot exist and precluding (1.11) constitutes a fundamental consistency requirement on the asset processes.

1.3 Equivalent Martingale Measures and Arbitrage

We turn to the question of characterizing the conditions under which the trading economy is free of arbitrage opportunities. A concise way to state these conditions involves *equivalent martingale measures*, a concept we shall work our way up to in a number of steps. First, we recall that two

probability measures P and $\widehat{\mathrm{P}}$ on the same measure space $(\Omega, \mathcal{F})$ are said to be *equivalent* if $\mathrm{P}(A) = 0 \Leftrightarrow \widehat{\mathrm{P}}(A) = 0$, $\forall A \in \mathcal{F}$; that is, the two measures have the same null-sets. An important result from measure theory states that equivalent measures are uniquely associated through a quantity known as a *Radon-Nikodym derivative*:

Theorem 1.3.1 (Radon-Nikodym Theorem). *Let* P *and* $\widehat{\mathrm{P}}$ *be equivalent probability measures on the common measure space* $(\Omega, \mathcal{F})$. *There exists a unique (a.s.) non-negative random variable* R *with* $\mathrm{E}^{\mathrm{P}}(R) = 1$, *such that*

$$\widehat{\mathrm{P}}(A) = \mathrm{E}^{\mathrm{P}}(R 1_{\{A\}}), \quad \textit{for all } A \in \mathcal{F}.$$

For a proof of Theorem 1.3.1, see e.g. Billingsley [1995]. The random variable R in the theorem is known as a *Radon-Nikodym derivative* and is denoted $d\widehat{\mathrm{P}}/d\mathrm{P}$. In the theorem we have used an *indicator* $1_{\{A\}}$; this quantity is 1 if the event A comes true, 0 if not.

For later use, we associate any probability measure $\widehat{\mathrm{P}}$ with a *density process*

$$\varsigma(t) = \mathrm{E}_t^{\mathrm{P}}\left(\frac{d\widehat{\mathrm{P}}}{d\mathrm{P}}\right), \quad \forall t \in [0, T]. \tag{1.12}$$

Clearly, $\varsigma(t)$ is a P-martingale with $\varsigma(0) = 1$ and $\varsigma(t) = \mathrm{E}_t^{\mathrm{P}}(\varsigma(T))$. A simple conditioning exercise demonstrates that for any $\mathcal{F}_T$-measurable random variable $Y(T)$, with $R = d\widehat{\mathrm{P}}/d\mathrm{P}$,

$$\begin{aligned}
\mathrm{E}^{\widehat{\mathrm{P}}}(Y(T)|\mathcal{F}_t) &= \frac{1}{\mathrm{E}^{\mathrm{P}}(R|\mathcal{F}_t)} \mathrm{E}^{\mathrm{P}}(RY(T)|\mathcal{F}_t) \\
&= \varsigma(t)^{-1} \mathrm{E}^{\mathrm{P}}\left(\mathrm{E}^{\mathrm{P}}(R|\mathcal{F}_T) Y(T)|\mathcal{F}_t\right) \\
&= \mathrm{E}^{\mathrm{P}}\left(Y(T)\frac{\varsigma(T)}{\varsigma(t)}\middle|\mathcal{F}_t\right).
\end{aligned} \tag{1.13}$$

We shall use this result on numerous occasions in this book.

We now introduce the important concept of a *deflator*, a strictly positive Ito process used to normalize the asset prices. Let the deflator be denoted $D(t)$, and define the normalized asset process $X^D(t) = (X_1(t)/D(t), \ldots, X_p(t)/D(t))^\top$. We say that a measure Q^D is an *equivalent martingale measure induced by* D if $X^D(t)$ is a martingale with respect to Q^D. If Q^D is a martingale measure, we say that a self-financing trading strategy is *permissible* if

$$\int_0^t \phi(s)^\top dX^D(s)$$

is a martingale. For the Ito setup discussed earlier, a permissible strategy[6] is obtained by, say, requiring that $\phi(t)^\top \sigma(t)$ is in H^2; see Theorem 1.1.3. An

[6]The technical restriction on trading positions imposed by only considering permissible trading strategies rules out certain pathological strategies, such as the

application of Ito's lemma combined with (1.9)–(1.10) implies that $\pi(t)/D(t)$ is a Q^D-martingale when the trading strategy is permissible.

For permissible trading strategies, the importance of equivalent martingale measures follows from the following theorem:

Theorem 1.3.2 (Sufficient Condition for No-Arbitrage). *Restrict attention to permissible trading strategies. If there is a deflator D such that the deflated asset price process allows for an equivalent martingale measure, then there is no arbitrage.*

For a proof we refer to Musiela and Rutkowski [1997]. We note that Theorem 1.3.2 only provides sufficient conditions for the absence of arbitrage, and known (and rather technical) counterexamples demonstrate that the existence of an equivalent martingale measure does not follow from the absence of arbitrage in a setting with permissible trading strategies. A body of results known as the *fundamental theorem of arbitrage* establishes the conditions under which the existence of an equivalent martingale measure is also a necessary condition for the absence of arbitrage. The results are rather technical, but generally state that absence of arbitrage and the existence of an equivalent martingale measure are "nearly" equivalent concepts. The exact notion of "nearly" equivalent is discussed in Duffie [2001] as well as in the authoritative reference[7] Delbaen and Schachermayer [1994]. For our purposes in this book, we ignore many of these technicalities and often simply treat the absence of arbitrage and the existence of a martingale measure as equivalent concepts.

Finally, if the deflator is one of the p assets, we call the deflator a *numeraire*. Let us, say, assume that X_1 is strictly positive and can be used as a numeraire. Also assume that a deflator D has been identified such that Theorem 1.3.2 holds. As $X_1(t)/D(t)$ is a Q^D-martingale, we can use the Radon-Nikodym theorem to define a new measure Q^{X_1} by the density $\varsigma(t) = (X_1(t)/D(t))/(X_1(0)/D(0))$. For an $\mathcal{F}_T$-measurable variable $Y(T)$, we then have, from (1.13),

$$X_1(t)\mathrm{E}_t^{\mathrm{Q}^{X_1}}\left(\frac{Y(T)}{X_1(T)}\right) = D(t)\mathrm{E}_t^{\mathrm{Q}^D}\left(\frac{Y(T)}{D(T)}\right). \tag{1.14}$$

In particular, if $Y(t)/D(t)$ is a Q^D-martingale, $Y(t)/X_1(t)$ must also be a Q^{X_1}-martingale. In practice, it normally suffices to only consider deflators from the set of available numeraires.

Remark 1.3.3. Some sources define $1/D(t)$ (rather than $D(t)$) as the deflator. The convention used in this book is more natural for our applications.

---

doubling strategy considered in Harrison and Kreps [1979]. A realistic resource-constrained economy will always bound the size of the positions one can take in an asset, sufficing to ensure that predictable trading strategies are permissible.

[7] In a nutshell, Delbaen and Schachermayer [1994] show that absence of arbitrage implies only the existence of a *local* martingale measure.

1.4 Derivative Security Pricing and Complete Markets

A T-maturity *derivative security* (also known as a *contingent claim*) pays out at time T an $\mathcal{F}_T$-measurable random variable $V(T)$, and makes no payments before T. We assume that $V(T)$ has finite variance, and say that the derivative security is *attainable* (or sometimes *redundant*) if there exists a permissible trading strategy ϕ such that $V(T) = \phi(T)^\top X(T) = \pi(T)$ a.s. The trading strategy is said to *replicate* the derivative security. Importantly, the absence of arbitrage dictates that the time 0 price of an attainable derivative security $V(0)$ must be equal to the cost of setting up the self-financing strategy, i.e. $V(0) = \pi(0)$. More generally, $V(t) = \pi(t)$, $t \in [0, T]$. This observation is the foundation of *arbitrage pricing* and allows us to price derivative securities as expectations under an equivalent martingale measure. Specifically, consider a deflator D and assume the existence of an equivalent martingale measure Q^D induced by D; the existence of Q^D guarantees that there are no arbitrages in the market, by Theorem 1.3.2. Now, from the martingale property of $\pi(t)/D(t)$ in the measure Q^D and the relation $V(t) = \pi(t)$ it immediately follows that

$$\frac{V(t)}{D(t)} = \mathrm{E}_t^{\mathrm{Q}^D}\left(\frac{V(T)}{D(T)}\right)$$

or

$$V(t) = D(t)\mathrm{E}_t^{\mathrm{Q}^D}\left(\frac{V(T)}{D(T)}\right). \tag{1.15}$$

If all finite-variance $\mathcal{F}_T$-measurable random variables can be replicated, the market is said to be *complete*. In a complete market, all derivatives are "spanned" and hence have unique prices. Interestingly, a similar uniqueness result holds for equivalent martingale measures:

Theorem 1.4.1. *In the absence of arbitrage, a market is complete if and only if there exists a deflator inducing a* unique *martingale measure.*

From (1.14) it follows that the martingale measures induced by all numeraires must then be unique as well.

In practical applications, we shall often manipulate the choice of numeraire asset to simplify computations. The following result is useful for this:

Theorem 1.4.2 (Change of Numeraire). *Consider two numeraires $N(t)$ and $M(t)$, inducing equivalent martingale measures Q^N and Q^M, respectively. If the market is complete, then the density of the Radon-Nikodym derivative relating the two measures is uniquely given by*

$$\varsigma(t) = \mathrm{E}_t^{\mathrm{Q}^N}\left(\frac{d\mathrm{Q}^M}{d\mathrm{Q}^N}\right) = \frac{M(t)/M(0)}{N(t)/N(0)}.$$

Proof. As the market is complete, all derivatives prices are unique. Consider an integrable $\mathcal{F}_T$-measurable payout $V(T) = Y(T)M(T)$, with time t price $V(t)$. From Theorem 1.4.1 and (1.15) we must have

$$V(t) = N(t)\mathrm{E}_t^{\mathrm{Q}^N}\left(\frac{M(T)Y(T)}{N(T)}\right) = M(t)\mathrm{E}_t^{\mathrm{Q}^M}\left(\frac{M(T)Y(T)}{M(T)}\right)$$

or

$$\mathrm{E}_t^{\mathrm{Q}^M}(Y(T)) = \mathrm{E}_t^{\mathrm{Q}^N}\left(Y(T)\frac{M(T)/N(T)}{M(t)/N(t)}\right).$$

Comparison with (1.13), and the fact that the density must be scaled to equal 1 at time 0, reveals that the Radon-Nikodym derivative for the measure shift is characterized by the density in the theorem. □

1.5 Girsanov's Theorem

The last two sections have demonstrated a close link between the concept of arbitrage and the existence and uniqueness of equivalent martingale measures. In this section, we consider i) the conditions on the asset prices that allow for an equivalent martingale measure; and ii) the effect on asset dynamics from a change of probability measure. We consider two measures P and $\mathrm{P}(\theta)$ related by a density $\varsigma^\theta(t) = \mathrm{E}_t^{\mathrm{P}}(d\mathrm{P}(\theta)/d\mathrm{P})$, where $\varsigma^\theta(t)$ is an *exponential martingale* given by the Ito process

$$d\varsigma^\theta(t)/\varsigma^\theta(t) = -\theta(t)^\top dW(t),$$

where $W(t)$ is a d-dimensional P-Brownian motion. The d-dimensional process θ is known as the *market price of risk*. By an application of Ito's lemma, we can write

$$\begin{aligned}\varsigma^\theta(t) &= \exp\left(-\int_0^t \theta(s)^\top dW(s) - \frac{1}{2}\int_0^t \theta(s)^\top\theta(s)\,ds\right)\\ &\triangleq \mathcal{E}\left(-\int_0^t \theta(s)^\top dW(s)\right)\end{aligned} \tag{1.16}$$

where $\mathcal{E}(\cdot)$ is the *Doleans exponential*. An often-quoted sufficient condition on $\theta(t)$ for (1.16) to define a proper martingale (and not just a local martingale) is the *Novikov condition*

$$\mathrm{E}^{\mathrm{P}}\left[\exp\left(\frac{1}{2}\int_0^t \theta(s)^\top\theta(s)\,ds\right)\right] < \infty. \tag{1.17}$$

The Novikov condition can often be difficult to verify in practical applications.

Armed with the notation above, we are now ready to state the main result of this section.

Theorem 1.5.1 (Girsanov's Theorem). *Suppose that $\varsigma^\theta(t)$ defined in* (1.16) *is a martingale. Then for all $t \in [0, T]$*

$$W^\theta(t) = W(t) + \int_0^t \theta(s)\, ds$$

is a Brownian motion under the measure $\mathrm{P}(\theta)$.

To discuss a strategy to prove Girsanov's theorem, assume for simplicity that the dimension of the Brownian motion is $d = 1$. One way to construct a proof for Theorem 1.5.1 is to demonstrate that the joint moment-generating function (mgf)[8] (under $\mathrm{P}(\theta)$) of the increments

$$W^\theta(t_1), W^\theta(t_2) - W^\theta(t_1), \ldots, W^\theta(t_n) - W^\theta(t_{n-1}), \quad 0 < t_1 < \ldots < t_n,$$

is the same as that of n independent Gaussian random variables with expectations 0 and variances $t_1, t_2 - t_1, \ldots$. That is, for any positive integer value of n and any set of values $\alpha_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, we need to show that,

$$\mathrm{E}^{\mathrm{P}^\theta}\left[\exp\left(\sum_{i=1}^n \alpha_i \left(W^\theta(t_i) - W^\theta(t_{i-1})\right)\right)\right] = \prod_{i=1}^n \exp\left(\alpha_i^2 \left(t_i - t_{i-1}\right)/2\right),$$

where we have defined $t_0 = 0$. While carrying out such a proof is not difficult, we here merely justify the final result by examining the case $n = 1$ only. Specifically, we consider

$$\mathrm{E}^{\mathrm{P}(\theta)}\left[\exp\left(\alpha W^\theta(t)\right)\right],$$

where $\alpha \in \mathbb{R}$ and $t > 0$. Shifting probability measure, we get

$$\begin{aligned}
\mathrm{E}^{\mathrm{P}(\theta)}\left[\exp\left(\alpha W^\theta(t)\right)\right] &= \mathrm{E}^{\mathrm{P}(\theta)}\left[\exp\left(\alpha W(t) + \alpha \int_0^t \theta(s)\, ds\right)\right] \\
&= \mathrm{E}^{\mathrm{P}}\left[\exp\left(\alpha W(t) + \alpha \int_0^t \theta(s)\, ds\right) \mathcal{E}\left(-\int_0^t \theta(s)\, dW(s)\right)\right] \\
&= e^{\alpha^2 t/2} \mathrm{E}^{\mathrm{P}}\left[\exp\left(\int_0^t (\alpha - \theta(s))\, dW(s) - \frac{1}{2}\int_0^t (\alpha - \theta(s))^2\, ds\right)\right] \\
&= e^{\alpha^2 t/2} \mathrm{E}^{\mathrm{P}}\left[\mathcal{E}\left(\int_0^t (\alpha - \theta(s))\, dW(s)\right)\right] \\
&= e^{\alpha^2 t/2},
\end{aligned}$$

[8] Recall that the moment-generating function of a random variable Y in some measure P is defined as the expectation $\mathrm{E}^{\mathrm{P}}(\exp(\alpha Y))$, $\alpha \in \mathbb{R}$. Unlike the characteristic function, the moment-generating function is not always well-defined for all values of the argument α.

as desired. In the last step, we used the fact that the Doleans exponential is a martingale with initial value 1.

Girsanov's theorem implies that we can shift probability measure to transform an Ito process with a given drift to an Ito process with nearly arbitrary drift. Specifically, we notice that our asset price process (under P)

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

can be written

$$dX(t) = (\mu(t) - \sigma(t)\theta(t))\,dt + \sigma(t)dW^{\theta}(t),$$

where $W^{\theta}(t)$ is a Brownian measure under the measure $\mathrm{P}(\theta)$. This process will be driftless provided that θ satisfies the "spanning condition" $\mu(t) = \sigma(t)\theta(t)$ for all $t \in [0, T]$. This gives us a convenient way to check for the existence of equivalent martingale measures:

Corollary 1.5.2. *For a given numeraire D, assume that the deflated asset process satisfies*

$$dX^D(t) = \mu^D(t)\,dt + \sigma^D(t)\,dW(t),$$

where $\sigma^D(t)$ is sufficiently regular to make $\int_0^t \sigma^D(s)\,dW(s)$ a martingale. Assume also that there exists a θ such that the density ς^{θ} is a martingale and (a.s.)

$$\sigma^D(t)\theta(t) = \mu^D(t), \quad t \in [0, T], \tag{1.18}$$

then D induces an equivalent martingale measure and there is no arbitrage.

Equation (1.18) is a system of linear equations and we can use rank results from linear algebra to determine the circumstances under which (1.18) will have solutions (no arbitrage) and when these are unique (complete market). For instance, a necessary condition for the market to be complete is that $\mathrm{rank}(\sigma) = d$. Further results along these lines can be found in Musiela and Rutkowski [1997] and Duffie [2001].

We conclude this section by noting that while a change of probability measure affects the drift μ of an Ito process, it does not change the diffusion coefficient σ. This is sometimes known as the *diffusion invariance principle.*

1.6 Stochastic Differential Equations

So far we have defined the asset process vector to be an Ito process with general measurable coefficients $\mu(t, \omega)$ and $\sigma(t, \omega)$. In virtually all applications, however, we restrict our attention to the case where these coefficients

are deterministic functions of time and the state of the asset process[9]. In other words, we consider a *stochastic differential equation* (SDE) of the form

$$dX(t) = \mu(t, X(t))\, dt + \sigma(t, X(t))\, dW(t), \quad X(0) = X_0, \tag{1.19}$$

with $\mu : [0,T] \times \mathbb{R}^p \to \mathbb{R}^p$; $\sigma : [0,T] \times \mathbb{R}^p \to \mathbb{R}^{p \times d}$; and X_0 an initial condition. A *strong solution*[10] to (1.19) is an Ito process

$$X(t) = X_0 + \int_0^t \mu(s, X(s))\, ds + \int_0^t \sigma(s, X(s))\, dW(s).$$

A number of restrictions on μ and σ are needed to ensure that the solution to (1.19) exists and is unique. A standard result is listed below.

Theorem 1.6.1. *In* (1.19) *assume that there exists a constant K such that for all $t \in [0,T]$ and all $x, y \in \mathbb{R}^p$,*

$$\begin{aligned} &|\mu(t,x) - \mu(t,y)| + |\sigma(t,x) - \sigma(t,y)| \le K|x-y|, && \text{(Lipschitz condition)},\\ &|\mu(t,x)|^2 + |\sigma(t,x)|^2 \le K^2\left(1 + |x|^2\right), && \text{(growth condition)}. \end{aligned}$$

Then there exists a unique solution to (1.19).

We notice that the dynamics of (1.19) do not depend on the past evolution of $X(t)$ beyond the state of X at time t. This lack of path-dependence suggests that X is a *Markov* process. We formalize this as follows.

Definition 1.6.2 (Markov Process). *The $\mathbb{R}^p$-valued stochastic process $X(t)$ is called a Markov process if for all $s, t \in [0,T]$ with $t \le s$,*

$$\mathrm{P}\left(X(s) \in B | \mathcal{F}_t\right) = \mathrm{P}\left(X(s) \in B | X(t)\right) \tag{1.20}$$

for all sets B in the p-dimensional σ-algebra of Borel set $\mathfrak{B}^p$. If (1.20) *holds with s replaced by a stopping time, the process is a* strong *Markov process.*

Expressed verbally, the Markov property implies that the past and future become statistically independent when we condition on the present.

Theorem 1.6.3 (Markov Property of SDEs). *Let the coefficient of the SDE for $X(t)$ satisfy the conditions in Theorem 1.6.1. Then $X(t)$ is a strong Markov process.*

---

[9]In this section, the process X is generic and need not represent financial assets.

[10]In a strong solution, the Brownian motion is given and the solution is adapted to the filtration generated by it. If we are free to pick our own Brownian motion on some different probability space, we say that (1.19) holds in a *weak sense*. For financial applications where we normally only need the law of the underlying process, weak solutions are typically sufficient. The distinction between weak and strong solutions is of little importance for our purposes and we shall ignore it going forward.

Let us consider the explicit solutions of a few simple SDEs. First, consider a *linear* SDE

$$dX(t) = (AX(t) + B(t))\, dt + C(t)\, dW(t),$$

where A is a constant $p \times p$ matrix, and B and C are deterministic matrices of dimension $p \times 1$ and $p \times d$, respectively. The solution to this equation can, by Ito's lemma, be verified to be

$$X(t) = e^{At}X(0) + \int_0^t e^{A(t-s)}\,(B(s)\, ds + C(s)\, dW(s)).$$

The term

$$\int_0^t e^{A(t-s)}C(s)\, dW(s)$$

is distributed as a p-dimensional Gaussian random variable with mean 0 and, from Theorem 1.1.3, covariance matrix

$$\Sigma = \int_0^t e^{A(t-s)}C(s)C(s)^\top e^{A^\top(t-s)}\, ds.$$

Extensions to time-varying A are straightforward, and basically involve replacing the exponential matrix e^{At} with the solution of a homogeneous ODE with time-dependent coefficients. Details can be found in, e.g., Arnold [1974] and key results are listed in Chapter 12.

Now let us specialize to the scalar case with $p = 1$. An SDE of great importance is the *geometric Brownian motion with drift* (GBMD):

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

where $\mu(t)$ and $\sigma(t)$ are *deterministic* (with $\sigma(t)$ having dimension $1 \times d$). An application of Ito's lemma to $\ln(X(t))$ reveals that

$$\begin{aligned} X(t) &= X(0)\exp\left(\int_0^t \left(\mu(s) - \frac{1}{2}\sigma(s)\sigma(s)^\top\right) ds + \int_0^t \sigma(s)\, dW(s)\right) \\ &= X(0)\exp\left(\int_0^t \mu(s)\, ds\right)\mathcal{E}\left(\int_0^t \sigma(s)\, dW(s)\right). \end{aligned} \tag{1.21}$$

Being an exponential of a Gaussian random variable, $X(t)$ follows a *log-normal* distribution, with moments (see Karatzas and Shreve [1997])

$$\mathrm{E}^{\mathrm{P}}\left(X(t)\right) = X(0)\exp\left(\int_0^t \mu(s)\, ds\right), \tag{1.22}$$

$$\mathrm{E}^{\mathrm{P}}\left(X(t)^2\right) = \mathrm{E}^{\mathrm{P}}\left(X(t)\right)^2 \exp\left(\int_0^t \sigma(s)\sigma(s)^\top\, ds\right). \tag{1.23}$$

1.7 Explicit Trading Strategies and PDEs

After the mathematical interlude of Section 1.6, we now return to financial markets and a more careful analysis of the trading strategies that replicate derivative securities. We have already established that in a complete market such strategies must exist for any given derivative, but it still remains to determine these strategies explicitly. Consider a Markovian setup where the asset vector X satisfies an SDE of the form (1.19). Let there be given a derivative security V paying out at time T an amount $V(T) = g(X(T))$, for some smooth payout function $g : \mathbb{R}^p \to \mathbb{R}$. The Markovian form of the asset dynamics suggests that the time t derivative price is a function of t and $X(t)$ only, $V(t) = V(t, X(t))$ for some deterministic function $V(t, x)$, $x \in \mathbb{R}^p$. Conjecturing that this function is smooth enough to allow for an application of Ito's lemma for all $t \in [0, T)$, Theorem 1.1.5 implies (suppressing dependence on $X(t)$ for brevity)

$$\begin{aligned} dV(t) = V_t(t)\,dt &+ \sum_{i=1}^{p} V_{x_i}(t)\mu_i(t)\,dt \\ &+ \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} V_{x_i x_j}(t)\Sigma_{i,j}(t)\,dt + \sum_{i=1}^{p} V_{x_i}(t)\sigma_i(t)\,dW(t), \end{aligned} \tag{1.24}$$

where σ_i is the i-th row of the $p \times d$ matrix σ and $\Sigma_{i,j}$ is the (i, j)-th element in $\sigma\sigma^\top$. We recall that subscripts like V_{X_i} denote partial differentiation, see Theorem 1.1.5.

If $V(t)$ can be replicated by a self-financing trading strategy ϕ in the p assets, we must also have, from (1.10),

$$dV(t) = \phi(t)^\top dX(t) = \sum_{i=1}^{p} \phi_i(t)\mu_i(t)\,dt + \sum_{i=1}^{p} \phi_i(t)\sigma_i(t)\,dW(t). \tag{1.25}$$

Comparing terms in (1.24) and (1.25) we see that both equations will hold, provided that for all $t \in [0, T]$

$$\phi_i(t) = \frac{\partial V(t, X(t))}{\partial x_i}, \quad i = 1, \ldots, p, \tag{1.26}$$

and

$$\frac{\partial V(t,x)}{\partial t} + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^2 V(t,x)}{\partial x_i \partial x_j}\Sigma_{i,j}(t,x) = 0. \tag{1.27}$$

To the extent that the system above allows for a solution (it may not if the market is not complete, from (1.26) we see that the trading strategy that replicates the derivative V holds $\partial V(t, X(t))/\partial x_i$ units of asset X_i at time

t. The quantity $\partial V/\partial x_i$ is often known as the *delta* with respect to X_i[11]. Note that, from (1.9) and (1.26) we have that

$$V(t, X(t)) = \sum_{i=1}^{p} \frac{\partial V(t, X(t))}{\partial x_i} X_i(t). \tag{1.28}$$

Besides identifying an explicit replication strategy, the arguments above have also produced (1.27), a partial differential equation (PDE) for the value function $V(t, x)$. The PDE is a second-order parabolic equation in p spatial variables, with known terminal condition $V(T, x) = g(x)$ (a so-called *Cauchy problem*). Solving this PDE provides an alternative way to price the derivative, as compared to the purely probabilistic expectations-based methods outlined earlier (see (1.15)). We shall investigate the link between expectations and PDEs in more detail in Section 1.8.

Inspection of the valuation PDE (1.27) reveals that the drifts μ_i of the asset price SDE (1.19) are notably absent, making the price of the derivative security independent of drifts. This is typical of derivatives in complete markets and follows from the fact that derivatives can be priced preference-free, by arbitrage arguments. In contrast, for the elements of the fundamental asset price vector, risk-averse investors would demand that assets with high volatilities $|\sigma_i|$ be rewarded with higher drifts (more precisely, higher rates of return) as compensation for the additional uncertainty.

1.8 Kolmogorov's Equations and the Feynman-Kac Theorem

In earlier sections, we have seen that derivatives prices can be expressed as expectations under certain probability measures or as solutions to PDEs. This hints at a deeper connection between expectations and PDEs, a connection we shall explore in this section. As part of this exploration, we list results for transition densities that will be useful later in model calibration.

As in Section 1.6, we consider a Markov vector SDE of the type (see (1.19))

$$dX(t) = \mu(t, X(t))\, dt + \sigma(t, X(t))\, dW(t), \quad X(0) = X_0, \tag{1.29}$$

where the coefficients are assumed smooth enough to allow for a unique solution (see Theorem 1.6.1). Now define a functional

$$u(t, x) = \mathrm{E}^{\mathrm{P}}\left(g(X(T)) \,|X(t) = x\right),$$

[11]Note that taking a position in V and following a trading strategy with $\phi_i = -\partial V/\partial x_i$, $i = 1, \ldots, p$ will effectively remove any exposure to V (as we simultaneously take a long position in V and, through a trading strategy, a short position in V). This strategy is known as a *delta hedge*.

for a function $g : \mathbb{R}^p \to \mathbb{R}$. Under regularity conditions on g, it is easy to see that the process $u(t, X(t))$, being a conditional expectation, must be a martingale. Proceeding informally, an application of Ito's lemma gives, for $t \in [0, T)$ (suppressing dependence on $X(t)$),

$$du(t) = u_t(t)\,dt + \sum_{i=1}^{p} u_{x_i}(t)\mu_i(t)\,dt + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} u_{x_i x_j}(t)\Sigma_{i,j}(t)dt + O(dW(t)),$$

where as before $\Sigma_{i,j}$ is the (i,j)-th element of $\sigma\sigma^\top$. From earlier results, we know that for $u(t, X(t))$ to be a martingale, the term multiplying dt in the equation above must be zero. Defining the operator

$$\mathcal{A} = \sum_{i=1}^{p} \mu_i(t,x)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \Sigma_{i,j}(t,x)\frac{\partial^2}{\partial x_i \partial x_j},$$

we deduce that $u(t,x)$ satisfies the PDE

$$\frac{\partial u(t,x)}{\partial t} + \mathcal{A}u(t,x) = 0, \tag{1.30}$$

with terminal condition $u(T,x) = g(x)$. The equation above is known as the *Kolmogorov backward equation* for the SDE (1.29). The operator $\mathcal{A}$ is known as the *generator* or *infinitesimal operator* of the SDE, and can be identified as

$$\mathcal{A}u(t,x) = \lim_{h \downarrow 0} \frac{\mathrm{E}^{\mathrm{P}}\left(u\left(t+h, X(t+h)\right) | X(t) = x\right) - u(t,x)}{h}.$$

In arriving at (1.30) we made several implicit assumptions, most notably that the function $u(t,x)$ exists and is twice differentiable. Sufficient conditions for the validity of (1.30) can be found in Karatzas and Shreve [1997], for instance. A relevant result is listed below.

Theorem 1.8.1. *Let the process $X(t)$ be given by the SDE* (1.29), *where the coefficients μ and σ are continuous in x and satisfy the Lipschitz and growth conditions of Theorem 1.6.1. Consider a continuous function $g(x)$ that is either non-negative or satisfies a* polynomial growth condition, *meaning that for some positive constants K and q*

$$g(x) \le K(1 + |x|^q), \quad x \in \mathbb{R}^p.$$

If $u(t,x)$ solves (1.30) *with boundary condition $u(T,x) = g(x)$, and $u(t,x)$ satisfies a polynomial growth condition in x, then*

$$u(t,x) = \mathrm{E}^{\mathrm{P}}\left(g\left(X(T)\right) | X(t) = x\right), \quad t \in [0,T]. \tag{1.31}$$

Conditions required to ensure existence of a solution to (1.30) are more involved, and we just refer to Karatzas and Shreve [1997] and the references therein.

A family of functions g of particular importance to many of our applications is

$$g(x) = e^{ik^\top x}, \quad k \in \mathbb{R}^p,$$

where $i = \sqrt{-1}$ is the imaginary unit. In this case $u(t,x)$ becomes the *characteristic function* of $X(T)$, conditional on $X(t) = x$. We refer to any standard statistics textbook (e.g. Ochi [1990]) for the many useful properties of characteristic functions.

For the Markov process $X(t)$ in (1.29), let us now introduce a *transition density*, given heuristically by

$$p(t,x;s,y)\,dy \triangleq \mathrm{P}\left(X(s) \in [y, y+dy] | X(t) = x\right), \quad 0 \le t \le s \le T.$$

We can loosely think of the transition density as a special case of the functional $u(t,x)$ above, with boundary condition $u(s,x) = \delta(x-y)$, where $\delta(\cdot)$ is the Dirac delta function. Sometimes $p(\cdot,\cdot;\cdot,\cdot)$ is called a *Green's function* or a *fundamental solution* to (1.30). Under certain regularity conditions discussed in Karatzas and Shreve [1997], the transition density solves the Kolmogorov backward equation

$$\frac{\partial p(t,x)}{\partial t} + \mathcal{A}p(t,x) = 0, \quad (s,y) \text{ fixed},$$

subject to the boundary condition $p(s,x;s,y) = \delta(x-y)$. Further, the general expectation $u(t,x) = \mathrm{E}^{\mathrm{P}}(g(X(T))|X(t) = x)$ in Theorem 1.8.1 can be written

$$u(t,x) = \int_{\mathbb{R}^p} g(y)p(t,x;T,y)\,dy, \quad t \in [0,T]. \tag{1.32}$$

In many applications, it is useful to have a result that produces transition densities at future times $s \ge t$ from a known state at time t, rather than vice-versa. For this, we first define an operator $\mathcal{A}^*$ by

$$\mathcal{A}^* f(s,y) = -\sum_{i=1}^{p} \frac{\partial\left[\mu_i(s,y)f(s,y)\right]}{\partial y_i} + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^2\left[\Sigma_{i,j}(s,y)f(s,y)\right]}{\partial y_i \partial y_j}.$$

In the transition density $p(t,x;s,y)$ now consider (t,x) fixed and let $\mathcal{A}^*$ operate on the resulting function of s and y. Under additional regularity conditions, we then have the *forward Kolmogorov equation*

$$-\frac{\partial p(s,y)}{\partial s} + \mathcal{A}^* p(s,y) = 0, \quad (t,x) \text{ fixed}, \tag{1.33}$$

subject to the boundary condition $p(t,x;t,y) = \delta(x-y)$.

The forward Kolmogorov equation is sometimes known as the *Fokker-Planck* equation. We stress that the backward equation is more general than the forward equation, in the sense that the former holds for general terminal conditions $g(x)$, whereas the latter only holds for δ-type initial conditions.

We round off this section by a useful extension to the Kolmogorov backward equation. Specifically, consider extending the PDE (1.30) to

$$\frac{\partial u(t,x)}{\partial t} + \mathcal{A}u(t,x) + h(t,x) = r(t,x)u(t,x), \tag{1.34}$$

where $h, r : [0,T] \times \mathbb{R}^p \to \mathbb{R}$. Given the boundary condition $u(T,x) = g(x)$, the *Feynman-Kac solution* to (1.34), should it exist, is given by

$$u(t,x) = \mathrm{E}^{\mathrm{P}}\left(\psi(t,T)g\left(X(T)\right) + \int_t^T \psi(t,s)h\left(s, X(s)\right)\, ds \middle| X(t) = x \right), \tag{1.35}$$

where

$$\psi(t,T) = \exp\left(-\int_t^T r\left(s, X(s)\right)\, ds \right), \quad t \in [0,T].$$

The result is easily understood from an application of Ito's lemma, similar to the one used above to motivate the backward Kolmogorov equation. Sufficient regularity conditions for the Feynman-Kac result to hold are identical to those of Theorem 1.8.1, supplemented with the requirement that r be nonnegative and continuous in x; and the requirement that h be continuous in x and either be nonnegative or satisfy a polynomial growth requirement in x. See Duffie [2001] for further details about the often delicate regularity issues surrounding the Feynman-Kac result.

For later use, let us finally note that when $g(x) = \delta(x-y)$ and $h(t,x) = 0$, $u(t,x)$ in (1.35) will equal

$$G(t,x;T,y) \triangleq \mathrm{E}^{\mathrm{P}}\left(e^{-\int_t^T r(s,X(s))\, ds} \delta\left(X(T) - y\right) | X(t) = x \right).$$

The function G is known as a *state-price density* or as an *Arrow-Debreu security price* function. In particular, notice that for an arbitrary $g(x)$, we then have

$$\mathrm{E}^{\mathrm{P}}\left(e^{-\int_t^T r(s,X(s))\, ds} g\left(X(T)\right) | X(t) = x \right) = \int_{\mathbb{R}} G(t,x;T,y)\, g(y)\, dy. \tag{1.36}$$

Comparison with (1.32) shows that the state-price density is, essentially, equivalent to a Green's function with built-in discounting.

1.9 Black-Scholes and Extensions

In reviews of asset pricing theory, a discussion of the seminal *Black-Scholes-Merton* model (sometimes just known as the *Black-Scholes* model) of Black and Scholes [1973] and Merton [1973] is nearly mandatory. As the Black-Scholes-Merton (BSM) model constitutes a well-behaved setting in which to tie elements of previous sections together, our text is no exception. To provide a smoother transition to material that follows, we do, however, extend the usual analysis to include a simple case of stochastic interest rates.

1.9.1 Basics

In the basic BSM economy, two assets are traded: a money market account β and a stock S. In previous notations, $X(t) = (\beta(t), S(t))^\top$ and $p = 2$. The money market account value is 1 at time 0 and accrues risk-free interest at a continuously compounded, non-negative rate of r, initially assumed constant. The dynamics for β are thus given by an ordinary differential equation (ODE)

$$d\beta(t)/\beta(t) = r\,dt, \quad \beta(0) = 1,$$

implying that simply $\beta(t) = \beta(0)e^{rt}$.

The stock dynamics are assumed to satisfy GBMD under measure P:

$$dS(t)/S(t) = \mu\,dt + \sigma\,dW(t), \tag{1.37}$$

where W is a Brownian motion of dimension $d = 1$, and μ and σ are constants.

Taking first a probabilistic approach, we notice that β is positive and can be used as a numeraire. Let $S^\beta(t) = S(t)/\beta(t)$ be the stock price deflated by β. By Ito's lemma,

$$dS^\beta(t)/S^\beta(t) = (\mu - r)\,dt + \sigma\,dW(t).$$

Applying Girsanov's theorem (see Theorem 1.5.1) and Corollary 1.5.2, we see that if $\sigma \neq 0$, β will induce a unique equivalent martingale measure, with the measure shift characterized by the density process[12]

$$d\varsigma(t)/\varsigma(t) = -\theta\,dW(t), \quad \theta = \frac{\mu - r}{\sigma}.$$

Clearly, $\varsigma(t)$ defines an exponential martingale. The probability measure induced by the money market account β is called the *risk-neutral martingale measure* and is traditionally denoted Q. Under Q, $W^\beta(t) = W(t) + \theta t$ is a Brownian motion, and

$$\begin{aligned} dS^\beta(t)/S^\beta(t) &= \sigma\,dW^\beta(t), \\ dS(t)/S(t) &= r\,dt + \sigma\,dW^\beta(t), \end{aligned} \tag{1.38}$$

or, from (1.21),

$$S(T) = S(t)e^{(r-\frac{1}{2}\sigma^2)(T-t)+\sigma\left(W^\beta(T)-W^\beta(t)\right)}, \quad t \in [0, T]. \tag{1.39}$$

We note that under Q, the drift μ of the stock process is replaced by the risk-free interest rate r. That is, under Q agents in the economy will

[12] The reader may recognize the market price of risk θ as the *Sharpe ratio* of the stock S, a measure of how well the risk of stock (represented by σ) is compensated by excess return (represented by $\mu - r$).

appear to be indifferent ("neutral") to the risk of the stock, content with an average growth rate of the stock equal to that of the money market account.

Before proceeding with the BSM analysis, we wish to emphasize that the drift restriction imposed on the stock in the risk-neutral measure Q is a general result. In a larger setting with a p-dimensional vector asset process X, if the Q-dynamics of the components of X are all of the form

$$dX_i(t) = rX_i\,dt + O(dW(t)), \quad i = 1, \ldots, p,$$

there is no arbitrage. This result holds unchanged if the interest rate is random (see Section 1.9.3).

Returning to the BSM setting, we note that the risk-neutral measure is unique, whereby the market is complete and all derivative securities on S (and β) are attainable. Let us consider a few such securities. First, we consider a security paying at time T \$1 for certain. Such a security is a *discount bond* and we shall denote its time t price by $P(t,T)$, $t \in [0,T]$. If the interest rate is positive, we would expect $P(t,T) \leq 1$ as a reflection of the time value of money, with equality only holding for $t = T$. Application of the basic derivative pricing equation (1.15) immediately gives

$$P(t,T) = \beta(t)\mathrm{E}_t^{\mathrm{Q}}\left(\frac{1}{\beta(T)}\right) = \mathrm{E}_t^{\mathrm{Q}}\left(e^{-r(T-t)}\right) = e^{-r(T-t)}.$$

This result is trivial, as it is easily seen that the amount $e^{-r(T-t)}$ invested in the money market account at time t will grow to exactly \$1 at time T.

Second, consider a derivative V paying $V(T) = S(T) - K$ at time T, with K being an arbitrary constant. Proceeding as above, at time $t \leq T$ the arbitrage-free price must be

$$\begin{aligned} V(t) &= \mathrm{E}_t^{\mathrm{Q}}\left(e^{-r(T-t)}\left(S(T) - K\right)\right) \\ &= e^{-r(T-t)}\left(\mathrm{E}_t^{\mathrm{Q}}\left(S(T)\right) - K\right) = S(t) - KP(t,T), \end{aligned} \tag{1.40}$$

where the last equality follows from property (1.22) of GBMD. We notice that $V(t) = 0$ if $K = S(t)/P(t,T)$. This value of K is known as the time t *forward price of* $S(T)$[13].

Third, consider the derivative that was the main focus of the original BSM analysis, a *European call option* paying[14] $c(T) = (S(T) - K)^+$, with K being a positive *strike price*. Following (1.40), we can write

$$c(t) = P(t,T)\mathrm{E}_t^{\mathrm{Q}}\left((S(T) - K)^+\right). \tag{1.41}$$

From the representation (1.39), basic probability theory allows us to write this expectation as

[13] We shall touch on the closely related concept of a *futures price* in Section 4.1.2.

[14] We use the notations $x^+ = \max(x,0)$, $x^- = \min(x,0)$ throughout this book.

$$c(t) = P(t,T)\int_{-\infty}^{\infty}\left(S(t)e^{(r-\frac{1}{2}\sigma^2)(T-t)+z\sigma\sqrt{T-t}} - K\right)^+ \phi(z)\,dz, \qquad (1.42)$$

where $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$ is the standard Gaussian density. A straightforward evaluation of the integral leads to the famous *Black-Scholes-Merton call pricing formula*:

Theorem 1.9.1. *In the BSM economy, the arbitrage-free time t price of a K-strike call option maturing at time T is*

$$c(t) = S(t)\Phi(d_+) - KP(t,T)\Phi(d_-), \qquad (1.43)$$

$$d_\pm \triangleq \frac{\ln(S(t)/K) + (r \pm \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \quad t < T,$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function.

A formula for a *European put option* $p(t)$ paying $(K - S(T))^+$ can be obtained from (1.43) by *put-call parity*:

$$c(t) - p(t) = V(t),$$

where $V(t)$ is the forward contract defined above.

Remark 1.9.2. At time t, call and put options with strikes equal to $S(t)$ are said to be *at-the-money* (ATM). If $S(t) > K$, the call option is *in-the-money* (ITM) and the put option is *out-of-the-money* (OTM). If $S(t) < K$, the call is OTM and the put is ITM. The ATM, ITM, and OTM monikers are sometimes used to refer to the ordering of the *forward value* $\mathrm{E}_t(S(T)) = S(t)e^{r(T-t)}$ (for a T-maturity option) rather than the spot $S(t)$, relative to the strike K.

In deriving (1.43), the choice of β as numeraire was arbitrary. If we instead use S (which is also strictly positive) as numeraire, we can write

$$c(t) = S(t)\mathrm{E}_t^{\mathrm{Q}^S}\left(\frac{(S(T)-K)^+}{S(T)}\right) = S(t)\mathrm{E}_t^{\mathrm{Q}^S}\left((1 - K/S(T))^+\right), \qquad (1.44)$$

where Q^S is the martingale measure induced by S. To identify the measure shift involved in moving from P to Q^S, consider that $\beta^S(t) = \beta(t)/S(t)$ must be a martingale in Q^S. By Ito's lemma, in measure P we have

$$d\beta^S(t)/\beta^S(t) = (r - \mu + \sigma^2)\,dt - \sigma\,dW(t),$$

such that $dW^S(t) = ((r-\mu)/\sigma + \sigma)\,dt + dW(t)$ is a Brownian motion under Q^S. Application of Ito's lemma on $1/S(t)$ yields, after a few rearrangements,

$$dS(t)^{-1}/S(t)^{-1} = -r dt - \sigma\,dW^S(t),$$

which is a GBMD as before. Evaluation of the expectation (1.44) can be verified to recover the BSM formula (1.43).

Our derivation of the BSM formula was so far entirely probabilistic. Writing $c(t) = c(t, \beta, S)$, the arguments in Section 1.7 allow us to write c as a solution to the PDE (see (1.27))

$$\frac{\partial c}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 c}{\partial S^2} = 0, \tag{1.45}$$

subject to the boundary condition $c(T, \beta, S) = (S-K)^+$. From (1.28) we also have that the replication positions in β and S are $\frac{\partial c}{\partial \beta}$ and $\frac{\partial c}{\partial S}$, respectively. That is,

$$c(t, \beta, S) = \frac{\partial c}{\partial \beta}\beta + \frac{\partial c}{\partial S}S. \tag{1.46}$$

As β is deterministic, we can actually eliminate c-dependence on this variable by a change of variables $\widetilde{c}(t, S) = c(t, \beta, S)$. By the chain rule

$$\frac{\partial \widetilde{c}}{\partial t} = \frac{\partial c}{\partial t} + \frac{\partial c}{\partial \beta}\frac{\partial \beta}{\partial t} = \frac{\partial c}{\partial t} + \frac{\partial c}{\partial \beta} r\beta = \frac{\partial c}{\partial t} + rc - \frac{\partial c}{\partial S} rS$$

where the last equation follows from (1.46). Inserting this into (1.45) yields the original *Black-Scholes PDE*

$$\frac{\partial \widetilde{c}}{\partial t} + rS\frac{\partial \widetilde{c}}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 \widetilde{c}}{\partial S^2} = r\widetilde{c}, \tag{1.47}$$

with $\widetilde{c}(T, S) = (S - K)^+$. We can solve this equation by classical methods (see Lipton [2001] for several techniques), or we can use the Feynman-Kac result to write it as an expectation. We leave it as an exercise to the reader to verify that Feynman-Kac leads to the same expectation as derived earlier by probabilistic means (see (1.41)).

A final note: the derivation of the Black-Scholes PDE above was somewhat non-standard due to the initial assumption of option price being a function of the deterministic numeraire β. A more conventional (but entirely equivalent) argument sets up a portfolio of the call option and a position in the stock, and demonstrates that the stock position can be set such that the total portfolio growth is deterministic (risk-free) on $[t, t+dt]$. Equating the portfolio growth with the risk-free rate yields the Black-Scholes PDE (1.47). See Hull [2006] for details of this approach.

1.9.2 Alternative Derivation

We have already demonstrated several different ways of proving the BSM call pricing formula, but as shown in Andreasen et al. [1998] there are many more. One particularly enlightening proof is based on the concept of *local time* and shall briefly be discussed in this section. The proof, which borrows

from the results in Carr and Jarrow [1990], will also allow us to demonstrate the Tanaka extension of Ito's lemma, mentioned earlier in Section 1.1.

As above, we assume that the stock price process is as in (1.38), and define the forward stock price $F(t) \triangleq S(t)/P(t,T)$. Clearly,

$$dF(t)/F(t) = \sigma\, dW^{\beta}(t), \quad t \leq T, \tag{1.48}$$

where W^{β} is a Brownian motion in the risk-neutral measure. Define the random variable $I(t) = (F(t) - K)^+$. The first derivative of I with respect to F is an indicator function $1_{\{F(t)>K\}}$ and the second derivative can be interpreted as the Dirac delta function, $\delta(F(t) - K)$. As I is clearly not twice differentiable, Ito's lemma formally does not apply, but the Tanaka extension nevertheless gives us permission to write

$$\begin{aligned} dI(t) &= 1_{\{F(t)>K\}}\, dF(t) + \frac{1}{2}\sigma^2 F(t)^2 \delta\left(F(t) - K\right) dt \\ &= 1_{\{F(t)>K\}}\, \sigma F(t)\, dW^{\beta}(t) + \frac{1}{2}\sigma^2 K^2 \delta\left(F(t) - K\right) dt. \end{aligned}$$

In integrated form,

$$I(T) = I(t) + \int_t^T 1_{\{F(u)>K\}}\, \sigma F(u)\, dW^{\beta}(u) + \frac{1}{2}\sigma^2 K^2 \int_t^T \delta\left(F(u) - K\right) du.$$

The second integral in this expression is a random variable known as the *local time of F spent at the level K*, on the interval $[t,T]$. Taking expectations, it follows that

$$\mathrm{E}_t^{\mathrm{Q}}\left(I(T)\right) = I(t) + \frac{1}{2}\sigma^2 K^2 \int_t^T \mathrm{E}_t^{\mathrm{Q}}\left(\delta\left(F(u) - K\right)\right) du.$$

Here, if $p(t,y;u,x)$ is the density of $F(u)$ given $F(t) = y$, $u \geq t$, then obviously

$$\mathrm{E}_t^{\mathrm{Q}}\left(\delta\left(F(u) - K\right)\right) = p\left(t, F(t); u, K\right).$$

By the definition of $F(T)$ we have $F(T) = S(T)$, such that $I(T) = (S(T) - K)^+$. From (1.41), we may therefore write the time t European call option price as

$$\begin{aligned} c(t) &= P(t,T)\mathrm{E}_t^{\mathrm{Q}}\left(I(T)\right) \\ &= \left(S(t) - KP(t,T)\right)^+ + \frac{P(t,T)}{2}\sigma^2 K^2 \int_t^T p\left(t, F(t); u, K\right) du. \end{aligned} \tag{1.49}$$

The formula (1.49) decomposes the call option into a sum of two terms, the *intrinsic value* and the *time value*, respectively. The time value can be made more explicit by observing from the representation (1.39) that[15]

[15] This also follows directly from the fact that $F(u)$ is a log-normal random variable with moments given by (1.22) and (1.23).

$$p\left(t, F(t); u, K\right) = \frac{1}{K\sigma\sqrt{u-t}\sqrt{2\pi}} \exp\left(-\frac{1}{2}d(u)^2\right),$$
$$d(u) \triangleq \frac{\ln\left(F(t)/K\right) - \frac{1}{2}\sigma^2(u-t)}{\sigma\sqrt{u-t}}.$$

In other words, we have arrived at the following result.

Proposition 1.9.3. *The European call option price $c(t)$ on the process (1.38) can be written as*

$$c(t) = \left(S(t) - KP(t,T)\right)^+ + \frac{P(t,T)\sigma K}{2}\int_t^T \frac{\phi\left(d(u)\right)}{\sqrt{u-t}}\,du, \tag{1.50}$$

where $\phi(x)$ is the Gaussian density.

Explicit evaluation of the integral in (1.50) can be verified to produce the BSM formula in Theorem 1.9.1. We leave this as an exercise to the reader.

1.9.3 Extensions

1.9.3.1 Deterministic Parameters and Dividends

In our basic BSM setup, consider now first a simple extension to a deterministic interest rate $r(t)$ and a deterministic volatility $\sigma(t)$. Carrying out the analysis as before, we see that discount bond prices now become

$$P(t,T) = e^{-\int_t^T r(s)\,ds}. \tag{1.51}$$

The BSM call pricing formula (1.43) holds unchanged provided $P(t,T)$ is changed according to (1.51), and we redefine

$$d_\pm \triangleq \frac{\ln\left(S(t)/K\right) + \int_t^T \left(r(s) \pm \sigma(s)^2/2\right)ds}{\sqrt{\int_t^T \sigma(s)^2\,ds}}.$$

Let us further assume that the stock pays dividends at a deterministic rate of $q(t)$. Our framework so far, however, has assumed that assets pay no cash over $[0,T]$. To salvage the situation, consider a fictitious asset S^* obtained by reinvesting all dividends into the stock S itself. It is easily seen that

$$S^*(t) = S(t)e^{\int_0^t q(s)\,ds},$$

and clearly $S^*(t)$ satisfies the requirements of generating no cash flows on $[0,T]$. Stating the call option payout as

$$c(T) = \left(S(T) - K\right)^+ = \left(S^*(T)e^{-\int_0^T q(s)\,ds} - K\right)^+$$

and performing the pricing analysis of Section 1.9.1 on $S^*(t)$, rather than $S(t)$, results in a dividend-extended BSM call option formula:

$$c(t) = S(t)e^{-\int_t^T q(s)\,ds}\Phi(d_+) - KP(t,T)\Phi(d_-),$$

$$d_\pm \triangleq \frac{\ln\left(S(t)/K\right) + \int_t^T \left(r(s) - q(s) \pm \sigma(s)^2/2\right) ds}{\sqrt{\int_t^T \sigma(s)^2\,ds}}.$$

When the stock pays a dividend rate of $q(t)$, note that the risk-neutral process for $S(t)$ is

$$dS(t)/S(t) = (r(t) - q(t))\,dt + \sigma(t)\,dW^\beta(t),$$

which extends (1.38). Note that for the special case where $r(t) = q(t)$, $S(t)$ becomes a martingale and the call option price formula simplifies to

$$c(t) = P(t,T)\left(S(t)\Phi(d_+) - K\Phi(d_-)\right), \tag{1.52}$$

where now

$$d_\pm \triangleq \frac{\ln\left(S(t)/K\right) \pm \frac{1}{2}\int_t^T \sigma(s)^2\,ds}{\sqrt{\int_t^T \sigma(s)^2\,ds}}.$$

Remark 1.9.4. The martingale call formula (1.52) typically emerges when pricing options on futures and forward prices (see (1.48)) and is often called the *Black formula*, in honor of the work in Black [1976].

1.9.3.2 Stochastic Interest Rates

We now get even more ambitious and wish to consider call option pricing in the case where the interest rate r is stochastic. The money market account β becomes

$$\beta(t) = e^{\int_0^t r(s)\,ds},$$

and is now assumed an $\mathcal{F}_t$-measurable random variable. Proceeding as in Section 1.9.1, we find that under the risk-neutral measure Q, the call option price expression is (assuming that the stock pays no dividends)

$$c(t) = \beta(t)\mathrm{E}_t^{\mathrm{Q}}\left(\frac{1}{\beta(T)}\left(S(T) - K\right)^+\right) = \mathrm{E}_t^{\mathrm{Q}}\left(e^{-\int_t^T r(s)\,ds}\left(S(T) - K\right)^+\right). \tag{1.53}$$

In (1.53), we emphasize that the numeraire no longer can be pulled out from the expectation. Still, to simplify call option computations, it would be convenient to somehow remove the term $\exp(-\int_t^T r(s)\,ds)$ from the expectation in (1.53). By substituting 1 for $(S(T) - K)^+$ in the expression above, we first notice that

$$P(t,T) = \mathrm{E}_t^{\mathrm{Q}}\left(e^{-\int_t^T r(s)\,ds}\right).$$

This inspires us to perform a new measure shift, where we use the discount bond $P(t,T)$, rather than $\beta(t)$, as our numeraire. Let the martingale measure induced by $P(t,T)$ be denoted Q^T, often termed the *T-forward measure*. By the standard result (1.15) we have

$$\begin{aligned} c(t) &= P(t,T)\mathrm{E}_t^{\mathrm{Q}^T}\left(P(T,T)^{-1}\left(S(T)-K\right)^+\right) \\ &= P(t,T)\mathrm{E}_t^{\mathrm{Q}^T}\left(\left(S(T)-K\right)^+\right), \end{aligned}$$

where we have used that $P(T,T) = 1$. From Theorem 1.4.2, Q^T and Q are related by the density

$$\varsigma(t) = \mathrm{E}_t^{\mathrm{Q}}\left(\frac{d\mathrm{Q}^T}{d\mathrm{Q}}\right) = \frac{P(t,T)/P(0,T)}{\beta(t)}. \tag{1.54}$$

To proceed, we need to add more structure to the model by making assumptions about the stochastic process for $P(t,T)$. We shall spend considerable effort in subsequent chapters on this issue, but for this initial application we simply assume that $P(t,T)$ has Q dynamics

$$dP(t,T)/P(t,T) = r(t)\,dt - \sigma_P(t,T)\,dW_P(t), \tag{1.55}$$

where $\sigma_P(t,T)$ is deterministic and $W_P(t)$ is a Brownian motion correlated to the stock Brownian motion. Notice that the drift of $P(t,T)$ under Q is not freely specifiable and must be equal to the risk-free rate; see the discussion following (1.38). For clarity, let the stock Brownian motion be renamed $W_S(t)$, and assume that the correlation between $W_P(t)$ and $W_S(t)$ is a constant ρ. In the setting of vector-valued Brownian motion with independent components used in earlier sections, we can introduce correlation by writing $W(t) = (W_1(t), W_2(t))^\top$ and setting, say,

$$\begin{aligned} W_P(t) &= W_1(t), \\ W_S(t) &= \rho W_1(t) + \sqrt{1-\rho^2}W_2(t). \end{aligned}$$

The filtration $\{\mathcal{F}_t\}$ of our extended BSM setting is the one generated by the 2-dimensional $W(t)$.

Under Q^T, the deflated process $S^P(t) = S(t)/P(t,T)$ is a martingale. An application of Ito's lemma combined with the Diffusion Invariance Principle shows that the Q^T process for $S^P(t)$ is

$$dS^P(t)/S^P(t) = \sigma_P(t,T)\,dW_1(t) + \sigma(t)\left(\rho\,dW_1(t) + \sqrt{1-\rho^2}\,dW_2(t)\right), \tag{1.56}$$

where $\sigma(t)$ as before is the deterministic volatility of the stock S. We recognize $S^P(t)$ as a drift-free geometric Brownian motion with instantaneous variance of

$$\sigma_P(t,T)^2 + \sigma(t)^2 + 2\rho\sigma(t)\sigma_P(t,T).$$

Exploiting the convenient fact that $S^P(T) = S(T)$ and $c(T) = (S^P(T) - K)^+$ (as $P(T,T) = 1$), we get

$$\begin{aligned} c(t) &= P(t,T)\mathrm{E}_t^{\mathrm{Q}^T}\left(\left(S^P(T) - K\right)^+\right) \\ &= P(t,T)\int_{-\infty}^{\infty}\left(S^P(t)e^{-\frac{1}{2}v(t,T)+z\sqrt{v(t,T)}} - K\right)^+ \phi(z)\,dz, \end{aligned} \tag{1.57}$$

where we have defined the "term", or total, variance

$$v(t,T) \triangleq \int_t^T \left(\sigma_P(s,T)^2 + \sigma(s)^2 + 2\rho\sigma(s)\sigma_P(s,T)\right)\,ds. \tag{1.58}$$

Completing the integration (compare with (1.42)) and using $S^P(t) = S(t)/P(t,T)$, we arrive at a modified BSM-type call option formula:

Proposition 1.9.5. *Consider a BSM economy with stochastic interest rates evolving according to (1.55). Define term variance $v(t,T)$ as in* (1.58)*. Then, the T-maturity European call option price is*

$$\begin{aligned} c(t) &= S(t)\Phi(d_+) - KP(t,T)\Phi(d_-), \\ d_\pm &= \frac{\ln\left(S(t)/\left(KP(t,T)\right)\right) \pm \frac{1}{2}v(t,T)}{\sqrt{v(t,T)}}. \end{aligned}$$

Proposition 1.9.5 was originally derived in Merton [1973], using PDE methods. Extensions to dividend-paying stocks are straightforward and follow the arguments shown in Section 1.9.3.1.

1.10 Options with Early Exercise Rights

In our previous definition of a contingent claim, we assumed that the claim involved a single $\mathcal{F}_T$-measurable payout at time T. In reality, a number of derivative contracts may have intermediate cash payments from, say, scheduled coupons or through "rebates" for barrier-style options. Mostly, such complications are straightforwardly incorporated; see for instance Section 2.7.3. Of particular interest from a theoretical perspective are the claims that allow the holder to accelerate payments through *early exercise*. Derivative securities with early exercise are characterized by an adapted payout process $U(t)$, payable to the option holder at a stopping time (or *exercise policy*) $\tau \leq T$, chosen by the holder. If early exercise can take place at any time in some interval, we say that the derivative security is an *American option*; if exercise can only take place on a discrete set of dates, we say that it is a *Bermudan option*.

Let the allowed (and deterministic) set of exercise dates larger than or equal to t be denoted $\mathcal{D}(t)$, and suppose that we are given at time 0 a particular exercise policy τ taking values in $\mathcal{D}(0)$, as well as a pricing numeraire N inducing a unique martingale measure Q^N. Let $V^\tau(0)$ be the time 0 value of a derivative security that pays $U(\tau)$. Under some technical conditions on $U(t)$, we can write for the value of the derivative security

$$V^\tau(0) = \mathrm{E}^{\mathrm{Q}^N}\left(\frac{U(\tau)}{N(\tau)}\right), \tag{1.59}$$

where we have assumed, with no loss of generality, that $N(0) = 1$. Let $\mathcal{T}(t)$ be the time t set of (future) stopping times taking value in $\mathcal{D}(t)$. In the absence of arbitrage, the time 0 value of a security with early exercise into U must then be given by the *optimal stopping problem*

$$V(0) = \sup_{\tau\in\mathcal{T}(0)} V^\tau(0) = \sup_{\tau\in\mathcal{T}(0)} \mathrm{E}^{\mathrm{Q}^N}\left(\frac{U(\tau)}{N(\tau)}\right), \tag{1.60}$$

reflecting the fact that a rational investor would choose an exercise policy to optimize the value of his claim.

We can extend (1.60) to future times t by

$$V(t) = N(t) \sup_{\tau\in\mathcal{T}(t)} \mathrm{E}_t^{\mathrm{Q}^N}\left(\frac{U(\tau)}{N(\tau)}\right), \tag{1.61}$$

where $\sup_{\tau\in\mathcal{T}(t)} \mathrm{E}_t^{\mathrm{Q}^N}(U(\tau)/N(\tau))$ is known as the *Snell envelope* of U/N under Q^N. The process $V(t)$ must here be interpreted as the value of the option with early exercise, *conditional* on exercise not having taken place before time t. To make this explicit, let $\tau^* \in \mathcal{T}(0)$ be the optimal exercise policy, as seen from time 0. We can then write, for $0 < t \leq T$,

$$V(0) = \mathrm{E}^{\mathrm{Q}^N}\left(1_{\{\tau^*\geq t\}}V(t)/N(t)\right) + \mathrm{E}^{\mathrm{Q}^N}\left(1_{\{\tau^*<t\}}U(\tau^*)/N(\tau^*)\right), \tag{1.62}$$

where we break the time 0 value into two components: one from the time t value of the option, should it not have been exercised before time t; and one from the right to exercise on $[0,t]$. As we can always elect — possibly suboptimally — to never exercise on $[0,t]$, from (1.62) we see that

$$V(0) \geq \mathrm{E}^{\mathrm{Q}^N}\left(V(t)/N(t)\right),$$

which establishes that $V(t)/N(t)$ is a *supermartingale* under Q^N. This result also follows directly from known properties of the Snell envelope; see, e.g., Musiela and Rutkowski [1997].

For later use, focus now on the Bermudan case and assume that $\mathcal{D}(0) = \{T_1, T_2, \ldots, T_B\}$, where $T_1 > 0$ and $T_B = T$. For $t < T_{i+1}$, define $H_i(t)$ as the time t value of the Bermudan option when exercise is restricted to the dates $\mathcal{D}(T_{i+1}) = \{T_{i+1}, T_{i+2}, \ldots, T_B\}$. That is

$$H_i(t) = N(t)\mathrm{E}_t^{\mathrm{Q}^N}\left(V(T_{i+1})/N(T_{i+1})\right), \quad i = 1, \ldots, B-1.$$

At time T_i, $H_i(T_i)$ can be interpreted as the *hold value* of the Bermudan option, that is, the value of the Bermudan option if not exercised at time T_i. If an optimal exercise policy is followed, clearly we must have at time T_i

$$V(T_i) = \max\left(U(T_i), H_i(T_i)\right), \quad i = 1, \ldots, B, \tag{1.63}$$

such that

$$H_i(t) = N(t)\mathrm{E}_t^{\mathrm{Q}^N}\left(\max\left(U(T_{i+1}), H_{i+1}(T_{i+1})\right)/N(T_{i+1})\right), \; i = 1, \ldots, B-1. \tag{1.64}$$

Starting with the terminal condition $H_B(T) = 0$, (1.64) defines a useful iteration backwards in time for the value $V(0) = H_0(0)$. We shall use this later for the purposes of designing valuation algorithms in Chapter 18, and for computing price sensitivities (deltas) in Chapter 24.

We note that the idea behind (1.63) is often known as *dynamic programming* or the *Bellman principle*. Loosely speaking, we here work "from the back" to price the Bermudan option. As we shall see later (in Chapter 2), this idea is particularly well-suited for numerical methods that proceed backwards in time, such as finite difference methods.

1.10.1 The Markovian Case

We now specialize to the Markovian case where $U(t) = g(t, x(t))$, where $g : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ is continuous and

$$dx(t) = \mu\left(t, x(t)\right) dt + \sigma\left(t, x(t)\right) dW(t) \tag{1.65}$$

is an n-dimensional Markovian process, where μ and σ satisfy the regularity conditions of Theorem 1.6.1. The n-dimensional process[16] $x(t)$ here defines the state of the exercise value $U(t)$, so we say that $x(t)$ is a *state variable process*. For concreteness let our numeraire $N(t)$ be the money market account

$$N(t) = \beta(t) = e^{\int_0^t r(u, x(u))\, du},$$

where the short interest rate $r : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ is here assumed a function of time and the state variable vector x. In (1.65), $W(t)$ is understood to be a d-dimensional Brownian motion in the risk-neutral measure Q.

Writing $V(t) = V(t, x(t))$, we have from (1.61)

$$V(t, x) = \sup_{\tau \in \mathcal{T}(t)} \mathrm{E}^{\mathrm{Q}}\left(e^{-\int_t^\tau r(u, x(u))\, du} g\left(\tau, x(\tau)\right) \middle| x(t) = x\right). \tag{1.66}$$

[16] Note that $x(t)$ is an abstract construct, and does not necessarily coincide with any asset price process.

For dates $t \in \mathcal{D}(0)$, clearly $V(t,x) \geq g(t,x)$, with equality holding only when time t exercise is optimal. This leads us to define the concept of an *exercise region* as

$$\mathcal{X} = \{(t,x) \in \mathcal{D}(0) \times \mathbb{R}^n : V(t,x) = g(t,x)\}.$$

Similarly, we define the complement of $\mathcal{X}$,

$$\mathcal{C} = \{(t,x) \in [0,T] \times \mathbb{R}^n : (t,x) \notin \mathcal{X}\},$$

to be the *continuation region*, i.e. the region where we wait (either because exercise is not optimal or because it is not allowed, $t \notin \mathcal{D}(0)$) rather than exercise the option.

For Markovian systems, rather than solving the optimization problem (1.66) directly, it is often particularly convenient to invoke the Bellman principle. Extending the ideas presented earlier, let us, somewhat loosely, state the Bellman principle as follows: for any $t \in \mathcal{D}(0)$,

$$V(t,x) = \lim_{\Delta \downarrow 0} \max\left(g(t,x), \mathrm{E}_t^{\mathrm{Q}}\left(e^{-\int_t^{t+\Delta} r(u,x(u))\,du} V\left(t+\Delta, x(t+\Delta)\right)\right)\right). \tag{1.67}$$

Again, this simply says that the option value at time t is the maximum of the exercise value and the hold value, that is, the present value of continuing to hold on to the option for a small period of time. As we have seen above, for a Bermudan option, (1.67) also holds for finite Δ (namely up to the next exercise date).

The Bellman principle provides us with a link between present (time t) and future (time $t+\Delta$) option values that we can often exploit in a numerical scheme. For this, however, we need further characterization of $V(t,x)$ in the continuation region. By earlier arguments, we realize that $V(t,x)/\beta(t)$ must be a Q-martingale on the continuation region. Assuming sufficient smoothness for an application of Ito's lemma, this leads to a PDE formulation, to hold for $(t,x) \in \mathcal{C}$,

$$\mathcal{J}V(t,x) = 0, \tag{1.68}$$

where

$$\mathcal{J} = \frac{\partial}{\partial t} + \mu(t,x)\frac{\partial}{\partial x} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sigma(t,x)\sigma(t,x)^{\top}\right)_{i,j}\frac{\partial^2}{\partial x_i \partial x_j} - r(t,x).$$

Assume first that our option is of the Bermudan type, and let T_i and T_{i+1} be subsequent exercise dates in the exercise schedule. For any function f of time, define $f(t\pm)$ to be the limits $\lim_{\varepsilon\downarrow 0} f(t \pm \varepsilon)$, and assume that $V(T_{i+1}-,x)$ is known for all x. As all values of $t \in (T_i, T_{i+1})$ by definition must be in the continuation region, we can use (1.68) to solve for $V(T_i+,x)$. Applying the Bellman principle (1.67) at time T_i then leads to the condition

$$V(T_i-,x) = \max\left(g(T_i,x), V(T_i+,x)\right).$$

In PDE parlance, this is a so-called *jump condition* which is straightforward to incorporate into a numerical solution; see Section 2.7.4 for details.

For American-style options, (1.68) continues to apply on $\mathcal{C}$. The Bellman principle here leads to the characterization that

$$\mathcal{J}V(t,x) < 0,$$

for $(t,x) \in \mathcal{X}$, i.e. we exercise when the rate of return from holding the option strictly fails to match $r(t,x)$. The American option pricing problem is often conveniently summarized in a *variational inequality,* to hold on $\mathcal{X} \cup \mathcal{C}$,

$$V(t,x) \geq g(t,x), \quad \mathcal{J}V(t,x) \leq 0, \quad (V(t,x) - g(t,x))\,\mathcal{J}V(t,x) = 0, \tag{1.69}$$

and subject to the boundary condition $V(T,x) = g(T,x)$. The first of these three conditions expresses that the option is always worth at least its exercise value; the second expresses the supermartingale property of $V(t,x)$; and the third implies (after a little thought) that $\mathcal{J}V(t,x) = 0$ on $\mathcal{C}$ and $\mathcal{J}V(t,x) < 0$ on $\mathcal{X}$. The system (1.69) is discussed more carefully in Duffie [2001], where additional discussion of regularity issues may also be found.

1.10.2 Some General Bounds

In many cases of practical interest, solving PDEs and/or variational inequalities is not computationally feasible. In such situations, we may be interested in at least bounding the value of an option with early exercise rights. Providing a lower bound is straightforward: postulate an exercise policy τ and compute the price $V^\tau(0)$ by direct methods. From (1.60), clearly this provides a lower bound

$$V^\tau(0) \leq V(0). \tag{1.70}$$

The closer the postulated exercise policy τ is to the optimal exercise policy τ^*, the tighter this bound will be. We shall later study a number of numerical techniques to generate good exercise strategies for fixed income options with early exercise rights, see Chapter 18.

To produce an upper bound, we can rely on duality results established in Rogers [2001], Haugh and Kogan [2004] and Andersen and Broadie [2004]. Let $\mathcal{K}$ denote the space of adapted martingales M for which $\sup_{\tau\in[0,T]} \mathrm{E}^{\mathrm{Q}^N}|M(\tau)| < \infty$. For a martingale $M \in \mathcal{K}$, we then write

$$\begin{aligned} V(0) &= \sup_{\tau\in\mathcal{T}(0)} \mathrm{E}^{\mathrm{Q}^N}\left(\frac{U(\tau)}{N(\tau)}\right) \\ &= \sup_{\tau\in\mathcal{T}(0)} \mathrm{E}^{\mathrm{Q}^N}\left(\frac{U(\tau)}{N(\tau)} + M(\tau) - M(\tau)\right) \\ &= M(0) + \sup_{\tau\in\mathcal{T}(0)} \mathrm{E}^{\mathrm{Q}^N}\left(\frac{U(\tau)}{N(\tau)} - M(\tau)\right). \end{aligned}$$

In the second equality, we have relied on the *optional sampling theorem*, a result that states that the martingale property is satisfied up to a bounded random stopping time, i.e. that $\mathrm{E}^{\mathrm{Q}^N}(M(\tau)) = M(0)$; see Karatzas and Shreve [1997] for details. We now turn the above result into an upper bound by forming a pathwise maximum at all possible future exercise dates $\mathcal{D}(0)$:

$$\begin{aligned} V(0) &= M(0) + \sup_{\tau \in \mathcal{T}(0)} \mathrm{E}^{\mathrm{Q}^N}\left(\frac{U(\tau)}{N(\tau)} - M(\tau)\right) \\ &\leq M(0) + \mathrm{E}^{\mathrm{Q}^N}\left(\max_{t \in \mathcal{D}(0)}\left(\frac{U(t)}{N(t)} - M(t)\right)\right). \end{aligned} \tag{1.71}$$

With (1.70) and (1.71) we have, as desired, established upper and lower bounds for values of options with early exercise rights. Let us consider how to make these bounds tight. As mentioned earlier, to tighten the lower bound we need to pick exercise strategies close to the optimal one. Tightening the upper bound is a bit more involved and requires the following basic theorem, proven in Karatzas and Shreve [1997]:

Theorem 1.10.1 (Doob-Meyer Decomposition). *Let* $\{Y(t),\ t \in [0,T]\}$ *be a positive* $\mathcal{F}_t$*-adapted supermartingale process with right-continuous sample paths. Then we can write*

$$Y(t) = m(t) - A(t),$$

where $m(t)$ *is a martingale process with* $m(0) = Y(0)$ *and* $A(t)$ *is an increasing predictable process with* $A(0) = 0$.

Applying the Doob-Meyer decomposition on the supermartingale process $V(t)/N(t)$ under Q^N shows that

$$V(t)/N(t) = m(t) - A(t),$$

and $V(0) = m(0)$. Consider taking $M(t) = m(t)$ in equation (1.71), to get

$$\begin{aligned} V(0) &\leq V(0) + \mathrm{E}^{\mathrm{Q}^N}\left(\max_{t \in \mathcal{D}(0)}\left(\frac{U(t)}{N(t)} - m(t)\right)\right) \\ &= V(0) + \mathrm{E}^{\mathrm{Q}^N}\left(\max_{t \in \mathcal{D}(0)}\left(\frac{U(t)}{N(t)} - \frac{V(t)}{N(t)} - A(t)\right)\right) \\ &\leq V(0). \end{aligned}$$

The last inequality follows from the fact that $V(t) \geq U(t)$ and $A(t) \geq 0$. In conclusion, we have arrived at a *dual* formulation of the option price

$$V(0) = \inf_{M \in \mathcal{K}}\left\{M(0) + \mathrm{E}^{\mathrm{Q}^N}\left(\max_{t \in \mathcal{D}(0)}\left(\frac{U(t)}{N(t)} - M(t)\right)\right)\right\}, \tag{1.72}$$

and have demonstrated that the infimum is attained when the martingale M is set equal to the martingale component of the deflated price process

$V(t)/N(t)$. In practice, we are obviously not privy to $V(t)/N(t)$ (which is a quantity that we are trying to estimate), but we are nevertheless provided with a strategy to make the upper bound (1.71) tight: use a martingale that is "close" to the martingale component of the true deflated option price process. In Chapter 18 we shall demonstrate how to make this strategy operational.

1.10.3 Early Exercise Premia

We finish our discussion of options with early exercise rights by listing some known results for puts and calls, including an interesting decomposition of American and Bermudan option prices into the sum of a European option price and an *early exercise premium*. For convenience, we work in a Markovian setting where the single state variable, denoted $S(t)$, follows one-dimensional GBMD. Specifically, we assume that

$$dS(t)/S(t) = (r - q)\,dt + \sigma\,dW^{\beta}(t), \tag{1.73}$$

with $W^{\beta}(t)$ being a one-dimensional Brownian motion in the risk-neutral measure, i.e. the measure induced by the money market account $\beta(t) = e^{rt}$. For simplicity we assume that the interest rate r, the dividend yield q, and the volatility σ are all constants; the extension to time-dependent parameters is straightforward.

Let $c(t)$, $C_A(t)$, and $C_B(t)$ be the time t European, American, and Bermudan prices of the call option with terminal maturity T, conditional on no exercise prior to time t. While obviously $c(t) \le C_B(t) \le C_A(t)$, in some cases these inequalities are equalities, as the following straightforward lemma shows.

Lemma 1.10.2. *Suppose that $r \ge 0$ and $q \le 0$ in (1.73). It is then never optimal to exercise a call option early, and*

$$c(t) = C_A(t) = C_B(t).$$

Proof. Notice that, by Jensen's inequality,

$$\begin{aligned} c(t) &= e^{-r(T-t)}\mathrm{E}_t^{\mathrm{Q}}\left((S(T) - K)^+\right) \\ &\ge e^{-r(T-t)}\left(\left(\mathrm{E}_t^{\mathrm{Q}}\left(S(T)\right) - K\right)^+\right) = \left(e^{-q(T-t)}S(t) - e^{-r(T-t)}K\right)^+. \end{aligned}$$

It is therefore clear that if $r \ge 0$ and $q \le 0$, then for any value of $T - t$,

$$c(t) \ge (S(t) - K)^+,$$

i.e. the European call option price dominates the exercise value. As the hold value of American and Bermudan options must be at least as large as the European option price, it follows that the option to exercise early is worthless. □

Remark 1.10.3. For the put option, early exercise is never optimal if $r \leq 0$ and $q \geq 0$. As this situation rarely happens in practice, American put options nearly always trade at a premium to their European counterparts.

Lemma 1.10.2 demonstrates the well-known fact that American or Bermudan call options on stocks that pay no dividends ($q = 0$) should never be exercised early. On the other hand, if the stock does pay dividends, for an American call option there will, at time t, be a critical value of the stock, $S_A(t)$, at which the value of the stream of dividends paid by the stock will compensate for the cost of accelerating the payment of the strike K. In other words, an American option should be exercised at time t, provided that $S(t) \geq S_A(t)$. The deterministic curve $S_A(t)$ is known as the *early exercise boundary* and marks the boundary between the exercise and continuation regions, $\mathcal{X}$ and $\mathcal{C}$. Writing $C_A(t) = C_A(t, S(t))$, we formally have

$$S_A(t) = \inf\left\{S : C_A(t,S) = (S-K)^+\right\}, \quad t \leq T.$$

For a Bermudan option, we may similarly define

$$S_B(t) = \inf\left\{S : C_B(t,S) = (S-K)^+\right\}, \quad t \in \mathcal{D}(0),$$

where we recall that $\mathcal{D}(0)$ is the (discrete) set of allowed exercise dates for the Bermudan option.

The following important result characterizes the exercise boundary of American call options.

Proposition 1.10.4. *For the American call option on a stock that follows (1.73), we have*

$$\frac{\partial S_A(t)}{\partial t} \leq 0, \quad t < T, \tag{1.74}$$

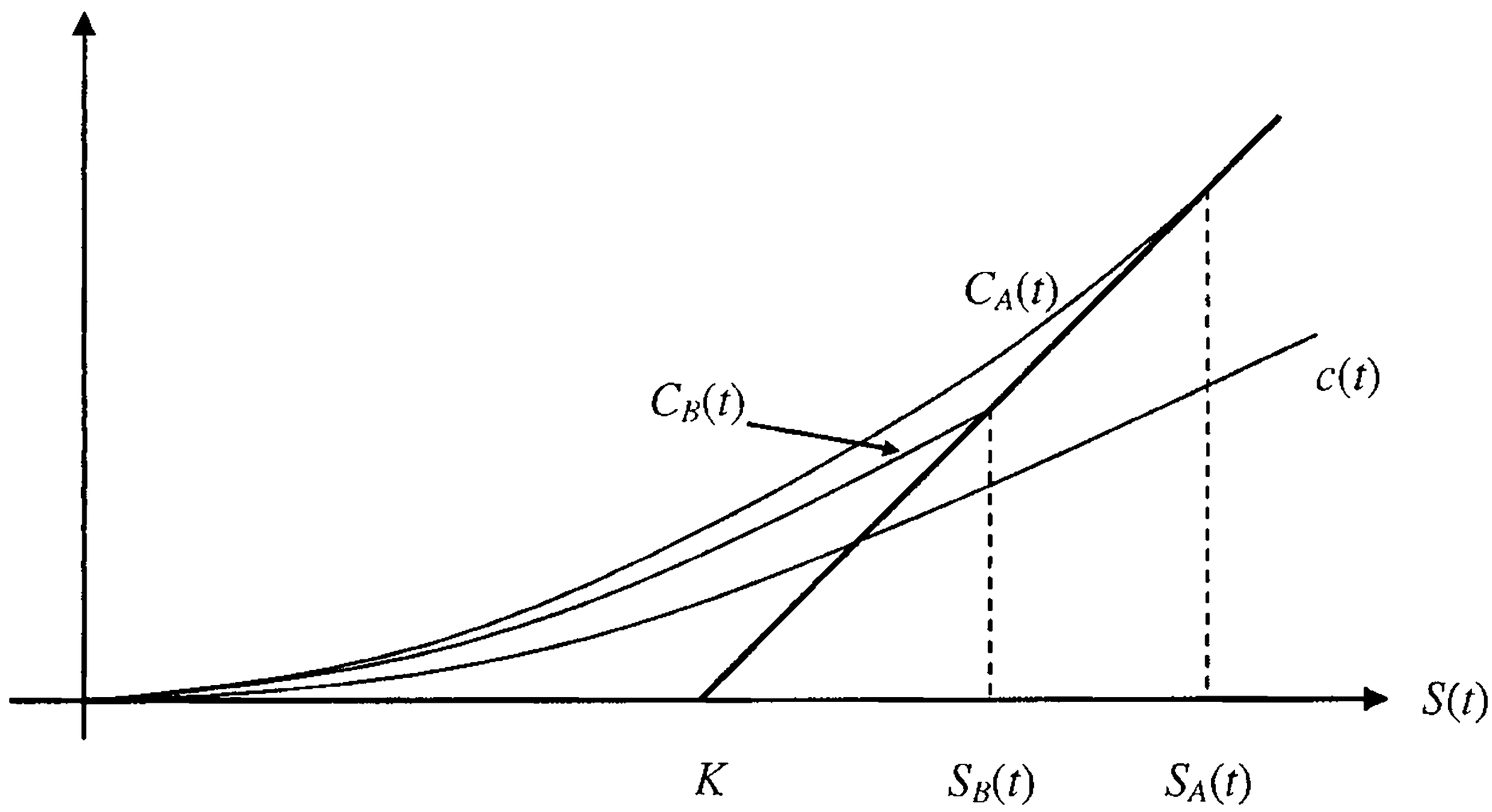
and

$$\left.\frac{\partial C_A(t,S)}{\partial S}\right|_{S=S_A(t)} = 1, \quad t < T. \tag{1.75}$$

Equation (1.74) states that the exercise boundary decreases as we approach maturity, a result that is easily understood. Statement (1.75) is more subtle, however, and amounts to a tangency condition that ensures that the American call option value transitions smoothly from hold value to exercise value across the early exercise boundary. As a consequence, (1.75) is often known as the *smooth pasting condition* or the *high contact condition.* A similar tangency condition does *not* hold for the Bermudan option value, which is not differentiable at the boundary but instead transitions into the exercise region at a "kink":

$$\lim_{\varepsilon \downarrow 0} \left.\frac{C_B(t,S) - C_B(t,S-\varepsilon)}{\varepsilon}\right|_{S=S_B(t)} < 1, \quad t \in \mathcal{D}(0).$$

Fig. 1.1. Call Option Prices



Notes: Time t prices of American, Bermudan, and European call options, as a function of the asset price. The Bermudan option is assumed to be exercisable at time t.

Figure 1.1 shows a typical value profile for a Bermudan call, along with the corresponding profiles for the European and American options.

Smooth pasting is essentially an optimality condition, which is how Proposition 1.10.4 is traditionally derived (see, e.g., Merton [1973] or the more recent Brekke and Øksendal [1991]). A more descriptive proof based on hedging arguments is given in Tavella and Randall [2000] and Wilmott et al. [1993]. Loosely speaking, the idea is here that a delta hedger should not be able to make riskless profits when the underlying asset crosses into the exercise region. This requires that the delta is continuous across the boundary, which is (1.75).

Remark 1.10.5. For the American put option, $\partial S_A(t)/\partial t \geq 0$ and the high contact condition states that the delta equals -1 at the exercise boundary.

Establishing the boundary $S_A(t)$ will virtually always require numerical methods, although asymptotic results are known for t close to T (see for instance Lipton [2001]). One simple result is listed below.

Lemma 1.10.6. *Assume that $r \geq 0$ and $q \geq 0$, such that the early exercise boundary exists for the American call option. The exercise boundary just prior to maturity is then*

$$\lim_{\varepsilon \downarrow 0} S_A(T-\varepsilon) = K \max\left(1, \frac{r}{q}\right).$$

Proof. An informal proof of Lemma 1.10.6 proceeds as follows. At time $T - dt$, assume that $S(T - dt) > K$; otherwise it clearly makes no sense to exercise the option. If we exercise the option, we receive $S(T - dt) - K$ at time $T - dt$. On the other hand, if we postpone exercise, at time $T - dt$ our hold value is

$$\begin{aligned} e^{-r\,dt}\mathrm{E}^{\mathrm{Q}}_{T-dt}\left(S(T) - K\right) &= S(T - dt)e^{-q\,dt} - Ke^{-r\,dt} \\ &= S(T - dt) - K - S(T - t)q\,dt + Kr\,dt. \end{aligned}$$

Clearly, we should then only exercise if

$$S(T - dt) - K > S(T - dt) - K - S(T - dt)q\,dt + Kr\,dt$$

or if

$$S(T - dt)q > Kr.$$

□

Notice that since clearly $S_A(T) = K$, the call option exercise boundary will have a *discontinuity* at time T, if $q < r$.

One might guess that complete knowledge of the curve $S_A(t)$ should suffice to price the American option analytically. This intuition is confirmed by the following result due to Jamshidian [1992], Carr et al. [1992], Kim [1990], and Jacka [1991].

Proposition 1.10.7. *The American option price $C_A(t)$ satisfies*

$$C_A(t) = c(t) + E_A(t), \quad t \le T, \tag{1.76}$$

where the (American) early exercise premium $E_A(t)$ *is defined as*

$$E_A(t) = \int_t^T e^{-r(u-t)}\mathrm{E}^{\mathrm{Q}}_t\left(1_{\{S(u)\ge S_A(u)\}}\left(qS(u) - rK\right)\right)du \tag{1.77}$$

$$= \int_t^T \left(qS(t)e^{-q(u-t)}\Phi\left(d_+(u)\right) - rKe^{-r(u-t)}\Phi\left(d_-(u)\right)\right)du, \tag{1.78}$$

where

$$d_\pm(u) = \frac{\ln\left(S(t)/S_A(u)\right) + \left(r - q \pm \frac{1}{2}\sigma^2\right)(u - t)}{\sigma\sqrt{u - t}}.$$

.

Proof. Due to the smooth pasting condition in Proposition 1.10.4, we are justified[17] in applying Ito's lemma. In informal notation,

$$\begin{aligned} dC_A(t) &= 1_{\{S(t)\ge S_A(t)\}}\,dS(t) \\ &\quad + 1_{\{S(t)<S_A(t)\}}\left\{\frac{\partial C_A(t)}{\partial t}\,dt + \frac{\partial C_A(t)}{\partial S}\,dS(t) + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C_A(t)}{\partial S^2}\,dt\right\}, \end{aligned} \tag{1.79}$$

[17] In particular, there is no local time contribution to $dC_A(t)$ at the boundary.

where we have used the fact that

$$1_{\{S(t)\geq S_A(t)\}}C_A(t) = 1_{\{S(t)\geq S_A(t)\}}(S(t)-K).$$

In the continuation region, $C_A(t,S)$ satisfies the PDE (1.47), i.e.

$$\frac{\partial C_A(t,S)}{\partial t} + (r-q)S\frac{\partial C_A(t,S)}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C_A(t,S)}{\partial S^2} = rC_A(t,S).$$

Inserting this into (1.79) we get, after a few rearrangements,

$$\begin{aligned} dC_A(t) &= rC_A(t)\,dt + 1_{\{S(t)<S_A(t)\}}(r-q)S(t)\frac{\partial C_A(t)}{\partial S}\,dW^\beta(t) \\ &\quad + 1_{\{S(t)\geq S_A(t)\}}\left\{((r-q)S(t) - rC_A(t))\,dt + \sigma S(t)\,dW^\beta(t)\right\} \\ &= rC_A(t)\,dt + 1_{\{S(t)<S_A(t)\}}(r-q)S(t)\frac{\partial C_A(t)}{\partial S}\,dW^\beta(t) \\ &\quad + 1_{\{S(t)\geq S_A(t)\}}\left\{(rK - qS(t))\,dt + \sigma S(t)\,dW^\beta(t)\right\} \end{aligned}$$

Setting $y(t) = C_A(t)/\beta(t)$, it follows from Ito's lemma that

$$\begin{aligned} dy(t) &= e^{-rt}1_{\{S(t)<S_A(t)\}}(r-q)S(t)\frac{\partial C_A(t)}{\partial S}\,dW^\beta(t) \\ &\quad + 1_{\{S(t)\geq S_A(t)\}}e^{-rt}\left\{(rK - qS(t))\,dt + \sigma S(t)\,dW^\beta(t)\right\}. \end{aligned}$$

Integrating and taking expectations leads to

$$\mathrm{E}_t^{\mathrm{Q}}\left(y(T)\right) = y(t) + \int_t^T e^{-ru}\mathrm{E}_t^{\mathrm{Q}}\left(1_{\{S(u)\geq S_A(u)\}}\left(rK - qS(u)\right)\right)du.$$

Applying the definition of $y(t)$ and the fact that $y(T) = e^{-rT}(S(T)-K)^+$ proves (1.76). The explicit form of the early exercise premium in (1.78) follows from the properties of GBMD. □

Remark 1.10.8. Combining results from Lemma 1.10.6 and Proposition 1.10.4, it follows that $E_A(t) \geq 0$, so $C_A(t) \geq c(t)$ as expected.

The integral representation of the American call option in Proposition 1.10.7 forms the basis for a number of proposed computational methods for American option pricing. Loosely speaking, these methods are based on the idea of iteratively estimating the exercise boundary $S_A(t)$, often working backwards from $t = T$, after which an application of Proposition 1.10.7 will yield the American option price. A representative example of these methods can be found in Ju [1998]. See Chiarella et al. [2004] for a survey of the literature, and Section 19.7.3 for an application in interest rate derivative pricing.

For a Bermudan option, an integral representation such as that in Proposition 1.10.7 is not possible. Nevertheless, it is still possible to break the

Bermudan call option into the sum of a European option and an early exercise premium. To show this, assume that the allowed exercise dates are $\mathcal{D}(0) = \{T_1, T_2, \ldots, T_B\}$, and let $S_B(T_i)$ be the exercise level above which the Bermudan option should be exercised at time T_i, $i = 1, \ldots, B$. Notice that if at time T_i we have $S(T_i) > S_B(T_i)$, then C_B will *jump down* in value when time progresses past time T_i, as a reflection of the missed exercise opportunity. Indeed, in the earlier notation of hold and exercise values, we have

$$\begin{aligned} C_B(T_i) &= \max\left(U(T_i), H(T_i)\right), \\ C_B(T_i+) &= H(T_i), \end{aligned}$$

which makes the jump in value evident. Given the existence of these jumps, we may write

$$\begin{aligned} dC_B(t) &= rC_B(t)\,dt + dM(t) \\ &\qquad + \sum_{i=1}^{B} 1_{\{S(T_i)>S_B(T_i)\}}\delta(T_i - t)\left(H(T_i) - U(T_i)\right)\,dt, \end{aligned}$$

where $H(T_i) = C_B(T_i+)$ is the hold value at time T_i, $U(T_i) = S(T_i) - K$, and $M(t)$ is a martingale,

$$dM(t) = \frac{\partial C_B(t)}{\partial S}(r - q)S(t)\,dW^{\beta}(t).$$

Deflating C_B by the money market account and forming expectations, we get, since $c(t) = e^{-r(T-t)}\mathrm{E}_t^{\mathrm{Q}}(C_B(T))$,

$$C_B(t) = c(t) + \sum_{T_i \geq t} e^{-r(T_i - t)}\mathrm{E}_t^{\mathrm{Q}}\left(1_{\{S(T_i)>S_B(T_i)\}}\left(U(T_i) - H(T_i)\right)\right).$$

As $H(T_i)$ must be less than the exercise value $U(T_i)$ whenever $S(T_i) > S_B(T_i)$ we can simplify this expression to the following result that we, in Section 18.2.3, call the *marginal exercise value decomposition*.

Proposition 1.10.9. *The Bermudan option price $C_B(t)$ satisfies*

$$C_B(t) = c(t) + E_B(t), \quad t \leq T,$$

where the (Bermudan) early exercise premium $E_B(t)$ *is defined as*

$$E_B(t) = \sum_{T_i \geq t} e^{-r(T_i - t)}\mathrm{E}_t^{\mathrm{Q}}\left(\left(U(T_i) - H(T_i)\right)^+\right),$$

with $T_1 < T_2 < \ldots < T_B = T$ being the set of exercise dates.

As shown in Section 18.2.3, the result in Proposition 1.10.9 may be extended to more complicated processes and payouts than those considered here.

2

Finite Difference Methods

In Chapter 1 we described how the pricing of a derivative security typically requires either the solution of a parabolic partial differential equation (PDE) or the evaluation of an expectation of a random variable. In realistic applications, both of these price formulations often do not allow for closed-form solution, in which case we must resort to either analytical approximations or, more generally, numerical techniques. In the next two chapters we will describe a number of numerical algorithms useful in derivatives pricing. Analytical approximations will receive ample treatment later in this book, in the context of specific problems.

Our treatment of numerical methods is broken into two main subjects. In this chapter, we cover finite difference solutions of PDEs; and in Chapter 3 we turn to Monte Carlo evaluation of expectations. Many excellent specialist books exist on both topics, including Mitchell and Griffiths [1980], Tavella and Randall [2000], and Glasserman [2004]; our treatment only surveys the most important concepts, as required for our needs in this book. We do provide, however, a number of schemes rarely described in detail in the finance literature and also supplement our analysis with a number of "tricks of the trade", particularly in the application of finite difference grids.

The analysis of numerical PDE solutions in this chapter is arranged in two blocks. First, in Sections 2.1–2.8 we study the basic mechanics of the finite difference grid method for one-dimensional PDEs. Subsequently, Sections 2.9–2.12 then apply operator splitting techniques to extend the finite difference method to PDE of dimensions two and higher. The analysis culminates with a presentation of *ADI schemes* for multi-dimensional PDEs with mixed partial derivatives.

2.1 1-Dimensional PDEs: Problem Formulation

Initially, we will consider the numerical solution of the general one-dimensional terminal value PDE problem

$$\frac{\partial V}{\partial t} + \mathcal{L}V = 0, \tag{2.1}$$

where $\mathcal{L}$ is the operator

$$\mathcal{L} = \mu(t,x)\frac{\partial}{\partial x} + \frac{1}{2}\sigma(t,x)^2\frac{\partial^2}{\partial x^2} - r(t,x),$$

and where $V = V(t,x)$ satisfies a terminal condition $V(T,x) = g(x)$. We recognize the PDE as being an extension of the Black-Scholes PDE (1.47) to general time- and state-dependent drift (μ), volatility (σ), and interest rate (r). Underneath the PDE lies a physical model where a state variable process $x(\cdot)$ follows an SDE of the form

$$dx(t) = \mu\left(t, x(t)\right)\, dt + \sigma\left(t, x(t)\right)\, dW(t) \tag{2.2}$$

where $W(t)$ is a Brownian motion in the risk-neutral probability measure Q. Let the range of values attainable by $x(t)$ on $t \in [0,T]$ be denoted $\mathcal{B} \subseteq \mathbb{R}$, and assume that the functions $\mu, \sigma, r : [0,T) \times \mathcal{B} \to \mathbb{R}$ are sufficiently regular to make (2.1) and (2.2) meaningful (see Chapter 1).

The terminal value problem above is, as discussed earlier, a *Cauchy problem* to be solved for $V(t,x)$ on $(t,x) \in [0,T) \times \mathcal{B}$. In many cases of practical interest, further boundary conditions are applied in the spatial (x) domain. If such boundary conditions are expressed directly in terms of V (rather than its derivatives) we have a *Dirichlet boundary problem*. For instance, a so-called *up-and-out barrier option* will pay out $g(x(T))$ at time T if and only if $x(t)$ stays strictly below a contractually specified barrier level H at all times $t \leq T$. If, on the other hand, $x(t)$ touches H at any time during the life of the contract, it will expire worthless (or "knock out"). In this case, the PDE is only to be solved on $(t,x) \in [0,T) \times (\mathcal{B} \cap (-\infty, H))$ and is subject to the Dirichlet boundary condition

$$V(t,H) = 0, \quad t \in [0,T],$$

which expresses that the option has no value for $x \geq H$. We note that it is not uncommon to encounter options where the spatial domain boundaries are functions of time, a situation we shall deal with in Section 2.7.1. Also, as we shall see shortly, sometimes boundary conditions are conveniently expressed in terms of derivatives of V.

For numerical solution of the PDE (2.1), we often need to assume that the domain of the state variable x is finite, even in situations where (2.1) is supposed to hold for an infinite domain. Suitable truncation of the domain can often be done probabilistically, based on a confidence interval for $x(T)$. To illustrate the procedure, consider the Black-Scholes PDE (1.47) applied to a call option with strike K. A common first step is to use the transformation $x = \ln S$, such that the PDE has constant coefficients,

$$\frac{\partial V}{\partial t} + \left(r - \frac{1}{2}\sigma^2\right)\frac{\partial V}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 V}{\partial x^2} - rV = 0, \tag{2.3}$$

with terminal value (for a call option) $V(T,x) = (e^x - K)^+$. The domain of x is here the entire real line, $\mathcal{B} = \mathbb{R}$. We know (from (1.39)) that

$$x(T) = x(0) + \left(r - \frac{1}{2}\sigma^2\right) T + \sigma\left(W(T) - W(0)\right), \tag{2.4}$$

which is a Gaussian random variable with mean $\overline{x} = x(0) + (r - \frac{1}{2}\sigma^2)T$ and variance $\sigma^2 T$. Consider now replacing the domain $(-\infty, \infty)$ with the finite interval $[\overline{x} - \alpha\sigma\sqrt{T}, \overline{x} + \alpha\sigma\sqrt{T}]$ for some positive constant α. The likelihood of $x(T)$ falling outside of this interval is easily seen to be $2\Phi(-\alpha)$ (where, as always, $\Phi(z)$ is the standard Gaussian cumulative distribution function). If, say, we set α to 4, $2\Phi(-\alpha) = 6.3 \times 10^{-5}$, which is an insignificant probability for most applications. Larger (smaller) values of α will make the truncation error smaller (larger) and will ultimately require more (less) effort in a numerical scheme. We recommend values of α somewhere between 3 and 5 for most applications. For the Black-Scholes case, a rigorous estimate of the error imposed by domain truncation is given in Kangro and Nicolaides [2000].

In many cases of practical interest, it is not possible to write down an exact confidence interval for $x(T)$. In such cases, one instead may use an approximate confidence interval, found by, for instance, using "average" values for $\mu(t,x)$ and $\sigma(t,x)$. High precision in these estimates is typically not needed.

2.2 Finite Difference Discretization

In order to solve the PDE (2.1) numerically, we now wish to discretize it on the rectangular domain $(t,x) \in [0,T] \times [\underline{M}, \overline{M}]$, where $\overline{M}$ and $\underline{M}$ are finite constants, possibly found by a truncation procedure such as the one outlined above. We first introduce two equidistant[1] grids $\{t_i\}_{i=0}^n$ and $\{x_j\}_{j=0}^{m+1}$ where $t_i = iT/n \triangleq i\Delta_t$, $i = 0, 1, \ldots, n$, and $x_j = \underline{M} + j(\overline{M} - \underline{M})/(m+1) \triangleq \underline{M} + j\Delta_x$, $j = 0, 1, \ldots, m+1$. The terminal value $V(T,x) = g(x)$ is imposed at $t_n = T$, and spatial boundary conditions are imposed at x_0 and x_{m+1}.

2.2.1 Discretization in x-Direction. Dirichlet Boundary Conditions

We first focus on the spatial operator $\mathcal{L}$ and restrict x to take values in the interior of the spatial grid $x \in \{x_j\}_{j=1}^m$. Consider replacing the first- and second-order partial derivatives with first- and second-order difference operators:

[1] Non-equidistant grids are often required in practice and will be covered in Section 2.4.

$$\delta_x V(t, x_j) \triangleq \frac{V(t, x_{j+1}) - V(t, x_{j-1})}{2\Delta_x}, \tag{2.5}$$

$$\delta_{xx} V(t, x_j) \triangleq \frac{V(t, x_{j+1}) + V(t, x_{j-1}) - 2V(t, x_j)}{\Delta_x^2}. \tag{2.6}$$

These operators are accurate to second order. Formally[2],

Lemma 2.2.1.

$$\delta_x V(t, x_j) = \frac{\partial V(t, x_j)}{\partial x} + O\left(\Delta_x^2\right),$$

$$\delta_{xx} V(t, x_j) = \frac{\partial^2 V(t, x_j)}{\partial x^2} + O\left(\Delta_x^2\right).$$

Proof. A Taylor expansion of $V(t, x)$ around the point $x = x_j$ gives

$$\begin{aligned} V(t, x_{j+1}) = V(t, x_j) &+ \Delta_x \frac{\partial V(t, x_j)}{\partial x} \\ &+ \frac{1}{2}\Delta_x^2 \frac{\partial^2 V(t, x_j)}{\partial x^2} + \frac{1}{6}\Delta_x^3 \frac{\partial^3 V(t, x_j)}{\partial x^3} + O\left(\Delta_x^4\right), \end{aligned}$$

and

$$\begin{aligned} V(t, x_{j-1}) = V(t, x_j) &- \Delta_x \frac{\partial V(t, x_j)}{\partial x} \\ &+ \frac{1}{2}\Delta_x^2 \frac{\partial^2 V(t, x_j)}{\partial x^2} - \frac{1}{6}\Delta_x^3 \frac{\partial^3 V(t, x_j)}{\partial x^3} + O\left(\Delta_x^4\right). \end{aligned}$$

Insertion of these expressions into (2.5) and (2.6) gives the desired result. □

In other words, if we introduce the discrete operator

$$\widehat{\mathcal{L}} = \mu(t, x)\delta_x + \frac{1}{2}\sigma(t, x)^2 \delta_{xx} - r(t, x),$$

we have, for $x \in \{x_j\}_{j=1}^m$,

$$\mathcal{L}V(t, x) = \widehat{\mathcal{L}}V(t, x) + O\left(\Delta_x^2\right).$$

With attention restricted to values on the grid $\{x_j\}_{j=1}^m$, we can view $\widehat{\mathcal{L}}$ as a matrix, once we specify the side boundary conditions at x_0 and x_{m+1}. For the Dirichlet case, assume for instance that

$$V(x_0, t) = \underline{f}(t, x_0), \quad V(x_{m+1}, t) = \overline{f}(t, x_{m+1}),$$

[2]Recall that a function $f(h)$ is of order $O(e(h))$ if $|f(h)|/|e(h)|$ is bounded from above by a positive constant in the limit $h \to 0$.

for given functions $\underline{f}, \overline{f} : [0,T] \times \mathbb{R} \to \mathbb{R}$. With[3] $\mathbf{V}(t) \triangleq (V(t,x_1),\ldots,V(t,x_m))^\top$ and, for $j = 1,\ldots,m$,

$$c_j(t) \triangleq -\sigma(t,x_j)^2\Delta_x^{-2} - r(t,x_j), \tag{2.7}$$

$$u_j(t) \triangleq \frac{1}{2}\mu(t,x_j)\Delta_x^{-1} + \frac{1}{2}\sigma(t,x_j)^2\Delta_x^{-2}, \tag{2.8}$$

$$l_j(t) \triangleq -\frac{1}{2}\mu(t,x_j)\Delta_x^{-1} + \frac{1}{2}\sigma(t,x_j)^2\Delta_x^{-2}, \tag{2.9}$$

we can write

$$\widehat{\mathcal{L}}\mathbf{V}(t) = \mathbf{A}(t)\mathbf{V}(t) + \mathbf{\Omega}(t), \tag{2.10}$$

where $\mathbf{A}$ is a *tri-diagonal matrix*

$$\mathbf{A}(t) = \begin{pmatrix} c_1(t) & u_1(t) & 0 & 0 & 0 & \ldots & 0 \\ l_2(t) & c_2(t) & u_2(t) & 0 & 0 & \ldots & 0 \\ 0 & l_3(t) & c_3(t) & u_3(t) & 0 & \ldots & 0 \\ 0 & 0 & l_4(t) & c_4(t) & u_4(t) & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & l_{m-1}(t) & c_{m-1}(t) & u_{m-1}(t) \\ 0 & 0 & 0 & 0 & 0 & l_m(t) & c_m(t) \end{pmatrix} \tag{2.11}$$

and $\mathbf{\Omega}(t)$ is a vector containing boundary values

$$\mathbf{\Omega}(t) = \begin{pmatrix} l_1(t)\underline{f}(t,x_0) \\ 0 \\ \vdots \\ 0 \\ u_m(t)\overline{f}(t,x_{m+1}) \end{pmatrix}.$$

As discussed earlier, sometimes one or both of the functions $\overline{f}$ and $\underline{f}$ are explicitly imposed as part of the option specification (as is the case for a knock-out options). In other cases, asymptotics may be necessary to establish these functions. For instance, for the case of a simple call option on a stock paying no dividends, we can set

$$\overline{f}(t,x) = e^x - Ke^{-r(T-t)},$$
$$\underline{f}(t,x) = 0,$$

where we, as before, have set $x = \ln S$ (S being the stock price) and assumed that the strike K is positive. The result for $\underline{f}$ is obvious; the result for $\overline{f}$ follows from the fact that a deep in-the-money call option will almost certainly pay at maturity the stock (the present value of which is just $S = e^x$) minus the strike (the present value of which is $Ke^{-r(T-t)}$).

[3]For clarity, this chapter uses boldface type for all vectors and matrices.

2.2.2 Other Boundary Conditions

Deriving asymptotic Dirichlet conditions can be quite involved for complicated option payouts and is often inconvenient in implementations. Rather than having to perform an asymptotic analysis for each and every type of option payout, it would be preferable to have a general-purpose mechanism for specifying the boundary condition. One common idea involves making assumptions on the form of the functional dependency between V and x at the grid boundaries, often from specification of relationships between spatial derivatives. For instance, if we impose the condition that the second derivative of V is zero at the upper boundary (x_{m+1}) — that is, V is a linear function of x — we can write (effectively using a downward discretization of the second derivative)

$$\frac{V(t,x_{m+1}) + V(t,x_{m-1}) - 2V(t,x_m)}{\Delta_x^2} = 0$$
$$\Rightarrow V(t,x_{m+1}) = 2V(t,x_m) - V(t,x_{m-1}).$$

A similar assumption at the lower spatial boundary yields

$$V(t,x_0) = 2V(t,x_1) - V(t,x_2).$$

For PDEs discretized in the logarithm of some asset, it may be more natural to assume that $V(t,x) \propto e^x$ at the boundaries; equivalently, we can assume that $\partial V/\partial x = \partial^2 V/\partial x^2$ at the boundary. When discretized in downward fashion at the upper boundary (x_{m+1}), this implies that

$$\frac{V(t,x_{m+1}) - V(t,x_m)}{\Delta_x} = \frac{V(t,x_{m+1}) + V(t,x_{m-1}) - 2V(t,x_m)}{\Delta_x^2}$$

or (assuming that $\Delta_x \neq 1$)

$$V(t,x_{m+1}) = V(t,x_{m-1})\frac{1}{\Delta_x - 1} + V(t,x_m)\frac{\Delta_x - 2}{\Delta_x - 1}.$$

Similarly,

$$V(t,x_0) = V(t,x_1)\frac{2+\Delta_x}{1+\Delta_x} - V(t,x_2)\frac{1}{\Delta_x + 1}.$$

Common for both methods above — and for the Dirichlet specification discussed earlier — is that they give rise to boundary specifications through simple linear systems of the general form

$$V(t,x_{m+1}) = k_m(t)V(t,x_m) + k_{m-1}(t)V(t,x_{m-1}) + \overline{f}(t,x_{m+1}), \quad (2.12)$$
$$V(t,x_0) = k_1(t)V(t,x_1) + k_2(t)V(t,x_2) + \underline{f}(t,x_0). \quad (2.13)$$

This boundary specification can be captured in the matrix system (2.10) by simply rewriting a few components of $\mathbf{A}(t)$; specifically, we must set

$$
\begin{aligned}
c_m(t) &= -\sigma(t,x_m)^2\Delta_x^{-2} - r(t,x_m) + k_m(t)u_m(t),\\
l_m(t) &= -\frac{1}{2}\mu(t,x_m)\Delta_x^{-1} + \frac{1}{2}\sigma(t,x_m)^2\Delta_x^{-2} + k_{m-1}(t)u_m(t),\\
c_1(t) &= -\sigma(t,x_j)^2\Delta_x^{-2} - r(t,x_j) + k_1(t)l_1(t),\\
u_1(t) &= \frac{1}{2}\mu(t,x_1)\Delta_x^{-1} + \frac{1}{2}\sigma(t,x_1)^2\Delta_x^{-2} + k_2(t)l_1(t).
\end{aligned}
$$

All other components of $\mathbf{A}$ remain as in (2.11); note that $\mathbf{A}$ remains tri-diagonal.

An alternative approach to specification of boundary conditions in the x-domain involves using the PDE itself to determine the boundary conditions, through replacement of all central difference operators with one-sided differences at the boundaries. Section 10.1.5.2 contains a detailed example of this idea; ultimately, this approach leads to boundary conditions that can also be written in the form (2.12)–(2.13).

2.2.3 Time-Discretization

To simplify notation, assume for now that $\mathbf{\Omega}(t) = 0$ for all t, as will be the case if, say, we use the linear or linear-exponential boundary conditions outlined earlier. On the spatial grid, our original PDE can be written

$$\frac{\partial \mathbf{V}(t)}{\partial t} = -\mathbf{A}(t)\mathbf{V}(t) + O\left(\Delta_x^2\right)$$

which, ignoring the error term[4], defines a system of coupled ordinary differential equations (ODEs).

A number of methods are available for the numerical solution of coupled ODEs; see, e.g., Press et al. [1992]. We here only consider basic two-level time-stepping schemes, where grid computations at time t_i involve only PDE values at times t_i and t_{i+1}. Focusing the attention on a particular bucket $[t_i, t_{i+1}]$, the choice for the finite difference approximation of $\partial V/\partial t$ is obvious:

$$\frac{\partial \mathbf{V}}{\partial t} \approx \frac{\mathbf{V}(t_{i+1}) - \mathbf{V}(t_i)}{\Delta_t}.$$

Not so obvious, however, is to which time in the interval $[t_i, t_{i+1}]$ we should associate this derivative. To be general, consider picking a time $t_i^{i+1}(\theta) \in [t_i, t_{i+1}]$, given by

$$t_i^{i+1}(\theta) = (1-\theta)t_{i+1} + \theta t_i, \tag{2.14}$$

where $\theta \in [0,1]$ is a parameter. We then write

$$\frac{\partial \mathbf{V}\left(t_i^{i+1}(\theta)\right)}{\partial t} \approx \frac{\mathbf{V}(t_{i+1}) - \mathbf{V}(t_i)}{\Delta_t}.$$

[4]Note that the error term $O(\Delta_x^2)$ is here to be interpreted as an m-dimensional vector. We will use such short-hand notation throughout this chapter.

By a Taylor expansion, it is easy to see that this expression is first-order accurate in the time step when $\theta \neq \frac{1}{2}$, and second-order accurate when $\theta = \frac{1}{2}$. Written compactly,

$$\frac{\partial \mathbf{V}\left(t_i^{i+1}(\theta)\right)}{\partial t} = \frac{\mathbf{V}(t_{i+1}) - \mathbf{V}(t_i)}{\Delta_t} + 1_{\{\theta \neq \frac{1}{2}\}} O\left(\Delta_t\right) + O\left(\Delta_t^2\right). \tag{2.15}$$

This result on the convergence order is intuitive since only in the case $\theta = \frac{1}{2}$ is the difference coefficient precisely central; for all other cases, the difference coefficient is either predominantly backward in time or predominantly forward in time.

The time-discretization technique introduced above is known as a *theta scheme*. The special cases of $\theta = 1$, $\theta = 0$, and $\theta = \frac{1}{2}$ are known as the *fully implicit scheme*, the *fully explicit scheme*, and the *Crank-Nicolson scheme*, respectively. In light of the convergence result (2.15), one may wonder why anything other than the Crank-Nicolson scheme is ever used. The CN method is, indeed, often the method of choice, but there are situations where a straight application of the Crank-Nicolson scheme can lead to oscillations in the numerical solution or its spatial derivatives. Judicial application of the fully implicit method can often alleviate these problems, as we shall discuss later. The fully explicit method should never be used due to poor convergence and stability properties (see Section 2.3), but has nevertheless managed to survive in a surprisingly large number of finance texts and papers.

2.2.4 Finite Difference Scheme

We now proceed to combine the discretizations (2.10) and (2.15) into a complete finite difference scheme. First, we expand

$$\begin{aligned}\mathbf{A}\left(t_i^{i+1}(\theta)\right)\mathbf{V}\left(t_i^{i+1}(\theta)\right) &= \theta \mathbf{A}\left(t_i^{i+1}(\theta)\right)\mathbf{V}(t_i) \\ &\quad + (1-\theta)\,\mathbf{A}\left(t_i^{i+1}(\theta)\right)\mathbf{V}(t_{i+1}) + 1_{\{\theta \neq \frac{1}{2}\}} O(\Delta_t) + O\left(\Delta_t^2\right),\end{aligned}$$

such that our PDE can be represented as

$$\begin{aligned}\frac{\mathbf{V}(t_{i+1}) - \mathbf{V}(t_i)}{\Delta_t} &+ 1_{\{\theta \neq \frac{1}{2}\}} O(\Delta_t) + O\left(\Delta_t^2\right) \\ &= -\,\mathbf{A}\left(t_i^{i+1}(\theta)\right)\mathbf{V}\left(t_i^{i+1}(\theta)\right) + O\left(\Delta_x^2\right) \\ &= -\,\theta \mathbf{A}\left(t_i^{i+1}(\theta)\right)\mathbf{V}(t_i) - (1-\theta)\,\mathbf{A}\left(t_i^{i+1}(\theta)\right)\mathbf{V}(t_{i+1}) \\ &\quad + 1_{\{\theta \neq \frac{1}{2}\}} O(\Delta_t) + O\left(\Delta_t^2\right) + O\left(\Delta_x^2\right).\end{aligned}$$

Multiplying through with Δ_t gives rise to the complete finite difference representation of the PDE solution at times t_i and t_{i+1}:

Proposition 2.2.2. *On the grid $\{x_j\}_{j=1}^m$, the solution to (2.1) at times t_i and t_{i+1} is characterized by*

$$\left(\mathbf{I}-\theta\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right)\mathbf{V}(t_i) = \left(\mathbf{I}+(1-\theta)\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right)\mathbf{V}(t_{i+1}) + e_i^{i+1}, \tag{2.16}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix, and e_i^{i+1} is an error term

$$e_i^{i+1} = \Delta_t O\left(\Delta_x^2\right) + 1_{\left\{\theta\neq\frac{1}{2}\right\}} O\left(\Delta_t^2\right) + O\left(\Delta_t^3\right). \tag{2.17}$$

Let $\widehat{V}(t_i, x_j)$ denote the approximation to the true solution $V(t_i, x_j)$ obtained by using (2.16) without the error term. Defining

$$\widehat{\mathbf{V}}(t) = \left(\widehat{V}(t, x_1), \ldots, \widehat{V}(t, x_m)\right)^\top,$$

we have

$$\left(\mathbf{I}-\theta\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right)\widehat{\mathbf{V}}(t_i) = \left(\mathbf{I}+(1-\theta)\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right)\widehat{\mathbf{V}}(t_{i+1}). \tag{2.18}$$

For a known value of $\widehat{\mathbf{V}}(t_{i+1})$, (2.18) defines a simple linear system of equations that can be solved for $\widehat{\mathbf{V}}(t_i)$ by standard methods. Simplifying matters is the fact that the matrix $(\mathbf{I}-\theta\Delta_t\mathbf{A}(t_i^{i+1}(\theta)))$ is tri-diagonal, allowing us to solve (2.18) in only $O(m)$ operations; see Press et al. [1992] for an algorithm[5].

Starting from the prescribed terminal condition $V(t_n, x_j) = g(x_j)$, $j = 1, \ldots, m$, we can now use (2.18) to iteratively step backward in time until we ultimately recover $\widehat{\mathbf{V}}(0)$. This procedure is known as *backward induction*.

Proposition 2.2.3. *The theta scheme (2.18) recovers $\widehat{\mathbf{V}}(0)$ in $O(mn)$ operations. If the scheme converges, the error on $\widehat{\mathbf{V}}(0)$ compared to the exact solution $\mathbf{V}(0)$ is of order*

$$O\left(\Delta_x^2\right) + 1_{\left\{\theta\neq\frac{1}{2}\right\}} O\left(\Delta_t\right) + O\left(\Delta_t^2\right).$$

Proof. The backward induction algorithm requires the solution of n tri-diagonal systems, one per time step, for a total computational cost of $O(mn)$. The local truncation error on $\widehat{\mathbf{V}}(t_i)$ is e_i^{i+1}, making the global truncation error after n time steps of order ne_i^{i+1}. Combining (2.17) with the fact that $n = T/\Delta_t = O(\Delta_t^{-1})$ gives the order result listed in the proposition. □

[5]The special case of an explicit scheme ($\theta = 0$) provides us with a direct expression for $V(t_i, x_j)$ in terms of $V(t_{i+1}, x_{j-1})$, $V(t_{i+1}, x_j)$, and $V(t_{i+1}, x_{j+1})$, a scheme that is easily visualized as a "trinomial tree". The intuitive nature of the explicit scheme coupled with the fact that no matrix equation must be solved may explain the popularity of this scheme in the finance literature, despite its poor numerical qualities (see Section 2.3). We stress that the workload of the explicit scheme is still $O(m)$ per time step, as is the case for all theta schemes.

It follows from Proposition 2.2.3 that the Crank-Nicolson scheme is second-order convergent in the time step, and all other theta schemes are first-order convergent in the time step. All theta-schemes are second-order convergent in the spatial step Δ_x.

In deriving (2.18), we assumed earlier that the boundary vector was zero, $\mathbf{\Omega}(t) = 0$. Including a non-zero boundary vector into the scheme is, however, straightforward and results in a time-stepping scheme of the form

$$\left(\mathbf{I}-\theta\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right)\widehat{\mathbf{V}}(t_i) = \left(\mathbf{I}+(1-\theta)\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right)\widehat{\mathbf{V}}(t_{i+1}) + (1-\theta)\mathbf{\Omega}(t_{i+1}) + \theta\mathbf{\Omega}(t_i). \quad (2.19)$$

Again, this system is easily solved for $\widehat{\mathbf{V}}(t_i)$ by a standard tri-diagonal equation solver.

As a final point, we stress that the finite difference scheme above ultimately yields a full vector of values $\widehat{\mathbf{V}}(0)$ at time 0, with one element per value of x_j, $j = 1, \ldots, m$. In general, we are mainly interested in $V(0, x(0))$, where $x(0)$ is the known value of x at time 0. There is no need to include $x(0)$ in the grid, as we can simply employ an interpolator (e.g., a cubic spline) on this vector $\widehat{\mathbf{V}}(0)$ to compute $V(0, x(0))$. Clearly, such an interpolator should be at least second-order accurate to avoid interfering with the overall $O(\Delta_x^2)$ convergence of the finite difference scheme. Assuming the interpolator is sufficiently smooth, we can also use it to compute various partial derivatives with respect to x that we may be interested in. Alternatively, these can be computed by the same type of finite difference coefficients discussed in Section 2.2.1. The derivative $\partial V(0, x(0))/\partial t$ — the *time decay* — can be picked up from the grid in the same fashion.

Remark 2.2.4. The scheme (2.18) may, without affecting convergence order, be replaced with

$$\left(\mathbf{I}-\theta\Delta_t\mathbf{A}(t_i)\right)\widehat{\mathbf{V}}(t_i) = \left(\mathbf{I}+(1-\theta)\Delta_t\mathbf{A}\left(t_{i+1}\right)\right)\widehat{\mathbf{V}}(t_{i+1}).$$

2.3 Stability

2.3.1 Matrix Methods

Ignoring the contributions from boundary conditions, the finite difference scheme developed in the previous section can be rewritten

$$\widehat{\mathbf{V}}(t_i) = \mathbf{B}_i^{i+1}\widehat{\mathbf{V}}(t_{i+1}), \quad (2.20)$$

where

$$\mathbf{B}_i^{i+1} \triangleq \left(\mathbf{I}-\theta\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right)^{-1}\left(\mathbf{I}+(1-\theta)\Delta_t\mathbf{A}\left(t_i^{i+1}(\theta)\right)\right).$$

That is, for any $0 \leq k < n$,

$$\widehat{\mathbf{V}}(t_k) = \mathbf{B}_k^n \widehat{\mathbf{V}}(t_n), \ \mathbf{B}_k^n \triangleq \mathbf{B}_k^{k+1} \mathbf{B}_{k+1}^{k+2} \ldots \mathbf{B}_{n-1}^n.$$

We say that the scheme is *stable* if $|\widehat{\mathbf{V}}(t_k)|$ is bounded for all $0 \leq k < n$. Assuming $|\widehat{\mathbf{V}}(T)| < \infty$, a necessary and sufficient condition for stability is that there exists a constant K such that for all $0 \leq k < n$

$$|\mathbf{B}_k^n| \leq K, \tag{2.21}$$

where $|\cdot|$ is any matrix norm, e.g. the spectral norm or the infinity norm[6]. See Mitchell and Griffiths [1980] for further details.

2.3.2 Von Neumann Analysis

For simple problems with time- and space-independent coefficients, it may be possible to establish the spectral norm of $\mathbf{B}_k^n$ by direct methods (see e.g. Mitchell and Griffiths [1980], Kraaijevanger et al. [1987], Lenferink and Spijker [1991], Spijker and Straetemans [1997]), but generally the stability criterion (2.21) is difficult to evaluate. While certain somewhat simpler matrix-based methods exist to establish necessary conditions for stability (again, see Mitchell and Griffiths [1980]), we shall here only consider a "local" method, known as the *von Neumann method*. In principle, the von Neumann method only holds for finite difference schemes where the underlying PDE has constant coefficients, but there is much numerical evidence to support wider application[7]. The von Neumann method does not directly consider the effect of boundary conditions on stability, but (for constant coefficient problems) provides a necessary condition for stability irrespective of the type of boundary condition.

The basis for the von Neumann analysis is the observation that a real function sampled on a finite number of points is uniquely defined by a complex Fourier series. For our PDE solution sampled on the spatial grid, the precise result is

$$V(t_k, x_j) = \sum_l H_l(t_k) e^{-i\omega_l j \Delta_x},$$

where $H_l(t_k)$ and ω_l are the amplification factor (discrete Fourier transform) and wave number for the l-th mode, respectively. Notice that i here denotes

[6]The spectral norm of a matrix $\mathbf{C}$ is defined as the largest absolute eigenvalue of $(\mathbf{C}^\top \mathbf{C})^{1/2}$. The infinity norm is defined as $\max_i \sum_j |C_{i,j}|$.

[7]In the application to PDEs with non-constant coefficients, it may help to think of the von Neumann analysis as being applied to the PDE locally with "frozen" coefficients, followed by an examination of the worst case among all frozen coefficients.

the imaginary unit, $i^2 = -1$, with k (momentarily) having taken the role of the time index in the finite difference grid. For the constant coefficient case, a key fact for our PDE problem is that

$$H_l(t_k) = H_l(t_{k+1})\xi_l^{-1},$$

where ξ_l is a mode-specific *amplification factor* independent of time. To determine how a solution is propagated back through the finite difference grid, it thus suffices to consider a test function of the form

$$v(t_k, x_j) = \xi(\omega)^{n-k} e^{i\omega j \Delta_x}. \tag{2.22}$$

According to the Von Neumann criterion, stability of (2.20) requires that the *modulus of the amplification factor* $\xi(\omega)$ *is less or equal to one*, independent of the wave number:

$$\forall \omega : |\xi(\omega)| \leq 1. \tag{2.23}$$

This criterion is natural and merely expresses that all eigenmodes should be dampened, and not exponentially amplified, by the finite difference scheme.

Turning to our system (2.20), assume for simplicity that $r(t,x) = 0$. A positive interest rate (we will nearly always have $r(t,x) > 0$) introduces some extra dampening through discounting effects and will, if anything, lead to better stability properties than the case of zero interest rates. Writing $v(t_k, x_j) = v_{k,j}$, $\sigma(t_k^{k+1}(\theta), x_j) = \sigma_{k,j}$, and $\mu(t_k^{k+1}(\theta), x_j) = \mu_{k,j}$, the von Neumann analysis gives the following result:

Proposition 2.3.1. *Define* $\alpha = \Delta_t/(\Delta_x)^2$. *For (2.20) with* $r(t,x) = 0$, *the von Neumann stability criterion is*

$$1 \geq \theta \geq \frac{1}{2} - \frac{1}{\alpha}\left(\frac{\sigma_{k,j}^2}{\sigma_{k,j}^4 + \mu_{k,j}^2\Delta_x^2 + \left|\mu_{k,j}^2\Delta_x^2 - \sigma_{k,j}^4\right|}\right), \tag{2.24}$$

to hold for all $k = 0, 1, \ldots, n-1$, $j = 1, 2, \ldots, m$.

Proof. Define $\varsigma_{k,j}^{\pm} = \sigma_{k,j}^2 \pm \Delta_x \mu_{k,j}$. A local application of (2.20) gives

$$v_{k,j-1}\left(-\frac{\alpha\theta}{2}\varsigma_{k,j}^-\right) + v_{k,j}(1 + \alpha\theta\sigma_{k,j}^2) + v_{k,j+1}\left(-\frac{\alpha\theta}{2}\varsigma_{k,j}^+\right) =$$
$$v_{k+1,j}\left(\frac{\alpha(1-\theta)}{2}\varsigma_{k,j}^-\right) + v_{k+1,j}(1 - \alpha(1-\theta)\sigma_{k,j}^2) + v_{k+1,j+1}\left(\frac{\alpha(1-\theta)}{2}\varsigma_{k,j}^+\right)$$

with α defined above. Inserting (2.22) and rearranging (using Euler's formulas for sin and cos) yields

$$\xi(\omega) = \frac{1 - (1-\theta)\alpha\sigma_{k,j}^2(1 - \cos\omega\Delta_x) + i(1-\theta)\alpha\Delta_x\,\mu_{k,j}\sin\omega\Delta_x}{1 + \theta\alpha\sigma_{k,j}^2(1 - \cos\omega\Delta_x) - i\theta\alpha\Delta_x\,\mu_{k,j}\sin\omega\Delta_x}.$$

Note that ξ is a function of k and j, due to the non-constant PDE parameters. As discussed earlier (see also Mitchell and Griffiths [1980]), we expect the system to be stable if the criterion (2.23) holds for all k and j in the grid. Computing the modulus of ξ and requiring that it does not exceed one leads, after straightforward manipulations, to the stability criterion

$$\forall\omega : 2\alpha\sigma_{k,j}^2 + (2\theta - 1)\alpha^2 \left[\sigma_{k,j}^4 + \mu_{k,j}^2\Delta_x^2 + \cos\omega\Delta_x \left(\mu_{k,j}^2\Delta_x^2 - \sigma_{k,j}^4\right)\right] \geq 0.$$

As $\cos\omega\Delta_x \in [-1, 1]$, this expression can be simplified to (2.24). □

From (2.24) we can immediately conclude that the finite difference scheme is always stable if $\frac{1}{2} \leq \theta \leq 1$, irrespective of the magnitudes of Δ_x and Δ_t. For $\frac{1}{2} \leq \theta \leq 1$, we therefore say that the theta scheme is *absolutely stable*, or simply *A-stable*. Both the fully implicit ($\theta = 1$) and the Crank-Nicolson ($\theta = \frac{1}{2}$) finite difference schemes are thus A-stable. For the explicit scheme ($\theta = 0$), however, stability is *conditional*, requiring

$$\frac{2}{\alpha}\sigma_{k,j}^2 \geq \sigma_{k,j}^4 + \mu_{k,j}^2\Delta_x^2 + \left|\mu_{k,j}^2\Delta_x^2 - \sigma_{k,j}^4\right|.$$

For small drifts, this expression amounts to the restriction $\sigma_{k,j}^2 \leq \Delta_x^2/\Delta_t$ which can be quite onerous, often requiring the (laborious) use of thousands of time steps in the finite difference grid. We shall not consider fully explicit methods any further in this book.

Returning to the case $\frac{1}{2} \leq \theta \leq 1$, let us introduce a stronger definition of stability. A time-stepping method is said to be *strongly A-stable* if the modulus of the amplification factor ξ is strictly below 1 for any value of the time step, including the limit[8] $\Delta_t \to \infty$. From (2.24), we see that if $\Delta_t \to \infty$ (which implies $\alpha \to \infty$), then the modulus of the amplification factor could reach 1 in the special case of $\theta = 1/2$. In other words, the Crank-Nicolson scheme is *not* strongly A-stable. For large time steps, harmonics in the Crank-Nicolson finite difference solution will effectively not be dampened from one time step to the next, opening up the possibility that unwanted high-frequency oscillations can creep into the numerical solution. In practice, this is primarily a problem if high-frequency eigenmodes have high amplification factors, as can happen if there is an outright discontinuity in the terminal value function g. The problem is especially noticeable if the discontinuity in the value function is "close" in both time and space to $t = 0$ and $x = x(0)$ (as would be the case for a short-dated option with a discontinuity close to the starting value of x). Oscillations can be prevented by setting the time step smaller than twice the maximum stable explicit time step (see Tavella and Randall [2000]), but this can often be computationally expensive. We shall deal with other methods to suppress oscillations in Section 2.5.

We conclude this section by noting a deep connection between the stability of a finite difference scheme and its convergence to the true solution

[8] If further $|\xi|$ approaches zero for $\Delta_t \to 0$, the scheme is said to be *L-stable*.

of the PDE as $\Delta_t \to 0$ and $\Delta_x \to 0$. First, we define a finite difference scheme to be *consistent* if local (Taylor) truncation errors approach zero for $\Delta_t \to 0$ and $\Delta_x \to 0$. All the schemes we have encountered so far are consistent. Further, define a finite difference scheme to be *convergent* if the difference between the numerical solution and the exact PDE solution at a fixed point in the domain converges to zero uniformly as $\Delta_t \to 0$ and $\Delta_x \to 0$ (not necessarily independently of each other). We then have

Theorem 2.3.2 (Lax Equivalence Theorem). *For a well-posed*[9] *linear terminal value PDE, a consistent 2-level finite difference scheme is convergent if and only if it is stable.*

A more precise statement of the above result, as well as a proof, can be found in Mitchell and Griffiths [1980].

2.4 Non-Equidistant Discretization

In practice, we often wish to align the finite difference grid to particular dates (e.g., those on which a coupon or a dividend is paid) and particular values of x (e.g., those on which strikes and barriers are positioned). Also, for numerical reasons we may want to make certain important parts of the finite difference grid more densely spaced to concentrate computational effort on domains of particular importance to the solution of the PDE. To do so, we will now relax our earlier assumption of equidistant discretization in time and space. Doing so for the time domain is actually trivial and merely requires us to replace Δ_t in (2.18) with $\Delta_{t,i} \triangleq t_{i+1} - t_i$, where the spacing of the time grid $\{t_i\}_{i=0}^n$ is now no longer constant. The backward induction algorithm can proceed as before. We note that the ability to freely select the time grid will allow us to line up perfectly with dates that carry high significance for the product in question (e.g. dates on which cash flows take place, see Section 2.7.3) or to, say, use coarser time steps for the part of the finite difference grid that is far in the future. For an adaptive algorithm to automatically select the time-step, see d'Halluin et al. [2001].

For the spatial step, we have a number of options to induce non-equidistant spacing. One method involves a non-linear change of variables $y = h(x)$ in the PDE, followed by a regular equidistant discretization in the new variable y. This maps into a non-equidistant discretization in x which, provided that $h(\cdot)$ is chosen carefully, will have the desired geometry. Discussion of this method along with guidelines for choosing $h(\cdot)$ can be found in Chapter 5 of Tavella and Randall [2000]. We will here pursue a more direct alternative, where we simply introduce an irregular grid $\{x_j\}_{j=0}^{m+1}$

[9] Well-posed means that the PDE we are solving has a unique solution that depends continuously on the problem data (PDE coefficients, domain, boundary conditions, etc.)

and redefine the finite difference operators (2.5)–(2.6) to achieve maximum precision. For this, define

$$\Delta^+_{x,j} \triangleq x_{j+1} - x_j, \quad \Delta^-_{x,j} \triangleq x_j - x_{j-1},$$

and set

$$\delta^+_x V(t,x_j) = \frac{V(t,x_{j+1}) - V(t,x_j)}{\Delta^+_{x,j}}, \quad \delta^-_x V(t,x_j) = \frac{V(t,x_j) - V(t,x_{j-1})}{\Delta^-_{x,j}}.$$

By a Taylor expansion, we get

$$\begin{aligned}\delta^+_x V(t,x_j) = \frac{\partial V(t,x_j)}{\partial x} &+ \frac{1}{2}\frac{\partial^2 V(t,x_j)}{\partial x^2}\Delta^+_{x,j} \\ &+ \frac{1}{6}\frac{\partial^3 V(t,x_j)}{\partial x^3}\left(\Delta^+_{x,j}\right)^2 + O\left(\left(\Delta^+_{x,j}\right)^3\right), \quad (2.25)\end{aligned}$$

$$\begin{aligned}\delta^-_x V(t,x_j) = \frac{\partial V(t,x_j)}{\partial x} &- \frac{1}{2}\frac{\partial^2 V(t,x_j)}{\partial x^2}\Delta^-_{x,j} \\ &+ \frac{1}{6}\frac{\partial^3 V(t,x_j)}{\partial x^3}\left(\Delta^-_{x,j}\right)^2 + O\left(\left(\Delta^-_{x,j}\right)^3\right). \quad (2.26)\end{aligned}$$

Maximum accuracy on the first-order derivative approximation is achieved by selecting a weighted combination of (2.25)–(2.26) such that the terms of order $O(\Delta^+_{x,j})$ and $O(\Delta^-_{x,j})$ cancel. That is, we set

$$\begin{aligned}\delta_x V(t,x_j) &= \frac{\Delta^-_{x,j}}{\Delta^+_{x,j} + \Delta^-_{x,j}} \cdot \delta^+_x V(t,x_j) + \frac{\Delta^+_{x,j}}{\Delta^+_{x,j} + \Delta^-_{x,j}} \cdot \delta^-_x V(t,x_j) \quad (2.27) \\ &= \frac{\partial V(t,x_j)}{\partial x} + O\left(\frac{\left(\Delta^+_{x,j}\right)^2 \Delta^-_{x,j} + \left(\Delta^-_{x,j}\right)^2 \Delta^+_{x,j}}{\Delta^+_{x,j} + \Delta^-_{x,j}}\right)\end{aligned}$$

which is second-order accurate, in the sense that reducing both $\Delta^+_{x,j}$ and $\Delta^+_{x,j}$ by a factor of k will reduce the error by a factor of k^2. To estimate the derivative $\partial^2 V(t,x_j)/\partial x^2$ we set

$$\begin{aligned}\delta_{xx} V(t,x_j) &= \frac{\delta^+_x V(t,x_j) - \delta^-_x V(t,x_j)}{\frac{1}{2}\left(\Delta^+_{x,j} + \Delta^-_{x,j}\right)} \quad (2.28) \\ &= \frac{\partial^2 V(t,x_j)}{\partial x^2} + O\left(\frac{\left(\Delta^+_{x,j}\right)^2 - \left(\Delta^-_{x,j}\right)^2}{\Delta^+_{x,j} + \Delta^-_{x,j}} + \frac{\left(\Delta^+_{x,j}\right)^3 + \left(\Delta^-_{x,j}\right)^3}{\Delta^+_{x,j} + \Delta^-_{x,j}}\right)\end{aligned}$$

which is only first-order accurate, unless $\Delta^+_{x,j} = \Delta^-_{x,j}$. Despite this, the global discretization error will typically remain second-order in the spatial step, even for a non-equidistant grid. A proof of this perhaps somewhat

surprising result can be found in the monograph Axelsson and Barker [1991] on finite element methods.

Development of a theta scheme around the definitions (2.27) and (2.28) proceeds in the same way as in Section 2.2. The resulting time-stepping scheme is identical to (2.18), after a modification of the matrix **A**. Specifically, we must simply redefine the c-, u-, and l-arrays in (2.7)–(2.9) as follows:

$$c_j(t) \triangleq \frac{\Delta^+_{x,j} - \Delta^-_{x,j}}{\Delta^+_{x,j}\Delta^-_{x,j}} - \frac{1}{\Delta^-_{x,j}\Delta^+_{x,j}}\sigma(t,x_j)^2 - r(t,x_j), \tag{2.29}$$

$$u_j(t) \triangleq \frac{\Delta^-_{x,j}}{\left(\Delta^+_{x,j} + \Delta^-_{x,j}\right)\Delta^+_{x,j}}\mu(t,x_j) + \frac{1}{\left(\Delta^+_{x,j} + \Delta^-_{x,j}\right)\Delta^+_{x,j}}\sigma(t,x_j)^2, \tag{2.30}$$

$$l_j(t) \triangleq -\frac{\Delta^+_{x,j}}{\left(\Delta^+_{x,j} + \Delta^-_{x,j}\right)\Delta^-_{x,j}}\mu(t,x_j) + \frac{1}{\left(\Delta^+_{x,j} + \Delta^-_{x,j}\right)\Delta^-_{x,j}}\sigma(t,x_j)^2. \tag{2.31}$$

For an example where having a non-equidistant grid is essential to the numerical performance of the scheme, see Section 9.4.3.

2.5 Smoothing and Continuity Correction

2.5.1 Crank-Nicolson Oscillation Remedies

As discussed earlier, for discontinuous terminal conditions, the Crank-Nicolson scheme may exhibit localized oscillations if the time step is too coarse relative to the spatial step. Depending on the timing and spatial position of the discontinuities, these spurious oscillations may negatively affect the computed option value or, more likely, its first ("delta") or second ("gamma") x-derivatives. Further, in the presence of discontinuous terminal conditions, the expected $O(\Delta_t^2)$ convergence order of the Crank-Nicolson scheme may not be realized. While $O(\Delta_t^2)$ convergence is possible without spurious oscillations in some multi-level time-stepping schemes, there is evidence that these schemes are less robust than the Crank-Nicolson scheme for many financially relevant problems, see, e.g., Windcliff et al. [2001]. Fortunately, it is relatively easy to remedy the problems in the Crank-Nicolson scheme. Specifically, a theoretical result by Rannacher [1984] shows that second-order convergence can be achieved for the Crank-Nicolson scheme, provided that two simple algorithm modifications are taken:

- The discontinuous terminal payout is least-squares (L^2) projected onto the space of linear Lagrange basis functions[10].

[10] Recall that the linear Lagrange basis functions (also called "hat" functions) are simply small triangles given by $l_j(x) = 1_{\{x_{j-1}<x\le x_j\}} \cdot \frac{x-x_{j-1}}{x_j-x_{j-1}} + 1_{\{x_j<x\le x_{j+1}\}} \cdot \frac{x_{j+1}-x}{x_{j+1}-x_j}$, $j = 1,\ldots,m$. For an algorithm to perform the L^2-projection, see Pooley et al. [2003].

- Two fully implicit time steps ($\theta = 1$) are taken before we switch to Crank-Nicolson ($\theta = \frac{1}{2}$) time stepping ("Rannacher stepping").

Both techniques effectively smoothen out the discontinuity before the Crank-Nicolson scheme is applied, dampening the problematic high-frequency modes of the numerical solution. As demonstrated in Pooley et al. [2003] (see also Giles and Carter [2006]), applying either technique in isolation will typically not suffice; both are jointly required to ensure smooth second-order convergence. That said, the application of Lagrange basis function projection may conveniently be substituted with simpler smoothing techniques, with no loss of convergence order. The usefulness of such payoff smoothing extends beyond the case of discontinuous boundary conditions, so we proceed to discuss a few common techniques next.

2.5.2 Continuity Correction

By the Shannon sampling theorem, (see Shannon [1949]) if the spectrum of $g(x)$ contains frequencies higher than $1/(2\Delta_x)$ (the *Nyquist frequency*), information is lost when we sample $g(x)$ on our mesh $\{x_j\}_{j=0}^{m+1}$. In other words, whenever $g(x)$ or its derivatives are non-smooth, we will incur a *quantization* error where important features of the payout (e.g., the discontinuity of the slope of a call option at the strike) will be lost between grid points. As the grid geometry is modified, and the location of critical points (strikes, barriers, etc.) relative to x-grid changes, the computed finite difference solution will jump back and forth in erratic fashion. This so-called *odd-even effect* will result in poor convergence and an undesirably strong dependence of the solution on the grid geometry.

One straightforward way to reduce the odd-even effect (and to smooth out the high-frequency components of the payoff) is to apply a common technique from probability theory known as a *continuity correction*. Here, we simply imagine that the value of g at a grid point x_j represents the average value of the function over the interval $[x_j - (x_j - x_{j-1})/2, x_j + (x_{j+1} - x_j)/2]$. In setting the terminal boundary value $V(T, x_j)$ we thus write

$$V(T, x_j) = \frac{1}{(x_{j+1} - x_{j-1})/2} \int_{x_j - (x_j - x_{j-1})/2}^{x_j + (x_{j+1} - x_j)/2} g(x)\, dx. \tag{2.32}$$

We note that this implies that $V(T, x_j) \neq g(x_j)$, unless g is linear in x. The application of continuity correction to parabolic PDE solvers was first proposed in Kreiss et al. [1970].

2.5.3 Grid Shifting

Consider the effect of using (2.32) on a *digital call option*, $g(x) = 1_{\{x>H\}}$, where the level H (the digital strike) is located between nodes x_k and

x_{k+1}. For nodes x_j, $j > k+1$, clearly $V(T, x_j) = 1$; for nodes x_j, $j < k$, $V(T, x_j) = 0$. The smoothing algorithm will have effect only at x_k or x_{k+1}, and will set either $V(T, x_k)$ or $V(T, x_{k+1})$ to a value somewhere between 0 and 1, depending on which of x_k or x_{k+1} is closest to H. If H happens to be exactly midway between x_k or x_{k+1}, the continuity correction is seen to have no effect whatsoever.

The digital option example above gives rise to a method listed in Tavella and Randall [2000] (see also Cheuk and Vorst [1996]). Here, we simply arrange the spatial grid such that the x-values where the payoff (or its derivatives) is discontinuous are exactly midway between grid nodes. If necessary, we can use a scheme with non-equidistant grid spacing to accomplish this (see Section 2.4). Our example above shows that aligning the grid in this way will, in a loose sense, make the payoff smooth.

For digital options, the grid shifting technique can be very efficient, and such "locking" of the location of strikes and barriers relative to the spatial grid can often reduce odd-even effects even better than the continuity correction discussed earlier. To demonstrate, consider the concrete task of using a finite difference grid to price a digital call option on a stock S in the Black-Scholes model. In this case, we conveniently have a theoretical option price to compare against, since it is easily shown that the time 0 value $V(0)$ must be

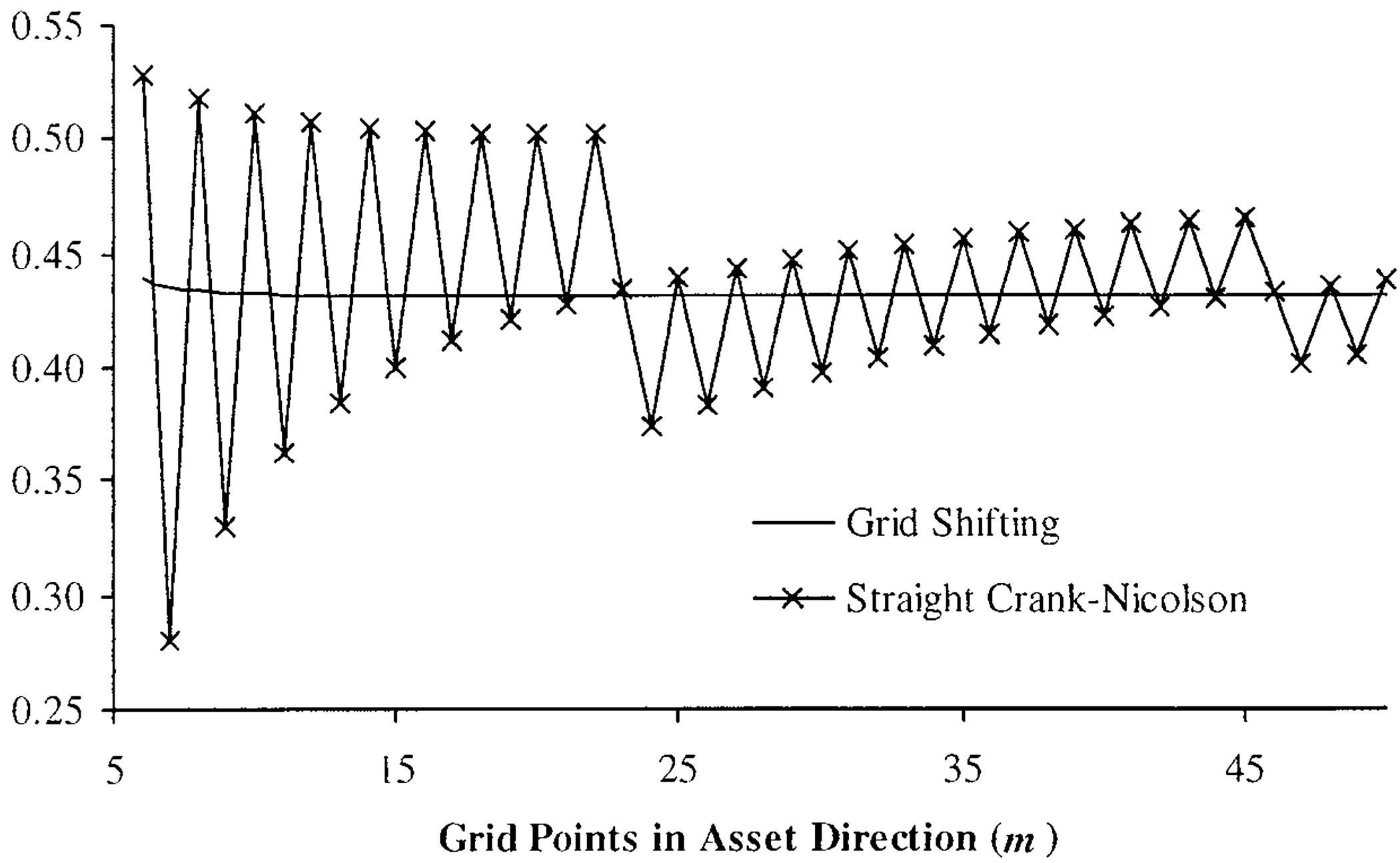
$$V(0) = e^{-rT} Q\left(S(T) > H\right) = e^{-rT} \Phi\left(\frac{\ln(S(0)/H) + (r - \sigma^2/2)T}{\sigma\sqrt{T}}\right). \tag{2.33}$$

For our numerical work, we discretize the asset equidistantly in log-space (i.e., we work with the PDE (2.3)) and determine the spatial grid boundaries by probabilistic means using a multiplier of $\alpha = 4.5$, see Section 2.1. Spatial boundary conditions are $\partial V/\partial x = \partial^2 V/\partial x^2$, implemented as described in Section 2.2.2. In one experiment, we apply a straight Crank-Nicolson approach, with no attempt to regularize the payoff condition. In a second experiment, we combine Crank-Nicolson with Rannacher stepping and also nudge the entire spatial grid upwards until the log-barrier $\ln(H)$ is located exactly half-way between two spatial grid points. Numerical results are shown in Figure 2.1.

As Figure 2.1 shows, a naive Crank-Nicolson implementation is plagued by severe odd-even effects and very slow convergence — 100's of spatial steps appear to be necessary before acceptable levels of the option price are reached. On the other hand, grid shifting combined with Rannacher stepping results in a perfectly smooth[11] convergence profile, and 5-digit price precision is here reached in less than 30 steps.

[11] It can be verified that the convergence order in m is, as expected, close to 2 in this experiment.

Fig. 2.1. 3 Year Digital Option Price



Notes: Finite difference estimates for the Black-Scholes price of a 3 year digital option with a strike of $H = 100$. The initial asset price is $S(0) = 100$, the interest rate is $r = 0$, and the volatility is $\sigma = 20\%$. Time stepping is performed with an equidistant grid containing $n = 50$ points. Spatial discretization in log-space is equidistant, as described in the main text; the number of grid points (m) is as listed on the x-axis of the figure. The "Straight Crank-Nicolson" graph shows the convergence profile for a pure Crank-Nicolson finite difference grid. The "Grid Shifting" graph shows the convergence profile for a Crank-Nicolson finite difference grid with Rannacher stepping and a shift of the spatial grid to center $\ln(H)$ midway between two grid points. From (2.33), the theoretical value of the option is 0.4312451.

2.6 Convection-Dominated PDEs

Recall from Section 2.3 that stability of the explicit finite difference scheme requires that (omitting grid subscripts on μ and σ)

$$\frac{2\Delta_x^2}{\Delta_t}\sigma^2 \geq \sigma^4 + \mu^2\Delta_x^2 + \left|\mu^2\Delta_x^2 - \sigma^4\right|.$$

As discussed, this condition can be violated if Δ_t is too large relative to Δ_x. However, for fixed Δ_t and Δ_x we notice that instability can also be triggered if the absolute value of the drift μ is raised to be sufficiently large relative to the diffusion coefficient σ.

While theta schemes with $\theta \geq 1/2$ are always stable, large drifts in the PDE can nevertheless cause spurious oscillations and an overall deterioration in numerical performance of these schemes. PDEs for which this effect

occurs are said to be *convection-dominated*. To quantify matters, assume for simplicity that the finite difference grid is equidistant in the x-direction, and consider the matrix $\mathbf{A}$ in (2.11) with tri-diagonal coefficients c, u, and l given by (2.7)–(2.9). As discussed in e.g. d'Halluin et al. [2005], spurious oscillations can occur when, for some t and some j, either $u_j(t) < 0$ or $l_j(t) < 0$. From (2.8) and (2.9), to avoid spurious oscillations we would thus need

$$\sigma(t, x_j)^2 \geq |\mu(t, x_j)| \Delta_x. \tag{2.34}$$

Intuitively, in convection-dominated systems, the central difference coefficient δ_x and δ_{xx} used to discretize the PDE can no longer fully contain the large expected up- or downward trend of the underlying process for x; as a result, spurious oscillations can occur.

2.6.1 Upwinding

There are a number of well-established techniques to deal with convection-dominated PDEs. First, we can obviously attempt to lower Δ_x such that (2.34) is satisfied. This, however, may not be practical from a computational standpoint (and may require that Δ_t is lowered as well to avoid spurious oscillations originating from the time-stepping scheme). An alternative is to modify the first-order discrete operator δ_x such that it points in the direction of the large absolute drift. For instance, we can simply elect to use a suitably oriented one-sided difference, rather than a central difference, whenever (2.34) is violated. This procedure is known as *upstream differencing* or *upwinding*. To formalize the idea, introduce a new first-order difference operator δ_x^* given as

$$\delta_x^* V(t, x_j) = \begin{cases} \frac{1}{2}\left(V(t, x_{j+1}) - V(t, x_{j-1})\right) \Delta_x^{-1}, & |\mu(t, x_j)| \Delta_x \leq \sigma(t, x_j)^2, \\ \left(V(t, x_j) - V(t, x_{j-1})\right) \Delta_x^{-1}, & \mu(t, x_j) \Delta_x < -\sigma(t, x_j)^2, \\ \left(V(t, x_{j+1}) - V(t, x_j)\right) \Delta_x^{-1}, & \mu(t, x_j) \Delta_x > \sigma(t, x_j)^2. \end{cases}$$

Using δ_x^* instead of δ_x modifies the matrix $\mathbf{A}$ in (2.11). Specifically, if $\mu(t, x_j)\Delta_x < -\sigma(t, x_j)^2$ we replace (2.7)–(2.9) with:

$$c_j(t) = \mu(t, x_j)\Delta_x^{-1} - \sigma(t, x_j)^2 \Delta_x^{-2} - r(t, x_j), \tag{2.35}$$

$$u_j(t) = \frac{1}{2}\sigma(t, x_j)^2 \Delta_x^{-2}, \tag{2.36}$$

$$l_j(t) = -\mu(t, x_j)\Delta_x^{-1} + \frac{1}{2}\sigma(t, x_j)^2 \Delta_x^{-2}. \tag{2.37}$$

And when $\mu(t, x_j)\Delta_x > \sigma(t, x_j)^2$, we use

$$c_j(t) = -\mu(t, x_j)\Delta_x^{-1} - \sigma(t, x_j)^2 \Delta_x^{-2} - r(t, x_j), \tag{2.38}$$

$$u_j(t) = \mu(t, x_j)\Delta_x^{-1} + \frac{1}{2}\sigma(t, x_j)^2 \Delta_x^{-2}, \tag{2.39}$$

$$l_j(t) = \frac{1}{2}\sigma(t, x_j)^2 \Delta_x^{-2}. \tag{2.40}$$

For non-equidistant grids, a similar modification to (2.29)–(2.31) is required. We omit the straightforward details.

Let us try to gain some further understanding of the upwind algorithm. Comparison of (2.35)–(2.40) with (2.7)–(2.9), shows that upwinding amounts to using a regular central difference operator δ_x on a PDE with a diffusion coefficient modified to be $\sigma(t,x) + \sqrt{|\mu(t,x)|\Delta_x}$. The numerical scheme in effect introduces enough artificial diffusion into the PDE to satisfy (2.34). Doing so, however, comes at a cost: the convergence order of the scheme will be reduced to $O(\Delta_x)$ if one-sided differencing ends up being activated in a significant part of the grid. We note that higher-order upwinding schemes are possible if the finite difference operator δ_x^* is allowed to act on more than three neighboring points. For such schemes, the matrix $\mathbf{A}$ will no longer be tri-diagonal.

2.6.2 Other Techniques

As discussed earlier, upwinding amounts to adding numerical diffusion at nodes where the scheme is convection dominated. Alternatively, we can increase $\sigma(t,x)$ directly, to $\sigma(t,x) + \varepsilon$ where ε is chosen to be large enough for the scheme to satisfy (2.34). By solving the resulting PDE for different values of ε, it may be possible to determine how the error associated with ε scales in ε. This, in turn, will allow us to extrapolate to the limit $\varepsilon = 0$. See p. 135 of Tavella and Randall [2000] for an example.

The upwinding scheme presented in Section 2.6.1 switches abruptly from central differencing to one-sided differencing when the condition (2.34) is violated. In some schemes, the switch from central to one-sided differencing is made smooth by using a weighted average of a one-sided and a central difference operator. The weight on the central difference is close to one when $\sigma(t,x)^2 \gg |\mu(t,x)|\Delta_x$, but decreases smoothly to zero as $\sigma(t,x)^2/|\mu(t,x)|$ tends to zero. While it is unclear whether a smooth transition to upwinding is truly important (the convergence order is typically not improved over straight upwinding), Duffy [2000] suggests that the class of exponentially fitted schemes (see Duffy [2000] and Stoyan [1979]) may be quite robust in derivatives pricing applications.

In some finance applications, multi-dimensional PDEs might arise where $\sigma(t,x) = 0$ for one of the underlying variables; see for instance Section 2.7.5. While upwinding techniques still apply here, we note that specialized methods exist with better ($O(\Delta_x^2)$) convergence, should they become necessary. See, for instance, d'Halluin et al. [2005] for details on the so-called *semi-Lagrangian* methods.

2.7 Option Examples

In our discussion so far, we have assumed that options are characterized by a single terminal payoff function $g(x)$ and a set of spatial boundary

conditions determining the option price at the boundaries of the x-domain. In reality, many options are more complicated than this and may involve early exercise decisions, pre-maturity cash flows, path dependency, and more. In this section, we provide some relatively straightforward examples of such complications and how to modify the basic finite difference algorithm to deal with them. More examples will be provided later, in the context of specific fixed income securities.

2.7.1 Continuous Barrier Options

We have already touched upon the concept of an up-and-out knock-out option, an option that expires worthless if the x-process ever rises above a critical level H. As we described, we here must simply solve the PDE (2.1) on a domain $[\underline{M}, H]$, where $\underline{M}$ represents the lowest attainable value of the process $x(t)$ on $[0, T]$. The boundary condition at the upper boundary is then dictated to be $V(t, H) = 0$, i.e. of the Dirichlet type. We can generalize this to allow both "up" and "down" type barriers, and to perhaps give a non-zero payout (a "rebate") at the time the barrier(s) are hit (provided this happens before the option maturity). Specifically, if we have a lower barrier at $\underline{H}$, an upper barrier of $\overline{H}$, a time-dependent lower rebate function of $\underline{f}(t)$, and a time-dependent upper rebate function of $\overline{f}(t)$, we must dimension our spatial grid $\{x_j\}_{j=0}^{m+1}$ to have $x_0 = \underline{H}$, $x_{m+1} = \overline{H}$, and we then simply impose the Dirichlet boundary conditions $V(t, x_0) = \underline{f}(t)$ and $V(t, x_{m+1}) = \overline{f}(t)$. See (2.10) and the definition of $\mathbf{\Omega}$ for the algorithm required to incorporate such Dirichlet boundary conditions into the finite difference scheme.

In practice, barrier options sometimes involve time-dependent barriers, possibly with discontinuities. For instance, *step-up* and *step-down* barrier options will have piecewise flat barriers that increase (step-up) or decrease (step-down) at discrete points in time. Extension of the finite difference algorithm to cover step-up and step-down options is relatively straightforward. As an illustration, consider a zero-rebate up-and-out single-barrier option where the (upper) barrier is flat, except for a discontinuous change at time $T^* < T$, at which point the barrier moves from a value of H^* to a value of H, with $H > H^*$. We set the x-domain of our finite difference grid to $x \in [\underline{M}, H]$, with $\underline{M}$ a probabilistic lower limit, as defined above; accordingly, our spatial grid would be $\{x_j\}_{j=0}^{m+1}$, where $x_0 = \underline{M}$ and $x_{m+1} = H$. In preparation for the shift in barrier levels at time T^*, we make sure that one level in the spatial grid — say x_{k+1}, $k < m$, — is set exactly at the level H^*. Similarly, we make sure that one level in the time grid is set exactly to T^*. Starting at time T, we then iterate backwards in time by repeated solution of m-dimensional tri-diagonal systems of equations, at each step integrating a prescribed rebate function by supplying the Dirichlet boundary condition $V(t, x_{m+1}) = 0$. The moment we hit T^*, the PDE now only applies to the smaller region $[\underline{M}, H^*]$, covered by the reduced spatial grid $\{x_j\}_{j=0}^{k+1}$ with

$x_{k+1} = H^*$. From T^* back to time 0, the backward induction algorithm then involves only k-dimensional tri-diagonal systems of equations, with the Dirichlet boundary condition $V(t, x_{k+1}) = 0$. Spatial nodes above x_{k+1} correspond to zero option value and can be ignored[12]. Modification of the algorithm outlined above to handle more than two barrier discontinuities is straightforward.

We can extend our definition of barrier options even further by making the topology of "alive" and "dead" regions more complicated. At time t, assume for instance that the PDE applies in an "alive" region of $x \in L(t)$ and a rebate function $R(t, x)$ that applies in the "dead" region $D(t) = \mathcal{B} \backslash L(t)$. Assume that we discretize the problem on a single rectangular finite difference grid spanning the spatial domain $[\underline{M}, \overline{M}]$, where $\underline{M}$ and $\overline{M}$ are set such that the alive regions are covered, up to probabilistic limits (if necessary). Given option values at time t_{i+1}, we then only need to run the basic matrix equation (2.18) for values in our grid $\{x_j\}$ that lie inside $L(t_i)$. This requires scaling down the dimension of the matrix $\mathbf{A}$ as needed, and providing the relevant boundary conditions (given through $R(t_i, x_j)$) at the boundary (or boundaries) of $L(t_i)$. The parts of the spatial grid that lie outside of $L(t_i)$ can be directly filled in with values provided by the rebate function R. Notice that, if possible, the spatial grid should be set such that the boundaries of $L(t_i)$ are contained in the mesh; this will likely require us to use the techniques outlined in Section 2.4.

If the alive region has the simple form $L(t) = [\alpha(t), \beta(t)]$ for smooth deterministic functions α and β, an alternative to the scheme above is to introduce a time-dependent transformation that straightens out the barriers, allowing us to return to the standard finite difference setup where the PDE applies to a single rectangular (t, x) domain. One possible transformation involves using a spatial variable of

$$y = y(t, x) = \frac{x - \alpha(t)}{\beta(t) - \alpha(t)}, \tag{2.41}$$

which transforms the curved x-barriers $\alpha(t)$ and $\beta(t)$ into flat y-barriers at $y = 0$ and $y = 1$, respectively. The linearity of the transformation (2.41) makes it easy to work with; see Tavella and Randall [2000] for details and a discussion of extensions to multi-dimensional PDEs and to barriers with discontinuities.

[12] An obvious twist to the algorithm involves using different spatial grids over $[0, T^*]$ and $[T^*, T]$, allowing for more flexibility in node placement. In this case, values computed by backward induction must, at time T^*, be interpolated from one x-grid to another. The interpolation rule should be at least third-order accurate; see the discussion in Section 2.7.3.

2.7.2 Discrete Barrier Options

The barrier options considered in Section 2.7.1 are continuously monitored, in the sense that the barrier condition is observed for all times in a given interval. In practice, monitoring the barrier condition continuously can be impractical, and it may instead only be imposed on a discrete set of dates $T_1 < T_2 < \ldots < T_K$, with $T_K \leq T$ and $T_1 > 0$. For the sake of concreteness, let us consider a discretely monitored up-and-out option with a constant barrier H. For a continuously monitored up-and-out barrier option it would suffice to solve the PDE on a domain $x \in [\underline{M}, H]$, where $\underline{M}$ is a probabilistic lower limit. This is, however, no longer the case for a discretely monitored option where we need to allow the value function to "diffuse" above the barrier levels between dates in the monitoring set $\{T_k\}_{k=1}^K$. To allow for this, we discretize the PDE on a larger domain $x \in [\underline{M}, \overline{M}]$, $\overline{M} > H$. We can determine $\overline{M}$ probabilistically by determining a confidence interval for how far above the barrier $x(t)$ can rise between monitoring dates. For instance, for the Black-Scholes PDE (2.3), assume that $\max_{k=2,\ldots,K}(T_k - T_{k-1}) = \Delta_T$. Conditioned on $x(t) = H$, the probability that $x(t + \Delta_T)$ exceeds

$$x_\alpha = H + \left(r - \frac{1}{2}\sigma^2\right)\Delta_T + \alpha\sigma\sqrt{\Delta_T}$$

is $\Phi(-\alpha)$. As in Section 2.1, we recommend setting $\overline{M} = x_\alpha$ for values of α somewhere between 3 and 5. To properly capture diffusion between barrier observation dates, we should also dimension the time grid of the finite difference scheme such that multiple time steps (at least two or three, say) are taken between observation dates. All observation dates $\{T_k\}_{k=1}^K$ should obviously be contained in the time grid.

Between barrier observation dates, we solve our PDE by the standard finite difference algorithm outlined in Section 2.2.4, as always imposing either an asymptotic Dirichlet condition at $x = \overline{M}$ or a condition on the x-derivatives of the value function. At each barrier observation time T_k, we must impose a *barrier jump condition*

$$V(T_k-, x) = V(T_k+, x)1_{\{x<H\}}, \quad k = 1, \ldots, K, \tag{2.42}$$

where the notation $T_k\pm$ was introduced in Section 1.10.1 to denote the limit $T_k \pm \varepsilon$ for $\varepsilon \downarrow 0$. This merely states that all values $V(T_k, x)$ are zero for $x \geq H$, consistent with the definition of an up-and-out option. In our finite difference scheme, we incorporate this jump condition by simply interpreting the vector $\widehat{\mathbf{V}}(T_k)$ as found by regular backward induction as $\widehat{\mathbf{V}}(T_k+)$ and then replacing

$$\widehat{\mathbf{V}}(T_k+) = \left(\widehat{V}_1(T_k+), \ldots, \widehat{V}_m(T_k+)\right)^\top$$

with

$$\widehat{\mathbf{V}}(T_k-) = \left(\widehat{V}_1(T_k+)1_{\{x_1<H\}}, \ldots, \widehat{V}_m(T_k+)1_{\{x_m<H\}}\right)^\top$$

before continuing the algorithm backwards from T_k.

The jump condition (2.42) will generally produce a discontinuity in V as a function of x, around the barrier level H. If we use Crank-Nicolson time-stepping, it will then be prudent to employ a fully implicit scheme for the first few backwards time steps (Rannacher stepping) past each barrier observation date T_k. As discussed in Section 2.5, ideally this should be combined with a smoothing algorithm acting on $\widehat{\mathbf{V}}(T_k-)$ or, perhaps more conveniently, a shift of the spatial grid such that H lies exactly mid-way between two spatial nodes in the grid.

We round off by noting that the discussion above for an up-and-out option easily extends to more complicated discrete barrier options, including those with time-varying barrier levels and rebates. For instance, assume that an option involves upper and lower time-varying barriers of $\overline{H}(t)$ and $\underline{H}(t)$, respectively, as well as a time- and state-dependent rebate of $R(t,x)$. In this case, we simply replace the jump condition (2.42) with

$$\begin{aligned} V(T_k-,x) = V(T_k+,x)1_{\{\underline{H}(T_k)<x<\overline{H}(T_k)\}} \\ + R(T_k,x)\left(1_{\{x\geq\overline{H}(T_k)\}} + 1_{\{x\leq\underline{H}(T_k)\}}\right), \end{aligned}$$

and otherwise proceed as above. We note that time-dependent barriers will typically require flexibility in setting the spatial grid, as there are now multiple critical x-levels to consider. The discretization in Section 2.4 can obviously assist with this.

2.7.3 Coupon-Paying Securities and Dividends

Many fixed-income securities are coupon-bearing and involve periodic transfer of a cash amount between the buyer and the seller. This can easily be incorporated into a finite difference grid, through a jump condition. Specifically, consider a security that pays its owner a single cash amount of $p(T^*,x)$ at time $T^* < T$, where p is a deterministic function $p : [0,T] \times \mathcal{B} \to \mathbb{R}$. We dimension our time grid such that T^* is contained in the grid, and then apply at time T^* the condition

$$V(T^*-,x) = V(T^*+,x) + p(T^*,x). \tag{2.43}$$

This simply expresses that V will decrease by an amount p immediately after p is paid (and thereby no longer contained in V). In a finite difference algorithm, (2.43) is incorporated by replacing $\widehat{\mathbf{V}}(T^*+)$, as found by regular backward induction, with

$$\widehat{\mathbf{V}}(T^*-) = \left(\widehat{V}_1(T^*+) + p\left(T^*,x_1\right), \ldots, \widehat{V}_m(T^*+) + p\left(T^*,x_m\right)\right)^\top$$

before continuing the algorithm backwards from T^*. Extensions to multiple coupons are trivial.

In some cases a derivative security does not itself pay coupons, but is written on a security that does. This involves no particular complications, except for the case where payments may affect the state variable underlying the PDE. For instance, consider the classical case of a stock paying a dividend: at the time of the dividend payment, the stock jumps down by an amount equal to the dividend payment. For a model that uses the stock price (or a transformation of the stock price) as the state variable x, a dividend payment at time T^* would thus be associated with a discontinuity in the state variable, $x(T^*+) = x(T^*-) - d(T^*, x(T^*-))$, where d is the magnitude of the jump[13]. As long as the dividend-payment does not come as a surprise (i.e., at a random time), it must already be contained into the option price at T^*-, and will have no price effect as we move forward from T^*- to T^*+. We can express this continuity restriction through yet another jump condition

$$V(T^*-, x) = V\left(T^*+, x - d\left(T^*, x\right)\right). \tag{2.44}$$

See Wilmott et al. [1993] for more discussion. Implementation of (2.44) in a finite difference grid proceeds as follows. First, we use regular backward induction to establish

$$\begin{aligned}\widehat{\mathbf{V}}(T^*+) &= \left(\widehat{V}_1(T^*+), \ldots, \widehat{V}_m(T^*+)\right)^\top \\ &= \left(\widehat{V}(T^*+, x_1), \ldots, \widehat{V}(T^*+, x_m)\right)^\top.\end{aligned}$$

Then we write

$$\widehat{\mathbf{V}}(T^*-) = \left(\widehat{V}\left(T^*+, x_1 - d\left(T^*, x_1\right)\right), \ldots, \widehat{V}\left(T^*+, x_m - d\left(T^*, x_m\right)\right)\right)^\top.$$

The values $\widehat{V}_j(T^*+, x_j - d)$ here can be found by interpolation in the x-direction on the $\widehat{\mathbf{V}}(T^*+)$-array. As shown in Tavella and Randall [2000], the order of the interpolator should be strictly higher than two, to avoid inducing spurious numerical diffusion into our θ-style finite difference schemes. We note that this rules out the piecewise linear interpolation rule proposed in Wilmott et al. [1993]. A common choice is to use cubic spline interpolation; see Chapter 6 for much information on cubic splines.

2.7.4 Securities with Early Exercise

In Section 1.10 we introduced the concept of Bermudan and American securities with early exercise features. Under the assumption that exercise

[13] To prevent negative stock prices, it may be necessary to truncate the size of d locally in the finite difference grid. For simplicity, we ignore this complication here.

values are determined by a deterministic function[14] $h(t,x)$, $h:[0,T]\times\mathcal{B}\to\mathbb{R}$, finite difference grids are ideal for pricing of such securities. Let us first consider a Bermudan option with exercise opportunities restricted to the finite set $\{T_k\}_{k=1}^K$. The Bellman principle (1.67) in Section 1.10 can, as shown there, be expressed as a simple jump condition

$$V(T_k-,x)=\max\left(V\left(T_k+,x\right),h\left(T_k,x\right)\right),\quad k=1,\ldots,K, \tag{2.45}$$

which can be incorporated into a finite difference solver precisely the same way as in previous sections. The condition (2.45) will result in a kink in the value function around the level of x at which we shift from the hold region into the exercise region. If Crank-Nicolson time-stepping is used, one should ideally apply smoothing on the finite difference value vector $\widehat{\mathbf{V}}(T^*-)$, particularly around the kink.

If exercise can take place continuously (that is, American-style) on a given time interval, a crude way to incorporate this into a finite difference grid is by simply applying (2.45) to every point in the time grid of the finite difference scheme. By not specifically imposing the partial differential inequalities (see Section 1.10.1), this algorithm, however, will generally only be accurate to first order in the time step, even if a Crank-Nicolson scheme is used; see Carverhill and Clewlow [1990] for a proof. As American-style exercise is rarely used in fixed income markets, we shall not pursue this issue further but just point out that a number of schemes exist to restore second-order time convergence to finite difference pricing of American options, see, e.g., Forsyth and Vetzal [2002].

2.7.5 Path-Dependent Options

Finite difference methods are normally limited to Markovian problems where dynamics are characterized by SDEs and where payouts are simple deterministic functions of the underlying state variables. A number of options, however, have terminal time T payouts that depend not only on the state of x at time T, but on the entire path $\{x(t),\ t\in[0,T]\}$. In general, such options must be priced by Monte Carlo methods (see Chapter 3), but exceptions exist. Indeed, barrier and American options can be considered path-dependent options, yet, as we have seen, can still be priced in a finite difference grid. Even stronger path-dependence can sometimes be handled, through the introduction of new state variables to the PDE.

To give an example, consider a path-dependent contract where the terminal payout at time T can be written as

[14] If h represents the value of a derivative security that has no closed-form pricing formula, it may be necessary to estimate this function by backward induction in the finite difference grid itself. Such a "preprocessing" step is typically straightforward to execute.

$$V(T) = g\left(x(T), I(T)\right), \tag{2.46}$$

where I is a path integral of the type

$$I(t) = \int_0^t h\left(x(s)\right)\, ds, \tag{2.47}$$

for some deterministic function h. For instance, if $h(x) = x$, we say that the option is a continuously sampled *Asian option*.

For the payout (2.46) we have $V(t) = V(t, x(t), I(t))$ where $x(t)$ satisfies the SDE (2.2) and

$$dI(t) = h\left(x(t)\right)\, dt, \quad I(0) = 0.$$

From the backward Kolmogorov equation, it follows that $V(t, x, I)$ solves

$$\frac{\partial V}{\partial t} + \mu(t,x)\frac{\partial V}{\partial x} + \frac{1}{2}\sigma(t,x)^2\frac{\partial^2 V}{\partial x^2} + h(x)\frac{\partial V}{\partial I} = r(t,x)V, \tag{2.48}$$

subject to the terminal condition $V(T, x, I) = g(x, I)$. There are several complications with this PDE. First, it involves *two* spatial variables, x and I, requiring the use of a two-dimensional PDE solver. Second, the PDE contains no second-order derivative in the variable I, i.e. it is convection dominated in the I-direction. We have discussed methods to handle the latter issue in Section 2.6.1 and will turn to address the former in Section 2.9. Another complication is the fact that the term $h(x)$ multiplying $\partial V/\partial I$ may be of a different order of magnitude than the other coefficients in (2.48), increasing the difficulty of solving the equation numerically. We refer to Zvan et al. [1998] for a more detailed discussion of PDEs of the type (2.48).

In practice, it is rare that a continuous-time integral such as (2.47) is used in an option payout. Instead, one normally samples the function $h(x(t))$ only on a discrete set of dates, i.e. we replace $I(T)$ with

$$I(T) = \sum_{i=1}^{n} h\left(x(T_i)\right)\left(T_i - T_{i-1}\right),$$

where $T_0 < T_1 < \ldots < T_n$ is a discrete schedule, with $T_0 = 0$ and $T_n = T$. Informally, we now have

$$dI(t) = \delta\left(T_i - t\right) \cdot h\left(x(T_i)\right)\left(T_i - T_{i-1}\right), \quad I(0) = 0, \tag{2.49}$$

where $\delta(\cdot)$ is the Dirac delta function. In a PDE setting, we incorporate a process such as (2.49) through appropriate jump conditions, writing

$$V(T_i-, x, I) = V(T_i+, x, I + h(x)\left(T_i - T_{i-1}\right)). \tag{2.50}$$

In the same fashion as for discrete dividends (Section 2.7.3), the jump condition enforces continuity of the option price across the dates where I

gets updated. The condition is applied at each date in the discrete schedule, $i = 1, \ldots, n$; in between schedule dates (where now $dI(t) = 0$), we solve the PDE

$$\frac{\partial V}{\partial t} + \mu(t,x)\frac{\partial V}{\partial x} + \frac{1}{2}\sigma(t,x)^2\frac{\partial^2 V}{\partial x^2} = r(t,x)V,$$

which has no term involving I. When the I-direction is discretized in, say, m_I different values, the solution scheme thus involves solving m_I different *one-dimensional* PDEs backward in time; the solutions of these m_I PDEs exchange information with each other at each date in the schedule, in accordance with (2.50). As was the case for cash dividends, implementation of (2.50) will normally require support from an interpolation scheme, to align the (x-dependent) jumps in I with the knots of the discretized I-grid used in the finite difference scheme. See, e.g., Zvan et al. [1999] or Wilmott et al. [1993] for further details. An application of this idea in the context of interest rate derivatives is given in Section 18.4.5.

On rare occasions — basically when the homogeneity condition $V(\eta x, \lambda I, t) = \lambda^\eta V(x, I, t)$, $\lambda, \eta > 0$, holds — it is possible to make a change of variables or a change of probability measure that will reduce (2.48) or its discrete-time version to a one-dimensional PDE; see e.g. Rogers and Shi [1995] or Andreasen [1998] for the case of various Asian options. Section 18.4.5 demonstrates one such method, sometimes called the method of *similarity reduction*, for pricing of "weakly path-dependent" securities, including certain callable interest rate derivatives where the notional accretes at a stochastic coupon rate (see Section 5.14.5 for definitions).

2.7.6 Multiple Exercise Rights

Certain financial products with early exercise rights allow the holder to exercise more than once. Such "multi-exercise" options are relatively rare, but the so-called *chooser cap* (also known as a *flexi-cap*) is occasionally traded and constitutes a good example for describing how to handle multi-exercise options in a PDE setting. Let there be given a set of L possible exercise dates, $T_1 < T_2 < \ldots < T_L$, and assume that we have the right to exercise no more than l times, with $l < L$. Provided that we exercise at time T_i, in a chooser cap we are paid[15] $(S(T_i) - K)^+$, where $S(\cdot)$ is some interest rate index and K is the strike. Clearly, we would never exercise at time T_i unless $S(T_i) > K$, but how much larger than K the rate $S(T_i)$ needs to be to trigger optimal exercise is not obvious, and must at least depend on i) how many of our l exercise opportunities we have already used up at time T_i; and ii) how much value is lost by using (rather than postponing) one of the remaining exercise opportunities.

[15]We have ignored a day count scaling constant in the payout. Also, in most cases payment takes place at time T_{i+1}, rather than at T_i; such a payment delay can be handled by a discount operation.

While the question of how to exercise optimally on a chooser cap may appear quite complex, it is surprisingly easy to implement in a finite difference setting by combining techniques from Sections 2.7.4 and 2.7.5 above. The key to the method is to introduce an additional state variable I to keep track of how many exercise opportunities are left. Assume that all interest rates are functions of a Markov state variable $x(\cdot)$, and let therefore $V(t, x, I)$ denote the value of the chooser cap at time t, given $x(t) = x$ and given that there are still I exercise opportunities left. Notice that the variable I can only take $l + 1$ distinct values: $0, 1, \ldots, l$; notice also that $V(t, x, 0) = 0$ for all t and x, since $I = 0$ corresponds to the situation where there are no exercise opportunities left. Additionally, at the terminal time T_L we clearly have

$$V(T_L, x, I) = (S(T_L, x) - K)^+, \quad I = 1, 2, \ldots, l, \tag{2.51}$$

where we have written $S(T_L) = S(T_L, x)$ to emphasize the deterministic dependence of S on the state variable x.

For given dynamics of $x(t)$, starting with the terminal conditions in (2.51), we may roll the l different value functions $V(\cdot, x, I)$, $I = 1, 2, \ldots, l$, back through time in standard finite difference manner. At each time T_i, $i = 1, \ldots, L - 1$, jump conditions similar to (2.45) must be applied, for all $I = 1, 2, \ldots, l$:

$$V(T_i-, x, I) = \max\left(V(T_i+, x, I), V(T_i+, x, I - 1) + (S(T_i, x) - K)^+\right).$$

Notice that these conditions simply express that exercise is optimal only if the exercise value (the cap payout plus the value of a chooser cap with one less exercise opportunity) exceeds the hold value (the non-exercised chooser cap). Once we have rolled all the way back to $t = 0$, the chooser cap value at time $t = 0$ may be identified as $V(0) = V(0, x(0), l)$.

We should note that the "chooser" or "flexi" feature can be added to securities other than caps (and floors). For instance, in Section 19.5 we study the so-called *flexi-swap*, another security with multiple embedded exercise rights.

2.8 Special Issues

In this section, we briefly show a few techniques that may come in handy for certain applications.

2.8.1 Mesh Refinements for Multiple Events

As discussed in Section 2.1, the domain of the state variable x is often determined as an exact or approximate confidence interval for the random variable $x(T)$, where T is the final time of interest for a particular valuation problem we want to solve. Given the number of desired spatial steps in the

scheme, the discretization step in x-direction is then obtained by dividing the size of the confidence interval by the number of steps. Similarly, the discretization step in t-direction is typically obtained by dividing T by the number of desired time steps. This is a standard procedure for building a simple rectangular mesh, and it works well if the derivative we wish to value does not have any "interesting" features between the valuation time 0 and the final time T (e.g., for a simple European option). However, as should be evident from the examples in Section 2.7, many real-life derivative securities are characterized by a multitude of events during their lifetimes, all of which must be adequately captured in the PDE scheme. It is not hard to see that a grid dimensioning scheme based solely on the last event date may yield inappropriate mesh resolution at earlier dates.

To make the discussion above concrete, let us consider the example of a Bermudan option (see Section 2.7.4) with two exercise dates, T_1 and T_2. Assume that $0 < T_1 \ll T_2$, i.e. that the first exercise date is much closer to the valuation date than the second (and last) one. Also assume that there is a decent chance that the option actually will be exercised at time T_1, making it important to capture to good precision the value of the option expiring at T_1. Now, if we build our mesh based only on the distribution of the state variable $x(T_2)$ at time T_2, there would typically be too few t-points in the interval $[0, T_1]$. Also, the x-direction discretization step would be too large compared to the range of possible values of the state variable $x(T_1)$ at time T_1, i.e. the x-grid would be too coarse for the process $x(\cdot)$ on the time interval $[0, T_1]$. Both issues would typically lead to a large discretization error in the finite difference stepping of the option over the time period $[0, T_1]$, leading to problems with accuracy in values and risk sensitivities.

The issue of the sparsity of the time grid is fairly easy to deal with, as we are free to add extra points to the time grid before time T_1. This by itself, however, will not solve precision problems, as the space step remains large. Any proper solution should, of course, come in the form of refining both the t- and x-grids at the same time.

One possible way of refining the x-discretization is to abandon the usage of a single rectangular (t, x)-domain, and instead link together different equidistant rectangular meshes for different periods in the life of the derivative. These mesh "blocks" would generally increase in spatial width with time and would connect to each other via an interpolation scheme. To be more specific, let us assume, as in Section 2.1, that the state variable $x(\cdot)$ is the logarithm of the stock in the Black-Scholes model and is given by (2.4), with the PDE to solve given by (2.3). We extend our simple two-period example above to a derivative with K times of interest, $0 < T_1 < \ldots < T_K$; these times could be specified as an additional input into valuation, or derived from the trade description (e.g. they could represent the exercise dates for a Bermudan option, or the knock-out dates for the discretely-monitored barrier option of Section 2.7.2). Suppose we are given values of m and n, and now wish to construct the mesh for the time period $[T_{k-1}, T_k]$, by using

the same time and space steps Δ_t^k, Δ_x^k as would be used in the standard scheme of Section 2.1 for a derivative security with the terminal payoff at T_k. That is, having fixed the cutoff α we would set

$$\Delta_t^k = T_k/n, \quad \Delta_x^k = 2\alpha\sigma\sqrt{T_k}/(m+1). \tag{2.52}$$

Then the rectangular, equidistant mesh for the time period $[T_{k-1}, T_k]$ is given by

$$\left\{t_i^k\right\}_{i=0}^{\lfloor(1-T_{k-1}/T_k)n\rfloor} \times \left\{x_j^k\right\}_{j=0}^{m+1}, \quad t_i^k = T_{k-1} + i\Delta_t^k, \quad x_j^k = x_{\min}^k + j\Delta_x^k, \tag{2.53}$$

where $\lfloor\cdot\rfloor$ denotes the integer part of a real number and (see (2.4))

$$x_{\min}^k = x(0) + \left(r - \frac{1}{2}\sigma^2\right)T_k - \alpha\sigma\sqrt{T_k}. \tag{2.54}$$

Note that in reality we would want to make sure that the point T_k is also in the mesh for the time period $[T_{k-1}, T_k]$, even though for simplicity of notations we did not reflect it in (2.53). It is also useful to note that the total number of time points is not going to be n, but is actually equal to

$$\sum_{k=1}^{K} \lfloor(1 - T_{k-1}/T_k)\, n\rfloor,$$

which scales linearly with n. Clearly, if exactly n points were required, a simple adjustment to the definition of the time step in (2.52) could be applied.

With a mesh as defined above, when arriving at time T_k in a backward induction scheme the solution $V(T_k, \cdot)$ would be discretized on the x-grid $\{x_j^{k+1}\}_{j=0}^{m+1}$. To solve the PDE backwards over the time period $[T_{k-1}, T_k]$, we would need to resample it on the different x-grid $\{x_j^k\}_{j=0}^{m+1}$. As with interpolation across dividends (Section 2.7.3), simple cubic interpolation would be a good choice here. Specifically, one would fit a cubic spline to the values $V(T_k, x_j^{k+1})$, $j = 0, \ldots, m+1$, and then calculate $V(T_k, x_j^k)$, $j = 0, \ldots, m+1$, by valuing the spline at the required grid points.

The "interpolated mesh" scheme above is rather intuitive and straightforward, but it does suffer from the need to do interpolation work that could slow down the PDE (especially in dimensions higher than 1 and/or for a large number of interface points K). Also, it is not entirely clear how interpolation will affect stability and convergence properties of the PDE. Finally, linking the interface mesh geometry to the trade specifics (such as exercise dates) may not be ideal from the point of view of designing an efficient valuation flow in a risk management system. These considerations lead us to an alternative approach that relies on *non*-equidistant discretization as developed in Section 2.4. The idea of this method is to use non-uniform

discretization to concentrate more points, both in time and space, around the initial point $t = 0$, $x = x(0)$. Clearly many ways of achieving this are possible — below we present a simple scheme we have used with good results.

We define K, the user input, to be the number of spatial refinement levels (with $K = 2$ or 3 typically used), and τ, another user input, to be a time scaling constant (typically $\tau = 4$). If T is the final horizon for valuation, we then introduce times

$$0 = T_0 < T_1 < \ldots < T_K = T$$

by

$$T_k = \frac{T}{\tau^{K-k}}, \quad k = 1, \ldots, K.$$

Then, the time grid for the time period $[T_{k-1}, T_k]$ is given by uniformly distributing $\widetilde{n} \triangleq \lfloor n/K \rfloor$ points[16] over $[T_{k-1}, T_k]$, i.e. is given by $\{t_i^k\}_{i=0}^{\widetilde{n}}$ with

$$t_i^k = T_{k-1} + i\frac{T_k - T_{k-1}}{\widetilde{n}} = T_{k-1} + i\frac{T_k - T_{k-1}}{\lfloor n/K \rfloor}.$$

(Note that we can use this specification with the interpolated mesh as well, instead of the time grid definition in (2.53)). The fact that the width of the intervals $[T_{k-1}, T_k]$ grow with k means that the time grid is more finely spaced in the beginning of the interval $[0, T]$ than at the end.

The x-grid we are going to define will be universal — i.e. the same for all time steps on the whole time interval $[0, T]$ — and non-uniform. To construct it, we first define a set of nested x-subdomains $[x_{\min}^k, x_{\max}^k]$, with $x_{\min}^k$ defined by (2.54) and $x_{\max}^k$ defined accordingly, i.e.

$$x_{\max}^k = x(0) + \left(r - \frac{1}{2}\sigma^2\right) T_k + \alpha\sigma\sqrt{T_k},$$

for $k = 0, \ldots, K$. Then we define step sizes by

$$\Delta_x^k = \frac{x_{\max}^k - x_{\min}^k}{\widetilde{m} + 1}, \quad \widetilde{m} = \left\lfloor \frac{m}{K} \right\rfloor.$$

The x-grid is then constructed by distributing grid points uniformly in subintervals $[x_{\min}^k, x_{\min}^{k-1}]$ and $[x_{\max}^{k-1}, x_{\max}^k]$ with the space step Δ_x^k, and is given by

$$\left(\bigcup_{k=1}^{K} \left\{x_j^{\min,k}\right\}_{j=0}^{m^k+1}\right) \cup \left(\bigcup_{k=1}^{K} \left\{x_j^{\max,k}\right\}_{j=0}^{m^k+1}\right),$$

where

$$x_j^{\min,k} = x_{\min}^k + j\Delta_x^k, \quad x_j^{\max,k} = x_{\max}^{k-1} + j\Delta_x^k,$$

[16] And, as advised earlier, adding trade event dates that fall into this period — although we do not reflect this in our notations for simplicity.

and

$$m^k = \left\lfloor \frac{x_{\min}^{k-1} - x_{\min}^k}{\Delta_x^k} \right\rfloor - 1 = \left\lfloor \frac{x_{\max}^k - x_{\max}^{k-1}}{\Delta_x^k} \right\rfloor - 1.$$

This distribution of space points results in an x-grid that is more dense around the point $x = x(0)$ than at the edges. It is worth noting that with only one refinement level $K = 1$, the standard rectangular uniform mesh sized by the terminal distribution of the state variable is recovered.

2.8.2 Analytics at the Last Time Step

In cases where the dynamics of underlying PDE variables are tractable, one naturally wonders whether finite difference methods could somehow be improved by incorporating analytical results into the scheme. Here, and in the next section, we discuss two simple ideas.

Suppose that we are faced with the problem of pricing a contingent claim with terminal boundary condition $g(x(T))$, where $x(t)$ is a Markovian process with known Arrow-Debreu state prices:

$$G(t, x; s, y) = \mathrm{E}^{\mathrm{Q}} \left(\delta\left(x(s) - y\right) e^{-\int_t^s r(u, x(u))\, du} | x(t) = x \right), \quad s > t.$$

Assume also that the claim in question involves a jump condition at time $0 < T^* < T$ (but no jump conditions between T^* and T). If our finite difference grid is $\{x_j\}_{j=0}^{m+1}$, we can now use a series of $m + 2$ outright convolutions to compute

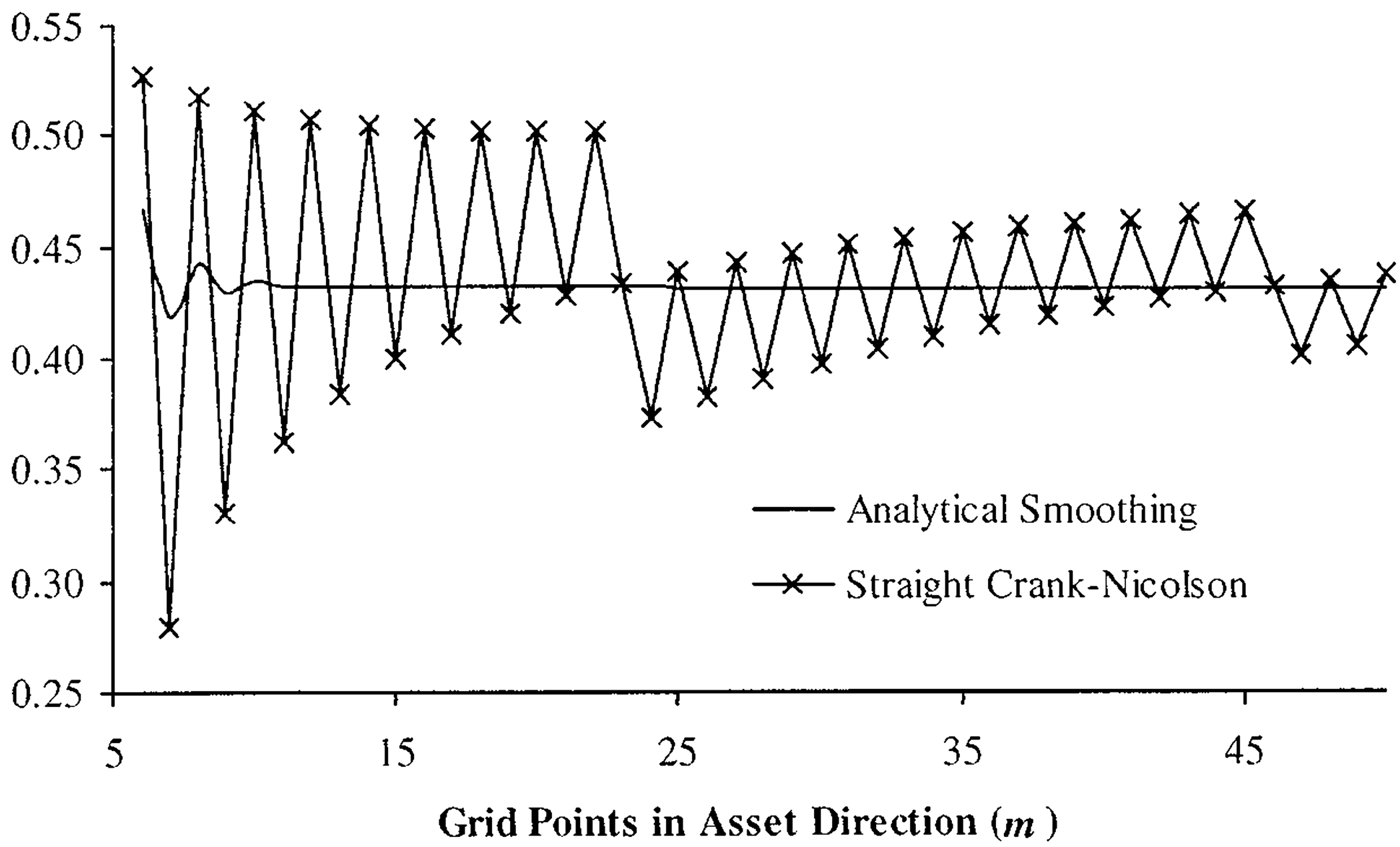
$$V(T^*, x_j) = \int_{\mathbb{R}} G\left(T^*, x_j; T, y\right) g(y)\, dy, \quad j = 0, 1, \ldots, m+1. \tag{2.55}$$

If we are lucky (i.e., if both g and G are sufficiently simple), then the integral on the right-hand side may be known in closed form for all values of x_j. If not, we can always perform a series of numerical integrations, the total cost of which is typically[17] $O(m^2)$, i.e. more expensive than the typical $O(m)$ cost of a single time step in a finite difference method. There are several reasons why we may want to perform the numerical integrations nevertheless. First, the convolution expression (2.55) is exact, as it is based on the true transition density. Second, if the gap between T^* and T is large, an ordinary finite difference grid would need to roll back from T to T^* using multiple time steps n^*, at a total cost of $O(n^* m)$; if n^* is of the same magnitude as m, the computational effort of the convolution scheme would be comparable to that of a finite difference grid. Third, for discontinuous payouts, the integration in (2.55) will have a naturally smoothing effect, similar to (but often better than) the continuity correction method of Section 2.5.2. The smoothing

[17] There are exceptions. For instance, if fast Fourier transform (FFT) methods are applicable, the cost may be reduced to $O(m \ln(m))$. See Section 8.4 for details.

effect is discussed in more detail in Section 23.2.4 and is also demonstrated below, in Figure 2.2, where we have continued our investigation of the 3 year digital option considered earlier in Section 2.5.3. Since the model used in Figure 2.2 is ordinary Black-Scholes and $g(x) = 1_{\{x>H\}}$, the integrals in (2.55) can here be computed in closed form from (2.33).

Fig. 2.2. 3 Year Digital Option Price



Notes: Finite difference estimates for the Black-Scholes price of a 3 year digital option. All contract and model parameters are as in Figure 2.1. Time stepping is performed with an equidistant grid containing $n = 50$ points. Spatial discretization in log-space is equidistant, as described in the main text; the number of grid points (m) is as listed on the x-axis of the figure. The "Straight Crank-Nicolson" graph shows the convergence profile for a pure Crank-Nicolson finite difference grid. The "Analytical Smoothing" graph shows the convergence profile for a Crank-Nicolson finite difference grid starting at $T^* = 2.5$ years, with the terminal boundary condition set equal to a 0.5 year digital option price (as in (2.55)). The theoretical value of the option is 0.4312451.

In principle, we could continue rolling back from T^* (through, possibly, jump conditions at earlier times) by performing convolutions, rather than solving finite difference grids. In practice, this rarely leads to improvements over a finite difference grid, unless the densities and payoffs are quite simple[18]. Moreover, in many cases we may not have *exact* Arrow-Debreu

[18]For simple densities (especially Gaussian), special-purpose methods exist to compute convolutions rapidly, typically involving payoff approximations through piecewise polynomials or other simple functions. We do not cover these methods in

prices, only approximate ones based on, say, a small-time expansion (see, e.g., Section 13.1.9.1). In this case, a one-time convolution may be safe — especially if $T - T^*$ is small — whereas repeated convolutions may lead to unacceptable biases.

2.8.3 Analytics at the First Time Step

The idea in Section 2.8.2 of replacing the finite difference stepping with analytical integration is even easier to apply over the *first*, rather than the *last*, time step. Suppose T^* is the first "interesting" time for a given derivative security, i.e. there might be a jump condition at time T^* but none over the time interval $[0, T^*]$. Then, rather than stepping the finite difference scheme from T^* to 0, we can perform a *single* integration to calculate the value $V(0, x(0))$ of the derivative at time zero from the discretized values $\{V(T^*, x_j)\}_{j=0}^{m+1}$ of the derivative at time T^* (using the same notations as in Section 2.8.2),

$$V(0, x(0)) = \int_{\mathbb{R}} G\,(0, x(0); T^*, y)\, \widetilde{V}(T^*, y)\, dy,$$

where $\widetilde{V}(T^*, y)$ is interpolated (using cubic splines, say) from the values $\{V(T^*, x_j)\}_{j=0}^{m+1}$ on the grid. If the integral is computed numerically — as is most often the case — the numerical cost is often comparable with that of the finite difference stepping because only one value $V(0, x(0))$ is required at time 0, not the whole slice.

While there are typically no numerical cost savings that arise from using integration over the first time step, there are accuracy and stability considerations that favor this approach. We have already seen in Section 2.8.1 that the standard discretization of a PDE often leads to insufficient fidelity in resolving any features of the payoff that are close to today, and numerical integration can be of considerable help in this regard. Moreover, as we discuss in much detail later in Chapter 23, an integration scheme typically allows us to treat discontinuities in the value $V(T^*, x)$ arising from the jump condition at time T^* explicitly. If the discontinuity is introduced at the value of the state variable x^*, then the integration scheme can (and should) explicitly take this information into account. For example we would write

$$\begin{aligned} V(0, x(0)) = &\int_{-\infty}^{x^*} G\,(0, x(0); T^*, y)\, \widetilde{V}^-(T^*, y)\, dy \\ &+ \int_{x^*}^{\infty} G\,(0, x(0); T^*, y)\, \widetilde{V}^+(T^*, y)\, dy \end{aligned}$$

---

this book except for a brief mention in Section 11.A. For a representative example see Hu et al. [2006].

and calculate $\widetilde{V}^-(T^*, y)$ by interpolating the grid values in the time interval $(-\infty, x^*)$, and $\widetilde{V}^+(T^*, y)$ by interpolating the grid values in (x^*, ∞), separately[19].

The usefulness of the method is only limited by the availability of the closed-form expression for the time 0 Arrow-Debreu prices $G(0, x(0); T^*, \cdot)$. For some models this is not an issue; for most others, sufficiently close approximations could be obtained in a small-time limit (see e.g. Section 13.1.9.1 for a typical approach) that can be useful for times T^* that are not too large. By a change of measure, we see that

$$\begin{aligned} V(0, x(0)) &= \mathrm{E}\left(e^{-\int_0^{T^*} r(s)\,ds} V(T^*, x(T^*))\right) \\ &= P(0, T^*)\mathrm{E}^{T^*}\left(V(T^*, x(T^*))\right), \end{aligned}$$

where E^{T^*} is the expected value operator under the T^*-forward measure Q^{T^*}; so we really only need the expression for the *density* (rather than Arrow-Debreu security prices) of $x(T^*)$ under Q^{T^*}, either exact or approximate.

Finally, we note that while the integration over the first time step can be seen to offer similar advantages to those of the methods in Section 2.8.1, the two approaches are not substitutes for each other, but are complementary. We typically recommend using direct integration over the time step $[0, T^*]$, where T^* is the smaller of the time of the first jump condition or the limit of applicability of the approximation to the density of $x(T^*)$, and then (if needed) use the methods in Section 2.8.1 over the time interval $[T^*, T]$, with T being the final maturity of the option in question.

2.9 Multi-Dimensional PDEs: Problem Formulation

We now turn our attention to the numerical solution of multi-dimensional terminal value problems. Let the spatial variable x be p-dimensional, $x = (x_1, \ldots, x_p)^\top$, and consider the PDE

$$\frac{\partial V}{\partial t} + \sum_{h=1}^{p} \mu_h(t, x)\frac{\partial V}{\partial x_h} + \frac{1}{2}\sum_{h=1}^{p}\sum_{l=1}^{p} s_{h,l}(t, x)\frac{\partial^2 V}{\partial x_h \partial x_l} - r(t, x) = 0, \quad (2.56)$$

where $s_{h,h}(t, x) \geq 0$ and $s_{h,l}(t, x) = s_{l,h}(t, x)$ for $h, l = 1, \ldots, p$. The PDE is assumed subject to the terminal value condition $V(T, x) = g(x)$, $g : \mathbb{R}^p \to \mathbb{R}$.

From the results in Chapter 1, we recognize that the PDE provides the solution to the expectation

[19] One of the functions $\widetilde{V}^-(T^*, y)$, $\widetilde{V}^+(T^*, y)$ is often known analytically and for all values of y (rather than sampled on the grid); this is for instance the case for the Bermudan options of Section 2.7.4. The integration algorithm should obviously take advantage of this.

$$V(t,x) = \mathrm{E}_t\left(e^{-\int_0^T r(u,x)\,du} g\left(x(T)\right)|x(t)=x\right),$$

where the components of $x(t)$ satisfy risk-neutral SDEs of the type

$$dx_h(t) = \mu_h\left(t, x(t)\right)\,dt + \sigma_h\left(t, x(t)\right)\,dW(t), \quad h = 1,\ldots,p. \tag{2.57}$$

Here $W(t)$ is a d-dimensional Brownian motion, $\mu_h : [0,T]\times\mathbb{R}^p \to \mathbb{R}$, $h = 1,\ldots,p$, are (scalar) drifts, and $\sigma_h : [0,T]\times\mathbb{R}^p \to \mathbb{R}^{1\times d}$, $h = 1,\ldots,p$, are d-dimensional (row vector) diffusion coefficients. The PDE coefficients $s_{h,l}$ in (2.56) represent the instantaneous covariance matrix for the components of $x(\cdot)$, i.e., $s_{h,l}(t,x) = \sigma_h(t,x)\sigma_l(t,x)^\top$. We assume enough regularity on μ_h, σ_h, r, and g to ensure that (2.56) has a unique solution.

For the purpose of solving (2.56) numerically, we assume that the PDE is to be solved on a (finite) spatial domain in x, $x \in [\underline{M}_1, \overline{M}_1]\times\ldots\times[\underline{M}_p, \overline{M}_p]$, where $\underline{M}_h, \overline{M}_h$, $h = 1,\ldots,p$, are constants either dictated by the contract at hand (barrier options) or found by a suitable probabilistic truncation (see Section 2.1).

2.10 Two-Dimensional PDE with No Mixed Derivatives

To illustrate the construction of finite difference discretization of (2.56), we start out with the simple case where $p = d = 2$ and there are no mixed partial derivatives in the PDE: $s_{1,2}(t,x) = s_{2,1}(t,x) = 0$ for all t and x. Probabilistically, the absence of mixed derivatives corresponds to the case where the stochastic process increments $dx_1(t)$ and $dx_2(t)$ are independent. Defining $\gamma_h(t,x)^2 = s_{h,h}(t,x)$, $h = 1,2$, the PDE to be solved now becomes

$$\frac{\partial V}{\partial t} + \left(\mathcal{L}_1 + \mathcal{L}_2\right)V = 0, \tag{2.58}$$

where

$$\mathcal{L}_h \triangleq \mu_h(t,x)\frac{\partial}{\partial x_h} + \frac{1}{2}\gamma_h(t,x)^2\frac{\partial^2}{\partial x_h^2} - \frac{1}{2}r(t,x), \quad h = 1,2.$$

Notice that we have divided the term $r(t,x)$ into equal pieces in $\mathcal{L}_1$ and $\mathcal{L}_2$.

To discretize (2.58) in x, introduce grids $x_1 \in \{x_1^{j_1}\}_{j_1=0}^{m_1+1}$ and $x_2 \in \{x_2^{j_2}\}_{j_2=0}^{m_2+1}$. To simplify notation, assume these grids are equidistant such that $x_1^{j_1} = \underline{M}_1 + j_1\Delta_1$ and $x_2^{j_2} = \underline{M}_2 + j_2\Delta_2$. Let $V_{j_1,j_2}(t) \triangleq V(t, x_1^{j_1}, x_2^{j_2})$. We define discrete central difference operators as before

$$\delta_{x_1}V_{j_1,j_2}(t) = \frac{V_{j_1+1,j_2}(t) - V_{j_1-1,j_2}(t)}{2\Delta_1},$$

$$\delta_{x_2}V_{j_1,j_2}(t) = \frac{V_{j_1,j_2+1}(t) - V_{j_1,j_2-1}(t)}{2\Delta_2},$$

and

$$\delta_{x_1x_1}V_{j_1,j_2}(t) = \frac{V_{j_1+1,j_2}(t) - 2V_{j_1,j_2}(t) + V_{j_1-1,j_2}(t)}{\Delta_1^2},$$
$$\delta_{x_2x_2}V_{j_1,j_2}(t) = \frac{V_{j_1,j_2+1}(t) - 2V_{j_1,j_2}(t) + V_{j_1,j_2-1}(t)}{\Delta_2^2}.$$

These operators, in turn, give rise to the discrete operators

$$\widehat{\mathcal{L}}_h \triangleq \mu_h(t,x)\delta_{x_h} + \frac{1}{2}\gamma_h(t,x)^2\delta_{x_hx_h} - \frac{1}{2}r(t,x), \quad h = 1,2,$$

where x is constrained to take values in the spatial grid. A Taylor expansion shows that this operator is second-order accurate (compare to Lemma 2.2.1),

$$(\mathcal{L}_1 + \mathcal{L}_2)V(t,x) = (\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2)V(t,x) + O\left(\Delta_1^2 + \Delta_2^2\right).$$

2.10.1 Theta Method

Turning to a theta-style time discretization, consider first proceeding exactly as in Section 2.2.3. Assuming equidistant time spacing Δ_t, we get for the period $[t_i, t_{i+1}]$,

$$\left(1 - \theta\Delta_t\left(\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2\right)\right)V_{j_1,j_2}(t_i)$$
$$= \left(1 + (1-\theta)\Delta_t\left(\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2\right)\right)V_{j_1,j_2}(t_{i+1}) + e_i^{i+1},$$

where

$$e_i^{i+1} = O\left(\Delta_t\left(\Delta_1^2 + \Delta_2^2 + 1_{\{\theta\neq\frac{1}{2}\}}\Delta_t + \Delta_t^2\right)\right),$$

and where it is understood that $\widehat{\mathcal{L}}_1$ and $\widehat{\mathcal{L}}_2$ are to be evaluated at $(t,x) = (t_i^{i+1}(\theta), x_1^{j_1}, x_2^{j_2})$ with $t_i^{i+1}(\theta)$ defined as in (2.14). If $\widehat{V}_{j_1,j_2}(t) \triangleq \widehat{V}(t, x_1^{j_1}, x_2^{j_2})$ is a finite difference approximation to $V_{j_1,j_2}(t)$, we thus get the scheme

$$\left(1 - \theta\Delta_t\left(\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2\right)\right)\widehat{V}_{j_1,j_2}(t_i) = \left(1 + (1-\theta)\Delta_t\left(\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2\right)\right)\widehat{V}_{j_1,j_2}(t_{i+1}), \tag{2.59}$$

to be solved for the m_1m_2 interior points $\widehat{V}_{j_1,j_2}(t_i)$, $j_1 = 1,\ldots,m_1$, $j_2 = 1,\ldots,m_2$, given the values of $\widehat{V}_{j_1,j_2}(t_{i+1})$, and given appropriate boundary conditions at $j_1 = 0$, $j_1 = m_1 + 1$, $j_2 = 0$, and $j_2 = m_2 + 1$.

The scheme (2.59) represents a system of linear equations in m_1m_2 unknowns $\{\widehat{V}_{j_1,j_2}(t_i)\}$. When written out as a matrix equation (which requires us to arrange the various $\widehat{V}_{j_1,j_2}(t_i)$ in some order in a (m_1m_2)-dimensional vector), the matrix to be inverted is sparse but, unfortunately, no longer tri-diagonal. Solution of the system of equations by standard methods (e.g., Gauss-Jordan elimination or LU decomposition) is out of the question due

to the size of the matrix[20]. We can proceed in two ways: either we use a specialized sparse-matrix solver; or we attempt to redo the discretization (2.59) to make it computationally efficient. We personally prefer the second approach and shall outline one method in the next section. As for the first approach, we simply note that a good iterative sparse solver should be able to solve (2.59) in order $O((m_1 m_2)^{5/4})$ operations. See Saad [2003] for concrete algorithms.

2.10.2 The Alternating Direction Implicit (ADI) Method

The ADI method is an example of a so-called *operator splitting* method, where the simultaneous application of two operators (here $\widehat{\mathcal{L}}_1$ and $\widehat{\mathcal{L}}_2$) is split into two *sequential* operator applications. To illustrate the idea, set $\theta = \frac{1}{2}$ (Crank-Nicolson scheme) in (2.59) and approximate

$$\left(1 - \frac{1}{2}\Delta_t\left(\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2\right)\right) \approx \left(1 - \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_1\right)\left(1 - \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_2\right), \tag{2.60}$$

$$\left(1 + \frac{1}{2}\Delta_t\left(\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2\right)\right) \approx \left(1 + \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_1\right)\left(1 + \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_2\right). \tag{2.61}$$

It is easy to see[21] (and to verify, by a Taylor expansion) that the operators on the right-hand sides of these approximations have the same order truncation error as do the left-hand sides, namely $O(\Delta_t(\Delta_1^2 + \Delta_2^2 + \Delta_t^2))$. To the order of our original scheme, no accuracy is gained or lost in using the right-hand sides of (2.60)–(2.61). What is gained, however, is a considerable improvement in computational efficiency, originating in the fact that the resulting scheme

$$\begin{aligned}\left(1 - \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_1\right)\left(1 - \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_i)& \\ = \left(1 + \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_1\right)\left(1 + \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_{i+1})&\end{aligned} \tag{2.62}$$

can be *split* into the system

$$\left(1 - \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_1\right)U_{j_1,j_2} = \left(1 + \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_{i+1}), \tag{2.63}$$

$$\left(1 - \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_i) = \left(1 + \frac{1}{2}\Delta_t\widehat{\mathcal{L}}_1\right)U_{j_1,j_2}, \tag{2.64}$$

[20]Recall that the solution of a general linear system with $m_1 m_2$ unknowns is an $O(m_1^2 m_2^2)$ operation. For, say, m_1 and m_2 in the order of 100, this would involve around 1,000,000 times more work than what is required for a one-dimensional (tri-diagonal) scheme ($O(m)$).

[21]To those versed in operator notation, we notice that the right- and left-hand sides both approximate, to identical order, $\exp(\pm 0.5\Delta_t(\widehat{\mathcal{L}}_1 + \widehat{\mathcal{L}}_2))$.

where we have introduced an *intermediate value* U_{j_1,j_2}. The advantage of this decomposition is the fact that in each of (2.63) and (2.64), there is only one operator on the left-hand side, leading to simple tri-diagonal equation systems. To formalize this, first define

$$\mathbf{U}_1^{j_2} = (U_{1,j_2}, U_{2,j_2}, \ldots, U_{m_1,j_2})^\top .$$

Then, for a fixed value of j_2 we can write for the first step

$$\left(\mathbf{I} - \frac{1}{2}\Delta_t \mathbf{A}_1^{j_2}\left(\frac{t_{i+1}+t_i}{2}\right)\right)\mathbf{U}_1^{j_2} = \mathbf{M}_2^{j_2}\left(\frac{t_{i+1}+t_i}{2}\right), \tag{2.65}$$

where $\mathbf{A}_1^{j_2}$ is an $(m_1 \times m_1)$-dimensional tri-diagonal matrix of the same form as (2.11) (to get $\mathbf{A}_1^{j_2}$, basically freeze $x_2 = x_2^{j_2}$ and substitute μ_1 and γ_1 for μ and σ in the definition of the one-dimensional matrix $\mathbf{A}$). The m_1-dimensional vector $\mathbf{M}_2^{j_2}$ has components $M_{2,j_1}^{j_2}$, $j_1 = 1, \ldots, m_1$, given by

$$\begin{aligned} M_{2,j_1}^{j_2}\left(\frac{t_{i+1}+t_i}{2}\right) &= \left(1 + \frac{1}{2}\Delta_t \widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_{i+1}) \\ &= \frac{1}{2}\varsigma_{j_1,j_2}^{-}\widehat{V}_{j_1,j_2-1}(t_{i+1}) + \frac{1}{2}\varsigma_{j_1,j_2}^{+}\widehat{V}_{j_1,j_2+1}(t_{i+1}) \\ &\quad + \left(1 - \frac{1}{2}\varsigma_{j_1,j_2}\right)\widehat{V}_{j_1,j_2}(t_{i+1}), \end{aligned} \tag{2.66}$$

where we have defined

$$\varsigma_{j_1,j_2}^{\pm} \triangleq \frac{\Delta_t}{2\Delta_2^2}\left(\gamma_2\left(\frac{t_{i+1}+t_i}{2}, x_1^{j_1}, x_2^{j_2}\right)^2 \pm \Delta_2\mu_2\left(\frac{t_{i+1}+t_i}{2}, x_1^{j_1}, x_2^{j_2}\right)\right),$$

$$\varsigma_{j_1,j_2} \triangleq \frac{\Delta_t}{\Delta_2^2}\left(\gamma_2\left(\frac{t_{i+1}+t_i}{2}, x_1^{j_1}, x_2^{j_2}\right)^2 + \frac{1}{2}\Delta_2^2 r\left(\frac{t_{i+1}+t_i}{2}, x_1^{j_1}, x_2^{j_2}\right)\right).$$

For known values of $\widehat{V}(t_{i+1})$, (2.65) defines a simple tri-diagonal equation system which can be solved for $\mathbf{U}_1^{j_2}$ in $O(m_1)$ operations. Repeating the procedure above for $j_2 = 1, \ldots, m_2$ allows us to find U_{j_1,j_2} for all $j_1 = 1, \ldots, m_1$, $j_2 = 1, \ldots, m_2$, at a total computational cost of $O(m_1 m_2)$.

Turning to the second step of (2.63)–(2.64), we first fix j_1 and define

$$\widehat{\mathbf{V}}_2^{j_1}(t) = \left(\widehat{V}_{j_1,1}(t), \widehat{V}_{j_1,2}(t), \ldots, \widehat{V}_{j_1,m_2}(t)\right)^\top .$$

In the same fashion as earlier, we can then write

$$\left(\mathbf{I} - \frac{1}{2}\Delta_t \mathbf{A}_2^{j_1}\left(\frac{t_{i+1}+t_i}{2}\right)\right)\widehat{\mathbf{V}}_2^{j_1}(t_i) = \mathbf{M}_1^{j_1}\left(\frac{t_{i+1}+t_i}{2}\right), \tag{2.67}$$

where $\mathbf{A}_2^{j_1}$ is an $(m_2 \times m_2)$-dimensional tri-diagonal matrix and where the right-hand side vector now has components

$$M_{1,j_2}^{j_1}\left(\frac{t_{i+1}+t_i}{2}\right) = \left(1+\frac{1}{2}\Delta_t\widehat{\mathcal{L}}_1\right)U_{j_1,j_2}, \quad j_2 = 1,\ldots,m_2.$$

For brevity we omit writing out the $M_{1,j_2}^{j_1}$ (which will be similar to (2.66)), but just notice that the right-hand side of (2.67) is known after the first step of the ADI algorithm (above) is complete. For a given value of j_1, we can solve the tri-diagonal system (2.67) for $\widehat{\mathbf{V}}_2^{j_1}(t_i)$ in $O(m_2)$ operations. Looping over all m_1 different values of j_1, the full matrix of time t_i values $\widehat{V}_{j_1,j_2}(t_i)$, $j_1 = 1,\ldots,m_1$, $j_2 = 1,\ldots,m_2$, can then be found at a total computational cost of $O(m_1m_2)$.

The scheme outlined above is known as the *Peaceman-Rachford* scheme. As is the case for all ADI schemes, the scheme works by alternating the directions that are treated fully implicitly in the finite difference grid: in the first step, the x_1-direction is fully implicit and the x_2-direction is fully explicit, and in the second step the order is reversed. In effect, both spatial variables end up being discretized "semi-implicitly", i.e. similar to a Crank-Nicolson scheme, resulting in convergence order is $O(\Delta_1^2+\Delta_2^2+\Delta_t^2)$. We emphasize, however, that whereas a direct application of the Crank-Nicolson scheme will involve (if an efficient sparse-matrix solver is used) a computational cost of $O((m_1m_2)^{5/4})$ per time step, the computational cost of the Peaceman-Rachford ADI scheme is only $O(m_1m_2)$. A (tedious) von Neumann analysis reveals that the scheme is A-stable, but, like the Crank-Nicolson scheme, not strongly A-stable.

While the Peaceman-Rachford scheme is a classical example of an ADI scheme, there are many others. For instance, consider a theta-version of the *Douglas-Rachford* scheme:

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)U_{j_1,j_2} = \left(1+(1-\theta)\Delta_t\widehat{\mathcal{L}}_1+\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_{i+1}), \tag{2.68}$$

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_i) = U_{j_1,j_2}-\theta\Delta_t\widehat{\mathcal{L}}_2\widehat{V}_{j_1,j_2}(t_{i+1}), \tag{2.69}$$

where we understand that in $\widehat{\mathcal{L}}_1$ and $\widehat{\mathcal{L}}_2$ the PDE coefficients are to be evaluated at time $t_i^{i+1}(\theta)$. Again, notice how the scheme consists of two steps, each involving the solution of tri-diagonal sets of equations along only one of the x_1- or x_2-directions. The computational cost thus remains at $O(m_1m_2)$. It can be shown that the convergence order of this scheme is $O(\Delta_1^2+\Delta_2^2+1_{\{\theta\neq\frac{1}{2}\}}\Delta_t+\Delta_t^2)$ and it is A-stable for $\theta\geq\frac{1}{2}$, and strongly A-stable for $\theta>\frac{1}{2}$. By elimination of U_{j_1,j_2} we note that the unsplit version of the Douglas-Rachford scheme is

$$\begin{aligned}&\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_i)\\ &\qquad=\left(\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)+\Delta_t\widehat{\mathcal{L}}_1+\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_{i+1}).\end{aligned}$$

It is not difficult to see that this approximates (2.59) to second order.

2.10.3 Boundary Conditions and Other Issues

The fact that ADI schemes reduce to solving sequences of matrix systems identical to the ones arising in the one-dimensional case is convenient, in the sense that many of the issues we have encountered for one-dimensional finite difference grids (oscillations, stability, convection dominance, etc.) and their remedies (smoothing, non-equidistant discretization, upwinding, etc.) carry over to the ADI setting with only minor modifications. Consider for instance the issue of applying spatial boundary conditions along the edges of the (x_1, x_2) domain, which we have so far not discussed. As for the one-dimensional PDEs, the most convenient way to express such boundary conditions is typically by imposing conditions on derivatives, like $\partial^2 V(t, x_1^0, x_2^{j_2})/\partial x_1^2 = \partial V(t, x_1^0, x_2^{j_2})/\partial x_1$ and so forth. For the Peaceman-Rachford scheme, say, such conditions can be incorporated directly into (2.65) and (2.67) by altering the matrices $\mathbf{A}_1^{j_2}$ and $\mathbf{A}_2^{j_1}$, as well as the boundary elements of $\mathbf{M}_1^{j_1}$ and $\mathbf{M}_2^{j_2}$, in the manner outlined in Section 2.2.1. If instead we wish to impose Dirichlet boundary conditions, we need to add corrective terms to the tri-diagonal systems, as in (2.19). To complete the first part of the split scheme, this then requires us to establish what boundary terms are needed for the intermediate quantity U_{j_1,j_2}, i.e. we must define $U_{j_1,0}$ and U_{j_1,m_2+1} for $j_1 = 1, \ldots, m_1$, as well as U_{0,j_2} and U_{m_1+1,j_2} for $j_2 = 1, \ldots, m_2$. While U_{j_1,j_2} is a purely mathematic construct, sometimes it is adequate to think of U_{j_1,j_2} as a proxy for V_{j_1,j_2} evaluated at $t_i^{i+1}(\theta)$, which obviously makes determination of boundary conditions straightforward. For maximum precision, however, we should use the ADI equations themselves to express the boundary conditions of U directly in terms of boundary conditions for $V(t_i)$ and $V(t_{i+1})$. Here, the Douglas-Rachford scheme is particularly easy to deal with, as a rearrangement of (2.69) directly relates U_{j_1,j_2} to $\widehat{V}_{j_1,j_2}(t_i)$ and $\widehat{V}_{j_1,j_2}(t_{i+1})$,

$$U_{j_1,j_2} = \left(1 - \theta \Delta_t \widehat{\mathcal{L}}_2\right) \widehat{V}_{j_1,j_2}(t_i) + \theta \Delta_t \widehat{\mathcal{L}}_2 \widehat{V}_{j_1,j_2}(t_{i+1}).$$

The Peaceman-Rachford scheme requires some further manipulations to express U in terms of $V(t_i)$ and $V(t_{i+1})$; see Mitchell and Griffiths [1980] for the details.

2.11 Two-Dimensional PDE with Mixed Derivatives

Consider now the case where the 2-dimensional PDE (2.58) has a mixed partial derivative,

$$\frac{\partial V}{\partial t} + (\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_{1,2}) V = 0, \tag{2.70}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are as in (2.58), and where

$$\mathcal{L}_{1,2} = s_{1,2}(t,x)\frac{\partial^2}{\partial x_1 \partial x_2} \triangleq \rho(t,x)\gamma_1(t,x)\gamma_2(t,x)\frac{\partial^2}{\partial x_1 \partial x_2}. \tag{2.71}$$

The quantity $\rho(t,x)$ is the instantaneous correlation between the processes $x_1(t)$ and $x_2(t)$ in (2.57), i.e. $\rho(t,x) \in [-1,1]$.

The presence of $\mathcal{L}_{1,2}$ prevents a direct application of the ADI methods in Section 2.10.2, since the mixed operator $\mathcal{L}_{1,2}$ is not amenable to operator splitting. We shall demonstrate two ways to overcome this problem: a) orthogonalization of the PDE; and b) predictor-corrector schemes.

2.11.1 Orthogonalization of the PDE

The idea here is to introduce new variables $y_1(t,x_1,x_2)$ and $y_2(t,x_1,x_2)$ such that the PDE loses its mixed derivative term when stated in terms of these variables. To demonstrate this idea, assume first that $\rho(t,x)$, $\gamma_1(t,x)$, and $\gamma_2(t,x)$ are all functions of time only and independent of x. Then define, say,

$$y_1(t,x_1,x_2) = x_1, \tag{2.72}$$

$$y_2(t,x_1,x_2) = -\rho(t)\frac{\gamma_2(t)}{\gamma_1(t)}x_1 + x_2 \triangleq a(t)x_1 + x_2, \tag{2.73}$$

where we must assume that $\gamma_1(t) \neq 0$ for all t.

Lemma 2.11.1. *Consider the PDE (2.70) subject to the terminal value condition $V(T,x) = g(x)$. Define $y = (y_1,y_2)^\top$ and $v(t,y) = V(t,x)$. With the variable change defined in (2.72)–(2.73), v satisfies*

$$\begin{aligned}\frac{\partial v}{\partial t} + \mu_1^y(t,y)\frac{\partial v}{\partial y_1} + \mu_2^y(t,y)\frac{\partial v}{\partial y_2} + \frac{1}{2}\gamma_1(t)^2\frac{\partial^2 v}{\partial y_1^2} \\ + \frac{1}{2}\left(1-\rho(t)^2\right)\gamma_2(t)^2\frac{\partial^2 v}{\partial y_2^2} - r(t,y_1,y_2-a(t)y_1) = 0,\end{aligned} \tag{2.74}$$

where

$$\mu_1^y(t,y) \triangleq \mu_1(t,x_1,x_2) = \mu_1(t,y_1,y_2-a(t)y_1), \tag{2.75}$$

$$\begin{aligned}\mu_2^y(t,y) &\triangleq \frac{da(t)}{dt}x_1 + a(t)\mu_1(t,x_1,x_2) + \mu_2(t,x_1,x_2) \\ &= \frac{da(t)}{dt}y_1 + a(t)\mu_1^y(t,y) + \mu_2(t,y_1,y_2-a(t)y_1).\end{aligned} \tag{2.76}$$

The equation (2.74) is subject to the terminal value condition $v(T,y_1,y_2) = g(x_1,x_2) = g(y_1,y_2-a(T)y_1)$.

Proof. While the result can be established by the usual mechanics of ordinary calculus, we will take the opportunity to show how stochastic calculus can

also conveniently prove results of this type. Going back to the processes underlying the PDE (see (2.57)), we write

$$dx_1(t) = \mu_1(t,x)\,dt + \gamma_1(t)\,dW_1(t), \tag{2.77}$$

$$dx_2(t) = \mu_2(t,x)\,dt + \gamma_2(t)\left(\rho(t)\,dW_1(t) + \sqrt{1-\rho(t)^2}\,dW_2(t)\right), \tag{2.78}$$

for independent scalar Brownian motions $W_1(t)$ and $W_2(t)$; this is easily seen to generate the correct correlation $\rho(t)$ between x_1 and x_2. An application of Ito's lemma then shows that the processes for y_1 and y_2 are

$$\begin{aligned}
dy_1(t) &= dx_1(t) = \mu_1(t,x)\,dt + \gamma_1(t)\,dW_1(t),\\
dy_2(t) &= \frac{da(t)}{dt}x_1(t)\,dt + a(t)\mu_1(t,x)\,dt + a(t)\gamma_1(t)\,dW_1(t)\\
&\quad + \mu_2(t,x)\,dt + \gamma_2(t)\left(\rho(t)\,dW_1(t) + \sqrt{1-\rho(t)^2}\,dW_2(t)\right)\\
&= \left(\frac{da(t)}{dt}x_1(t) + a(t)\mu_1(t,x) + \mu_2(t,x)\right)dt\\
&\quad + \gamma_2(t)\sqrt{1-\rho(t)^2}\,dW_2(t).
\end{aligned}$$

With the definitions (2.75)–(2.76), this becomes simply

$$dy_1(t) = \mu_1^y\left(t, y(t)\right)\,dt + \gamma_1(t)\,dW_1(t), \tag{2.79}$$

$$dy_2(t) = \mu_2^y\left(t, y(t)\right)\,dt + \gamma_2(t)\sqrt{1-\rho(t)^2}\,dW_2(t). \tag{2.80}$$

Equations (2.79)–(2.80) define a Markov SDE in $y_1(t)$ and $y_2(t)$ where, importantly, the Brownian motions on $y_1(t)$ and $y_2(t)$ are now *independent*. Writing $V(t,x) = v(t,y)$, it then follows immediately from the backward Kolmogorov equation (see Section 1.8) that v satisfies the PDE (2.74). □

Through the chosen transformation (2.72)–(2.73), our original PDE has now been put into a form where we can immediately apply the ADI schemes outlined in Section 2.10.2.

In performing the orthogonalization of the PDE in Lemma 2.11.1 we relied on $\rho(t,x)$, $\gamma_1(t,x)$, and $\gamma_2(t,x)$ all being independent of x. This can often be relaxed. Consider for instance the case where $\rho(t,x) = \rho(t)$, $\gamma_1(t,x) = \gamma_1(t,x_1)$, and $\gamma_2(t,x) = \gamma_2(t,x_2)$; here the correlation ρ is still assumed deterministic, but we now allow for some (though not full) x-dependence in γ_1 and γ_2. Assuming that $\gamma_1(t,x_1) > 0$ and $\gamma_2(t,x_2) > 0$ we can introduce new variables

$$z_1\left(t,x_1\right) = \int \frac{1}{\gamma_1(t,x_1)}dx_1, \tag{2.81}$$

$$z_2\left(t,x_2\right) = \int \frac{1}{\gamma_2(t,x_2)}dx_2. \tag{2.82}$$

Applying Ito's lemma to (2.77)–(2.78) we see that

$$dz_1(t,x_1) = \left(-\int \frac{\partial \gamma_1(t,x_1)}{\partial t} \frac{1}{\gamma_1(t,x_1)^2} dx_1 + \frac{\mu_1(t,x)}{\gamma_1(t,x_1)} - \frac{1}{2}\frac{\partial \gamma_1(t,x_1)}{\partial x_1} \right) dt + dW_1(t) \quad (2.83)$$

and

$$dz_2(t,x_2) = \left(-\int \frac{\partial \gamma_2(t,x_2)}{\partial t} \frac{1}{\gamma_2(t,x_2)^2} dx_2 + \frac{\mu_2(t,x)}{\gamma_2(t,x_2)} - \frac{1}{2}\frac{\partial \gamma_2(t,x_2)}{\partial x_2} \right) dt + \rho(t)\, dW_1(t) + \sqrt{1-\rho(t)^2}\, dW_2(t). \quad (2.84)$$

As we assumed that $\gamma_1(t,x_1) > 0$ and $\gamma_2(t,x_2) > 0$, the functions z_1 and z_2 are increasing in x_1 and x_2, respectively, and are thereby invertible. As such, we can rewrite (2.83)–(2.84) in the more appealing form

$$dz_1(t,x_1) = \mu_1^z\,(t,z_1,z_2)\, dt + dW_1(t),$$
$$dz_2(t,x_1) = \mu_2^z\,(t,z_1,z_2)\, dt + \rho(t)\, dW_1(t) + \sqrt{1-\rho(t)^2}\, dW_2(t).$$

Through the transformation (2.81)–(2.82), we have reduced our original system to one where the coefficients on $W_1(t)$ and $W_2(t)$ are no longer state-dependent, similar to the case that lead to Lemma 2.11.1. We can now proceed with another variable transformation, as in (2.72)–(2.73), to orthogonalize the system and prepare it for an application of the ADI method.

While the orthogonalization method outlined here can be very effective on a range of practical problems, it suffers from a few drawbacks. Most obviously, the method is not completely general and requires a certain structure on the parameters of the PDE. Another drawback is that the introduction of a time-dependent transformation on one or more variables (Lemma 2.11.1) often makes the alignment of the finite difference grid along (time-independent) critical level points in x-space impossible. Also, the introduction of terms like $y_1 da(t)/dt$ in the drift of y_2 (see (2.76)) can be problematic, particularly if the functions $\gamma_1(t)$ and $\gamma_2(t)$ are not smooth. For instance, it is not unlikely that $y_1 da(t)/dt$ will locally be of such magnitude that upwinding will be necessary to prevent oscillations; see Section 2.6.1. Further, we note that inversion of the transformations (2.81)–(2.82) will not always be possible to perform analytically and may require numerical (root-search) work, complicating the scheme and potentially slowing it down. Finally, as we shall highlight in future chapters, maintaining the "continuity" of a numerical scheme with respect to input parameters is of critical importance for the smoothness of risk sensitivities. Such continuity is difficult to ensure if complicated transformations are applied to model variables. So, in the end, we recommend formulating the PDEs in terms of financially meaningful variables, avoiding excessive transformations, and relying on methods such as developed in the next section when dealing with mixed derivatives and other numerical complications.

2.11.2 Predictor-Corrector Scheme

In this section we shall consider a completely general method for handling mixed derivatives in two-dimensional PDEs. While a bit slower than the method outlined in Section 2.11.1, it does not involve any variable transformations and, by extension, does not suffer from the drawbacks associated with such transformations. As a first step, consider the discretization of the mixed derivative $\partial^2 V/\partial x_1 \partial x_2$. There are a few possibilities (see Mitchell and Griffiths [1980]), but we shall just use

$$\begin{aligned}\delta_{x_1x_2} V_{j_1,j_2}(t) &= \delta_{x_1}\delta_{x_2} V_{j_1,j_2}(t) \\ &= \frac{V_{j_1+1,j_2+1}(t) - V_{j_1+1,j_2-1}(t) - V_{j_1-1,j_2+1}(t) + V_{j_1-1,j_2-1}(t)}{4\Delta_1\Delta_2}. \end{aligned} \tag{2.85}$$

Extensions to non-equidistant grids follow directly from (2.27) and the relation $\delta_{x_1x_2} V_{j_1,j_2}(t) = \delta_{x_1}\delta_{x_2} V_{j_1,j_2}(t)$. As we have not encountered mixed difference operators before, for completeness we show the following lemma.

Lemma 2.11.2. *For the discrete operator (2.85) we have*

$$\delta_{x_1x_2} V_{j_1,j_2}(t) = \frac{\partial^2 V(t, x_1^{j_1}, x_2^{j_2})}{\partial x_1 \partial x_2} + O\left(\Delta_1^2 + \Delta_2^2\right).$$

Proof. A Taylor expansion of $V(t,x)$ around the point $x = (x_1^{j_1}, x_2^{j_2})^\top$ gives

$$\begin{aligned}V_{j_1+1,j_2\pm1}(t) = V_{j_1,j_2}(t) &+ \Delta_1 \frac{\partial V}{\partial x_1} \pm \Delta_2 \frac{\partial V}{\partial x_2} + \frac{1}{2}\Delta_1^2 \frac{\partial^2 V}{\partial x_1^2} + \frac{1}{2}\Delta_2^2 \frac{\partial^2 V}{\partial x_2^2} \\ &\pm \Delta_1\Delta_2 \frac{\partial^2 V}{\partial x_1 \partial x_2} + \frac{1}{6}\Delta_1^3 \frac{\partial^3 V}{\partial x_1^3} \pm \frac{1}{6}\Delta_2^3 \frac{\partial^3 V}{\partial x_2^3} \\ &+ \frac{1}{2}\Delta_1\Delta_2^2 \frac{\partial^3 V}{\partial x_1 \partial x_2^2} \pm \frac{1}{2}\Delta_1^2\Delta_2 \frac{\partial^3 V}{\partial x_1^2 \partial x_2} + \ldots, \\ V_{j_1-1,j_2\pm1}(t) = V_{j_1,j_2}(t) &- \Delta_1 \frac{\partial V}{\partial x_1} \pm \Delta_2 \frac{\partial V}{\partial x_2} + \frac{1}{2}\Delta_1^2 \frac{\partial^2 V}{\partial x_1^2} + \frac{1}{2}\Delta_2^2 \frac{\partial^2 V}{\partial x_2^2} \\ &\mp \Delta_1\Delta_2 \frac{\partial^2 V}{\partial x_1 \partial x_2} - \frac{1}{6}\Delta_1^3 \frac{\partial^3 V}{\partial x_1^3} \pm \frac{1}{6}\Delta_2^3 \frac{\partial^3 V}{\partial x_2^3} \\ &- \frac{1}{2}\Delta_1\Delta_2^2 \frac{\partial^3 V}{\partial x_1 \partial x_2^2} \pm \frac{1}{2}\Delta_1^2\Delta_2 \frac{\partial^3 V}{\partial x_1^2 \partial x_2} + \ldots.\end{aligned}$$

A little thought then shows that

$$\begin{aligned}V_{j_1+1,j_2+1}(t) - V_{j_1+1,j_2-1}(t) - V_{j_1-1,j_2+1}(t) + V_{j_1-1,j_2-1}(t) \\ = 4\Delta_1\Delta_2 \frac{\partial^2 V}{\partial x_1 \partial x_2} + O\left(\Delta_1^3\Delta_2 + \Delta_1\Delta_2^3\right),\end{aligned}$$

as error terms of order Δ_1^4, Δ_2^4, and $\Delta_1^2\Delta_2^2$ will cancel. The result follows. □

Equipped with (2.85), we can approximate the operator $\mathcal{L}_{1,2}$ in (2.71) as

$$\widehat{\mathcal{L}}_{1,2}V_{j_1,j_2}(t) \triangleq \rho\left(t, x_1^{j_1}, x_2^{j_2}\right)\gamma_1\left(t, x_1^{j_1}, x_2^{j_2}\right)\gamma_2\left(t, x_1^{j_1}, x_2^{j_2}\right)\delta_{x_1x_2}V_{j_1,j_2}(t),$$

which is accurate to order $O(\Delta_1^2+\Delta_2^2)$. The first easy way to modify our ADI scheme to incorporate $\widehat{\mathcal{L}}_{1,2}$ is to treat the mixed derivative fully explicitly. In the Douglas-Rachford scheme (2.68)–(2.69), for instance, we thus modify the right-hand side of the first step as follows:

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)U_{j_1,j_2} = \left(1+(1-\theta)\Delta_t\widehat{\mathcal{L}}_1+\Delta_t\widehat{\mathcal{L}}_2+\Delta_t\widehat{\mathcal{L}}_{1,2}\right)\widehat{V}_{j_1,j_2}(t_{i+1}), \tag{2.86}$$

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_i) = U_{j_1,j_2}-\theta\Delta_t\widehat{\mathcal{L}}_2\widehat{V}_{j_1,j_2}(t_{i+1}). \tag{2.87}$$

The addition of $\widehat{\mathcal{L}}_{1,2}$ this way clearly preserves the ADI structure of the scheme which will continue to involve only sequences of tri-diagonal linear equations. However, having, in effect, only a one-sided time-differencing of the mixed derivative term will lower the convergence order of the time step to $O(\Delta_t)$, irrespective of the choice of θ.

To change the time at which the mixed operator $\widehat{\mathcal{L}}_{1,2}$ is evaluated, consider using a *predictor-corrector* scheme, where the results of (2.86)–(2.87) are re-used in a one-time[22] iteration. Specifically, we write, for some $\lambda \in [0,1]$,

Predictor:

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)U^{(1)}_{j_1,j_2} = \left(1+(1-\theta)\Delta_t\widehat{\mathcal{L}}_1+\Delta_t\widehat{\mathcal{L}}_2+\Delta_t\widehat{\mathcal{L}}_{1,2}\right)\widehat{V}_{j_1,j_2}(t_{i+1}), \tag{2.88}$$

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)U^{(2)}_{j_1,j_2} = U^{(1)}_{j_1,j_2}-\theta\Delta_t\widehat{\mathcal{L}}_2\widehat{V}_{j_1,j_2}(t_{i+1}). \tag{2.89}$$

Corrector:

$$\begin{aligned}\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)Z^{(1)}_{j_1,j_2} = \Big(&1+(1-\theta)\Delta_t\widehat{\mathcal{L}}_1+\Delta_t\widehat{\mathcal{L}}_2\\ &+(1-\lambda)\Delta_t\widehat{\mathcal{L}}_{1,2}\Big)\widehat{V}_{j_1,j_2}(t_{i+1})+\lambda\Delta_t\widehat{\mathcal{L}}_{1,2}U^{(2)}_{j_1,j_2},\end{aligned} \tag{2.90}$$

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)\widehat{V}_{j_1,j_2}(t_i) = Z^{(1)}_{j_1,j_2}-\theta\Delta_t\widehat{\mathcal{L}}_2\widehat{V}_{j_1,j_2}(t_{i+1}). \tag{2.91}$$

[22] We can run the iteration more than once if desired, but a single iteration will normally suffice.

Notice how the Douglas-Rachford scheme is first run once, in (2.88)–(2.89), to yield a first guess (a "predictor"), $U^{(2)}_{j_1,j_2}$, for the time t_i value $V_{j_1,j_2}(t_i)$. In a second run of the Douglas-Rachford scheme, in (2.90)–(2.91), this guess is used as a "corrector" to affect the time at which $\widehat{\mathcal{L}}_{1,2}$ is evaluated, by applying this operator to $(1-\lambda)\widehat{V}_{j_1,j_2}(t_{i+1}) + \lambda U^{(2)}_{j_1,j_2}$; when $\lambda = \frac{1}{2}$ we effectively center the time-differencing of the mixed term. The scheme now relies on three intermediate variables, $U^{(1)}_{j_1,j_2}$, $U^{(2)}_{j_1,j_2}$, and $Z^{(1)}_{j_1,j_2}$.

The combined predictor-corrector scheme above (in a slightly less general form, with $\Delta_1 = \Delta_2$) was suggested by Craig and Sneyd [1988]. It can be shown that the scheme has convergence order

$$O\left((\Delta_1+\Delta_2)^2 + 1_{\{\theta\neq\frac{1}{2}\}}\Delta_t + 1_{\{\lambda\neq\frac{1}{2}\}}\Delta_t + \Delta_t^2\right),$$

so second order convergence in the time domain is still achievable by setting $\theta = \lambda = \frac{1}{2}$. The scheme will be A-stable for $\theta \geq \frac{1}{2}$ and $\frac{1}{2} \leq \lambda \leq \theta$. The computational cost of the predictor-corrector is clearly still $O(m_1 m_2)$ per time step, as both the predictor and corrector schemes have $O(m_1 m_2)$ cost per time-step. Even though the standard Douglas-Rachford scheme is effectively run twice, we should point out that when intelligently implemented, (2.88)–(2.91) is typically only about 30-40% slower than the Douglas-Rachford scheme, as a number of results from the predictor step can be cached and reused in the corrector step.

As for the standard ADI grids, extensions to non-equidistant grids are straightforward using the techniques in Section 2.4. Boundary conditions in the x-domain are imposed along the lines outlined in Section 2.10.3.

2.12 PDEs of Arbitrary Order

We now turn our attention back to the general p-dimensional PDE (2.56). To prepare for a numerical scheme, let us rewrite the PDE as follows:

$$\frac{\partial V}{\partial t} + \sum_{h=1}^{p} \mathcal{L}_h V + \sum_{h=1}^{p}\sum_{l=h+1}^{p} \mathcal{L}_{h,l} V = 0, \tag{2.92}$$

where

$$\mathcal{L}_h = \mu_h(t,x)\frac{\partial}{\partial x_h} + \frac{1}{2}s_{h,h}(t,x)\frac{\partial^2}{\partial x_h^2} - p^{-1}r(t,x),$$
$$\mathcal{L}_{h,l} = s_{h,l}(t,x)\frac{\partial^2}{\partial x_h \partial x_l}.$$

The method we present here for solution of (2.92) is a p-dimensional version of the predictor-corrector scheme outlined above. The extension

is straightforward and we simply list it here without further discussion; see Craig and Sneyd [1988] for additional background. To simplify notation, we have omitted sub-indices everywhere (i.e., $\widehat{V}(t_i)$ is used instead of $\widehat{V}_{j_1,j_2,\ldots,j_p}(t_i)$).

Predictor:

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)U^{(1)}$$
$$=\Delta_t\left(\Delta_t^{-1}+(1-\theta)\widehat{\mathcal{L}}_1+\sum_{h=2}^{p}\widehat{\mathcal{L}}_h+\sum_{h=1}^{p}\sum_{l=h+1}^{p}\widehat{\mathcal{L}}_{h,l}\right)\widehat{V}(t_{i+1}),$$

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)U^{(2)}=U^{(1)}-\theta\Delta_t\widehat{\mathcal{L}}_2\widehat{V}(t_{i+1}),$$
$$\vdots$$
$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_p\right)U^{(p)}=U^{(p-1)}-\theta\Delta_t\widehat{\mathcal{L}}_p\widehat{V}(t_{i+1}).$$

Corrector:

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_1\right)Z^{(1)}$$
$$=\Delta_t\left(\Delta_t^{-1}+(1-\theta)\widehat{\mathcal{L}}_1+\sum_{h=2}^{p}\widehat{\mathcal{L}}_h\right.$$
$$\left.+(1-\lambda)\sum_{h=1}^{p}\sum_{l=h+1}^{p}\widehat{\mathcal{L}}_{h,l}\right)\widehat{V}(t_{i+1})+\lambda\Delta_t\sum_{h=1}^{p}\sum_{l=h+1}^{p}\widehat{\mathcal{L}}_{h,l}U^{(p)},$$

$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_2\right)Z^{(2)}=Z^{(1)}-\theta\Delta_t\widehat{\mathcal{L}}_2\widehat{V}(t_{i+1}),$$
$$\vdots$$
$$\left(1-\theta\Delta_t\widehat{\mathcal{L}}_p\right)\widehat{V}(t_i)=Z^{(p-1)}-\theta\Delta_t\widehat{\mathcal{L}}_p\widehat{V}(t_{i+1}).$$

With m_h points in the x_h-direction, $h=1,\ldots,p$, the computational cost of the predictor-corrector scheme is $O(\prod_{h=1}^{p}m_h)$. For $p\le 3$, sufficient conditions for A-stability are $\theta\ge\frac{1}{2}$ and $\frac{1}{2}\le\lambda\le\theta$. For $p\ge 4$, sufficient conditions are $\theta\le\frac{1}{2}$ and

$$\frac{1}{2}\le\lambda\le\frac{p^{p-1}}{(p-1)^p}\theta.$$

See Craig and Sneyd [1988] for a proof. Convergence is similar to the two-dimensional case.

As a final comment, let us note that as dimensionality increases, the computational complexity of an iterative sparse solver will start approaching that of ADI. Specifically, for a p-dimensional problem, the complexity of the former is $O(m_{\text{total}})$ and for the latter $O(m_{\text{total}}^{(2p+1)/2p})$, with $m_{\text{total}} = m_1 \cdot m_2 \cdot \ldots \cdot m_p$.

3

Monte Carlo Methods

While the finite difference method is flexible and powerful, it has a number of limitations. First, its usage is restricted to problems where the state variable dynamics are Markovian. Second, for strongly path-dependent problems, the method often does not apply. And third, it is unsuited for problems where the dimension of the underlying vector of state variables is high. To expand on the last point, recall from Section 2.9 that the (ADI) finite difference method applied to a p-dimensional problem has computational complexity $O(m^p)$ per time step, where m is the average number of spatial points per dimension. The exponential growth in p — the "curse of dimensionality" — is typical of grid-based methods and prevents the practical usage of the method for p larger than about 4 or 5.

In this chapter, we study the *Monte Carlo method*, a numerical technique where the computational effort grows only linearly in the problem dimension p. While convergence of the Monte Carlo method is relatively slow, it is nearly always the method of choice for high-dimensional pricing problems. Compared to finite difference methods, Monte Carlo methods are easy to apply to problems with non-Markovian dynamics as well as strong path-dependency in the payout. On the other hand, as Monte Carlo methods inherently run forward in time, dynamic programming techniques are challenging to implement, making Monte Carlo pricing of American and Bermudan options significantly more involved than for the naturally backward-working finite difference method.

3.1 Fundamentals

Consider a European-style derivative V with time T payout $V(T) = g(T)$, where $g(T)$ is an $\mathcal{F}_T$-measurable (and integrable) random variable. Where finite difference methods start with a PDE representation of the price of a contingent claim at times $t < T$, the starting point for the Monte Carlo method is the basic martingale relation (see (1.15))

$$V(t) = N(t)\mathrm{E}_t^{\mathrm{Q}^N}\left(g(T)/N(T)\right), \tag{3.1}$$

where $N(\cdot)$ is a numeraire and Q^N is the measure induced by $N(\cdot)$. To evaluate this expression numerically, we need a numerical technique to compute expectations of a random variable. For this, we turn to the law of large numbers:

Theorem 3.1.1 (Strong Law of Large Numbers). *Let $Y_1, Y_2, \ldots$ be a sequence of independent identically distributed (i.i.d.) random variables with expectation $\mu < \infty$. Define the* sample mean

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{3.2}$$

Then

$$\lim_{n\to\infty} \overline{Y}_n = \mu, \quad a.s.$$

This result forms the basis for the *Monte Carlo method*, which computes the expectation in (3.1) by simply i) generating independent realizations of $g(T)/N(T)$ under Q^N; and ii) forming their average. Specifically, let $g_1/N_1, \ldots, g_n/N_n$ denote n independent samples from the distribution of $g(T)/N(T)$, conditional on $\mathcal{F}_t$. Then our Monte Carlo estimator for $V(t)$ is the sample mean

$$\overline{V}(t) = N(t)\frac{1}{n}\sum_{i=1}^{n} g_i/N_i. \tag{3.3}$$

We shall delve into how to generate samples from the distribution of $g(T)/N(T)$ shortly, but before doing so let us consider the expected convergence rate of the Monte Carlo method as n is increased. The key result is here the central limit theorem:

Theorem 3.1.2 (Central Limit Theorem). *Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables with expectation μ and standard deviation $\sigma < \infty$. Let the sample mean be defined as in (3.2). Then, for $n \to \infty$,*

$$\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ is a standard Gaussian distribution and $\xrightarrow{d}$ denotes convergence in distribution[1]*. Further, if we define*

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2},$$

[1]Recall that a sequence of variables X_n with cumulative distribution functions F_n converge in distribution to a random variable X with distribution F if $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ at which $F(x)$ is continuous.

then also

$$\frac{\overline{Y}_n - \mu}{s_n/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Define the Gaussian percentile u_γ as $\Phi(u_\gamma) = 1 - \gamma$, where Φ is the Gaussian cumulative distribution function. From Theorem 3.1.2, and from the definition of convergence in distribution (see footnote 1), the probability that the confidence interval

$$\left[\overline{V}(t) - u_{\gamma/2} \cdot s_n/\sqrt{n}, \overline{V}(t) + u_{\gamma/2} \cdot s_n/\sqrt{n}\right] \tag{3.4}$$

fails to include the true value $V(t)$ approaches γ for large n. Here

$$s_n \triangleq \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{g_i N(t)}{N_i} - \overline{V}(t)\right)^2},$$

with the quantity $s_n/\sqrt{n}$ known as the *standard error*. For given γ, the rate at which the confidence interval for $V(t)$ contracts is $O(n^{-\frac{1}{2}})$. This is relatively slow: to reduce the width of the interval by a factor of 2, n must increase by a factor of 4. On the other hand, we notice that the (asymptotic) convergence rate only depends on n, not on the specifics of the g_i's. In particular, if $g(T) = g(X(T))$ where X is p-dimensional, the asymptotic convergence rate is independent of p. As we shall see shortly, in most applications the work required to generate samples of $g(X(T))$ is (at most) linear in p.

3.1.1 Generation of Random Samples

At the most basic level, the Monte Carlo method requires the ability to draw independent realizations of a scalar random variable Z with a specified cumulative distribution function $F(z) = \mathrm{P}(Z \leq z)$, where P is a probability measure. On a computer, the starting point for this exercise is normally a *pseudo-random number generator*, a software program that will generate a sequence of numbers uniformly distributed on $[0,1]$ (i.e. from $\mathcal{U}(0,1)$). Press et al. [1992] list a number of generators producing sequences of uniform numbers $u_1, u_2, \ldots$ from iterative relationships of the form

$$\begin{aligned} I_{i+1} &= (aI_i + c) \bmod(m), \\ u_{i+1} &= I_{i+1}/m. \end{aligned}$$

The externally specified starting point I_0 is the *seed* of the random number generator. In this so-called *general linear congruential generator*, the choice of the *multiplier* a, the *modulus* m, and the *increment* c must be done

with great care to ensure that the period length of the generator is large[2] and that the resulting algorithm is efficient on a computer. The latter, for instance, can be accomplished be setting m to be a power of 2 such that the modulo operation can be done by bit-shifting. For detailed discussion and a number of concrete algorithms (including computer code), we refer to Press et al. [1992]. The algorithms in Press et al. [1992] should suffice for most fixed income applications, but we should note the existence of more sophisticated methods that (theoretically, at least) have better performance than linear congruential generators. For instance, the so-called *Mersenne twister* proposed in Matsumoto and Nishimura [1998] has become popular, especially the specific variant MT19937 which has a period of $2^{19937} - 1$. For an extensive survey of pseudo-random number generators, see L'Ecuyer [1994].

So far we have only discussed techniques to generate $\mathcal{U}(0,1)$ numbers, but many methods exist to convert uniformly distributed numbers into draws from the distribution F of Z. We cover a few important techniques next.

3.1.1.1 Inverse Transform Method

The idea of the inverse transform method is straightforward. Let U be a random variable uniformly distributed on $[0,1]$, and consider setting

$$Z = F^{-1}(U), \tag{3.5}$$

where we assume that F^{-1} is well-defined, for all but a finite number of points[3]. As desired,

$$\mathrm{P}\left(Z \leq z\right) = \mathrm{P}\left(F^{-1}(U) \leq z\right) = \mathrm{P}\left(U \leq F(z)\right) = F(z),$$

where the last equality follows from the property of uniformly distributed random variables. The inverse transform method (3.5) is quite general, but its practical usefulness hinges on being able to compute F^{-1} fast. Many distributions allow for closed-form inversion; this includes the *exponential distribution* where $F(z) = 1 - e^{-z\lambda}$ for some positive constant λ, and the *Cauchy distribution* where $F(z) = 1/2 + (1/\pi)\arctan((z-t)/s)$ for constants t and $s > 0$.

For the important case of the Gaussian distribution, no closed-form expression for the inverse distribution exists. Nevertheless, the inverse transform

[2]Note that if a number $I_k = I_i$, the sequences starting from I_k and I_i are identical. In practice, we would want the generator to have *full period*, in the sense that the sequence would produce $m - 1$ distinct values before repeating the sequence.

[3]For discrete random variables, the distribution function is discontinuous around each of the possible (discrete) outcomes of Z. We can handle this by simply defining $F^{-1}(u) = \inf\{z : F(z) \geq u\}$.

method can still be applied as fast and extremely accurate approximations for Φ^{-1} exist. For instance, Beasley and Springer [1977] suggest the rational approximation

$$\Phi^{-1}(x) \approx \frac{\sum_{i=0}^{3} a_i \left(x - \frac{1}{2}\right)^{2i+1}}{1 + \sum_{i=0}^{3} b_i \left(x - \frac{1}{2}\right)^{2i}}, \quad 0.5 \leq x \leq 0.92, \tag{3.6}$$

for constants a_i, b_i, $i = 0, \ldots, 3$, listed in Appendix 3.A. For values of x greater than 0.92, Moro [1995] proposes the approximation

$$\Phi^{-1}(x) \approx \sum_{i=1}^{8} c_i \left[\ln\left(-\ln(1-x)\right)\right]^i, \quad 0.92 \leq x < 1, \tag{3.7}$$

for constants c_i, $i = 0, \ldots, 8$, given in Appendix 3.A. Taken together, (3.6) and (3.7) provide an approximation valid for $0.5 \leq x < 1$; when $0 < x < 0.5$ we can compute $\Phi^{-1}(x)$ by symmetry: $\Phi^{-1}(1-x) = -\Phi^{-1}(x)$. The precision of (3.6)–(3.7) is excellent[4], with the error less than 3×10^{-9} for x in the range $x \in [\Phi(-7), \Phi(7)]$. For alternative algorithms, see for instance Acklam [2003] and Wichura [1988].

Well-known alternative methods for sampling in the Gaussian distribution include the *Box-Muller method* and the related *Marsaglia polar method* (see Press et al. [1992]).

3.1.1.2 Acceptance-Rejection Method

In cases where F^{-1} is cumbersome to compute, the so-called *acceptance-rejection method* may be preferable. To describe the method, suppose that we want to sample from a density $f(z) = dF(z)/dz$, and further suppose that we have a good method to sample from a density $e(z)$, where

$$e(z)c \geq f(z), \quad z \in \mathbb{R}, \tag{3.8}$$

for some positive constant c. By necessity, $c \geq 1$ as both e and f integrate to 1. In the acceptance-rejection method, we

1. Draw a sample Z from $e(z)$.
2. Draw an independent uniform variable U, $U \sim \mathcal{U}(0,1)$.
3. Accept the sample Z if $U \leq f(Z)/(ce(Z))$; otherwise discard it.

[4]If even higher precision is required, we can use (3.6)–(3.7) as a guess for the root y in the equation $\Phi(y) = x$. Any number of numerical root search routines (e.g. Newton-Raphson) can then be applied to improve the precision of the solution further. Typically only one or two iterations will be required to get the solution to within machine precision on a PC.

The proof of why this algorithm works is straightforward and we omit it. Note that the third step of the acceptance-rejection method can be wasteful if too many samples need rejection. The key to the numerical efficiency of the acceptance-rejection method is thus evidently the ability to identify densities $e(z)$ that are "close" to $f(z)$, in the sense that c is close to 1 for all x. Indeed, it can easily be shown that the probability of rejecting a sample is $1/c$. Press et al. [1992] list good choices for $e(z)$ for a number of standard densities $f(z)$.

To demonstrate the mechanics of setting up an acceptance-rejection scheme for a particular distribution, let us consider sampling of a variable χ^2_ν from a *chi-square distribution* with ν degrees of freedom. This distribution arises in a number of interest rate applications and is characterized by the cumulative distribution function

$$\mathrm{P}\left(\chi^2_\nu \leq z\right) = \frac{1}{2^{\nu/2}\Gamma\left(\nu/2\right)} \int_0^z e^{-y/2} y^{(\nu/2)-1}\, dy, \quad \nu > 0,\ z \geq 0,$$

where Γ is the gamma function. For reasonably large degrees of freedom ν, the chi-square density is typically bell-shaped. The chi-square distribution is a special case of the *gamma distribution* with density

$$f(z; a, b) = \frac{a(az)^{b-1}e^{-az}}{\Gamma(b)}, \quad a, b > 0,\ z \geq 0. \tag{3.9}$$

The chi-square distribution corresponds to $a = \frac{1}{2}$ and $b = \frac{\nu}{2}$. Rather than considering how to simulate a chi-square distribution, we will consider the more general question of how to draw from (3.9). We note that if a variable X has gamma density $f(z; 1, b)$, then aX, $a > 0$, has gamma density $f(z; a, b)$, so, in fact, it suffices to consider a simulation algorithm for the unit-scale density

$$f(z) = \frac{z^{b-1}e^{-z}}{\Gamma(b)},$$

where we assume that $b \geq 1$. One simple choice of "comparison" density for an acceptance-rejection algorithm is the exponential density

$$e(z) = \lambda e^{-\lambda z},$$

which, as mentioned earlier, can easily be simulated by inverse transform techniques. Note that

$$\frac{f(z)}{e(z)} = \frac{1}{\lambda\Gamma(b)} z^{b-1} e^{(\lambda-1)z},$$

which can be checked to have a maximum value of

$$\sup\left(\frac{f(z)}{e(z)}\right) = \frac{1}{\lambda\Gamma(b)}\left(\frac{b-1}{e(1-\lambda)}\right)^{b-1}, \tag{3.10}$$

where we must assume that $\lambda < 1$. To satisfy (3.8) we take $c = \sup(f(z)/e(z))$ and now search for the value of λ that minimizes c, thereby optimizing computational speed. It is easy to see that (3.10) is minimized for $\lambda = 1/b$, corresponding to $c = b^b e^{1-b}/\Gamma(b)$. Note that

$$\frac{f(z)}{ce(z)} = \frac{z^{b-1}}{\lambda} e^{b-1+(\lambda-1)z} b^{-b},$$

with the third step of the acceptance-rejection algorithm best done in logarithms.

The algorithm outlined above was proposed by Fishman [1976] and works best for moderate values of b. For larger values, the Gamma distribution starts looking like a bell-shaped Gaussian distribution and is no longer well-approximated by an exponential distribution. Indeed, we notice that the probability of rejection $(1/c)$ is approximately $e\sqrt{b/(2\pi)}$, so of order $O(\sqrt{b})$. Modifications to the basic Fishman algorithm to accelerate sampling can be found in Cheng and Feast [1980]. Another common idea is to set $e(z)$ to the *Cauchy density*

$$e(z) = \frac{1}{s\pi\left(1 + ((z-t)/s)^2\right)},$$

where $s > 0$ and t are constants. This distribution is bell-shaped and, as discussed earlier, can be simulated by the inverse transform method. Press et al. [1992] list computer code and references for this case. For values $b \in [0, 1]$, the acceptance-rejection technique of Ahrens and Dieter [1974] can also be used.

3.1.1.3 Composition

A third and final method to generate random variables from a given distribution function exploits known functional relationships that map variables sampled from one or more distributions to variables sampled from a target distribution. This technique is known as *composition*. A classical example of composition is the *log-normal distribution* $\mathcal{LN}(\mu, \sigma^2)$ which, as we saw earlier in Chapter 1, is defined through the relation

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow e^X \sim \mathcal{LN}(\mu, \sigma^2),$$

where $\sim$ denotes "distributed as", and where $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian distribution with mean μ and variance σ^2. In other words, a sample Z from $\mathcal{LN}(\mu, \sigma^2)$ can be generated by drawing (by the inverse transformation method, say) a $\mathcal{N}(0, 1)$ variable X, and then setting $Z = e^{\mu+\sigma X}$.

Another classical example of a functional map is the *Student's t-distribution*, where samples can be generated by multiplying independent

samples from a standard Gaussian and a chi-square distribution; see Andersen et al. [2003] for a financial application of this. While we earlier demonstrated that the chi-square and gamma distributions can be generated by acceptance-rejection techniques, in fact we can also use composition for this. For instance, it is known that if $X_1, X_2, \ldots, X_\nu$ are independent standard Gaussian variables, then

$$Z = \sum_{i=1}^{\nu} X_i^2 \tag{3.11}$$

is distributed chi-square with ν degrees of freedom. Also, if $U_1, \ldots, U_b$ are independent uniformly distributed variables, then

$$Z = -a \sum_{i=1}^{b} \ln U_i \tag{3.12}$$

is gamma distributed with density (3.9). For small integer-valued distribution parameters b or ν, (3.11) or (3.12) often define a faster simulation scheme than acceptance-rejection methods.

For later use, we note that the relationship (3.11) can be generalized to

$$\widetilde{\chi}_\nu^2(\lambda) = \sum_{i=1}^{\nu} (X_i + a_i)^2$$

for a series of constants a_i, $i = 1, \ldots, \nu$. The random variable $\widetilde{\chi}_\nu^2(\lambda)$ follows a so-called *non-central chi-square distribution* with ν degrees of freedom and *non-centrality parameter* $\lambda = \sum_i a_i^2$. The distribution function is given by

$$\mathrm{P}\left(\widetilde{\chi}_\nu^2(\lambda) \le z\right) = e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{\left(\frac{1}{2}\lambda\right)^j}{j!\Gamma\left(\frac{\nu}{2}+j\right) 2^{(\nu/2)+j}} \int_0^z y^{\nu/2+j-1} e^{-y/2}\, dy, \tag{3.13}$$

an expression that also holds for non-integer ν. If $\nu > 1$, samples from a non-central chi-squared distribution can be generated by composition, using the relation

$$\widetilde{\chi}_\nu^2(\lambda) = \left(Z + \sqrt{\lambda}\right)^2 + \chi_{\nu-1}^2,$$

where Z is a standard Gaussian random variable independent of $\chi_{\nu-1}^2$. To handle the case $\nu \le 1$, one can observe from the expression (3.13) that a non-central chi-square variable can be expressed as a regular chi-square variable $\chi_{\nu+2N}^2$, where N is an independent *Poisson-distributed* discrete variable with intensity $\lambda/2$,

$$\mathrm{P}(N = j) = e^{-\lambda/2} \frac{(\lambda/2)^j}{j!}, \quad j = 0, 1, \ldots.$$

This suggests a composition rule for arbitrary ν: draw Poisson variables N (by the inverse transformation method, say) and then draw $\chi_{\nu+2N}^2$ using the methods in Section 3.1.1.2.

3.1.2 Correlated Gaussian Samples

The previous section dealt with the generation of scalar random variables. In applications, however, we may face the task of generating *vectors* of random variables, drawn from a joint multi-variate distribution. Of primary importance in financial applications is the multi-variate Gaussian distribution, so we devote this section to issues surrounding the generation of correlated Gaussian samples.

Recall that a p-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is characterized by a p-dimensional vector-valued mean μ and a $p \times p$ symmetric, positive semi-definite[5] covariance matrix Σ. The joint density is

$$\phi_p(z; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)\right), \quad z \in \mathbb{R}^p.$$

The following result is useful:

Lemma 3.1.3 (Linear Transformation). *Let $Z \sim \mathcal{N}(\mu, \Sigma)$ be p-dimensional. Given a $d \times p$ matrix A and a d-dimensional vector B, then*

$$AZ + B \sim \mathcal{N}\left(A\mu + B, A\Sigma A^\top\right).$$

We can use this lemma as follows. Suppose that we generate p independent standard (that is, $\mathcal{N}(0,1)$) Gaussian samples and collect them in a p-dimensional vector X. This can be accomplished using the techniques in Section 3.1.1. Clearly $X \sim \mathcal{N}(0, I)$, where I is the p-dimensional identity matrix. Define a $(p \times p)$-dimensional matrix C satisfying

$$CC^\top = \Sigma. \tag{3.14}$$

Then

$$Z = \mu + CX$$

is distributed $\mathcal{N}(\mu, \Sigma)$.

It remains to determine a matrix C that satisfies (3.14). While there is generally an infinite number of such matrices, two particular choices are of primary importance. We discuss these below.

3.1.2.1 Cholesky Decomposition

In the Cholesky decomposition, we impose the constraint that the matrix C be lower triangular (that is, having all zeros above the diagonal), thereby conveniently reducing the number of multiplications required to compute CX to $p(1 + (p-1)/2)$, rather than p^2. Assuming that the matrix is positive definite (not only positive semi-definite), the Cholesky decomposition is well-defined, and given by

[5] That is, all eigenvalues of Σ are non-negative.

$$C_{i,i} = \sqrt{\Sigma_{i,i} - \sum_{k=1}^{i-1} C_{i,k}^2}, \quad i = 1, \ldots, p,$$

$$C_{i,j} = \frac{1}{C_{j,j}} \left(\Sigma_{i,j} - \sum_{k=1}^{j-1} C_{i,k} C_{j,k} \right), \quad j = 1, \ldots, p-1, \quad j < i.$$

For instance, if

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where $\rho \in [-1, 1]$ and $\sigma_1, \sigma_2 > 0$, then

$$C = \begin{pmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{pmatrix},$$

a result that we have already used in Section 2.11. Press et al. [1992], among others, list computer code implementing the relations above.

If the matrix Σ is only positive semi-definite (but not positive definite), the Cholesky decomposition will fail. In this case, linear algebra tells us that the matrix Σ is rank-deficient, with rank $r < p$. As such, we must be able to set $Z = \mu + MY$, where M is a $p \times r$ matrix and $Y \sim \mathcal{N}(0, \Sigma_Y)$ is r-dimensional, with the covariance matrix having full rank r. Using Cholesky composition instead on Σ_Y, we can find a lower diagonal matrix C_Y satisfying $C_Y C_Y^\top = \Sigma_Y$. Thus, in this case

$$Z = \mu + MC_Y X$$

where X is a vector of r (not p) independent standard Gaussian samples. The matrix M can be found by the singular value decomposition (SVD) algorithm, see Press et al. [1992], or the algorithm in the next section.

3.1.2.2 Eigenvalue Decomposition

As an alternative to Cholesky decomposition, we can also consider diagonalizing Σ through an eigenvalue decomposition. Here, we write

$$\Sigma = E\Lambda E^\top, \tag{3.15}$$

where Λ is a diagonal matrix of eigenvalues λ_i, $i = 1, \ldots, p$, and the columns of E contain the orthonormal eigenvectors of Σ. Some eigenvalues may be zero, if Σ is rank-deficient (positive semi-definite). Comparison with (3.14) implies that one choice of C is

$$C = E\sqrt{\Lambda} = E \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{\lambda_p} \end{pmatrix}. \tag{3.16}$$

The eigenvalue decomposition (3.15) is relatively straightforward, at least as eigenvalue problems go, due to the fact that Σ is symmetric and positive semi-definite; see Press et al. [1992] for an algorithm. While both Cholesky decomposition and eigenvalue decompositions have computational complexity $O(n^3)$, in practice the Cholesky method is often much faster than the eigenvalue method, making the Cholesky method preferable in practice. Nevertheless, decompositions of the type (3.16) have certain appealing theoretical properties that shall be useful later, so the next section explores (3.16) further.

3.1.3 Principal Components Analysis (PCA)

Consider a p-dimensional Gaussian variable Z with a given covariance matrix Σ. Assume, with no loss of generality, that the mean of Z is 0 and that Σ has full rank (positive definite). Consider now writing, as an approximation,

$$Z \approx DX, \tag{3.17}$$

where X is an r-dimensional vector of independent standard Gaussian variables, $r \leq p$, and D is a $(p \times r)$-dimensional matrix. How should we choose D in an optimal way?

First, we obviously need to define what constitutes an "optimal" approximation in (3.17). We here have in mind L^2 closeness of the covariance matrix $DD^\top$ to Σ (see Lemma 3.1.3), so let us define the optimal D^* as the matrix that minimizes the norm

$$f(D) = \operatorname{tr}\left(\left(\Sigma - DD^\top\right)\left(\Sigma - DD^\top\right)^\top\right).$$

This is just the matrix representation of the usual Frobenius norm on the squared differences between Σ and $DD^\top$. The value of D that minimizes $f(D)$ can be shown to be

$$D^* = E_r\sqrt{\Lambda_r}, \tag{3.18}$$

where Λ_r is an $r \times r$ diagonal matrix containing the *largest* r eigenvalues of Σ, and E_r is a $p \times r$ matrix of r p-dimensional eigenvectors corresponding to the eigenvalues in Λ_r.

Equipped with the optimal D, we now go back to the approximation (3.17) and write

$$Z \approx \widetilde{Z} \triangleq E_r\sqrt{\Lambda_r}X = \sqrt{\lambda_1}e_1X_1 + \sqrt{\lambda_2}e_2X_2 + \ldots + \sqrt{\lambda_r}e_rX_r, \tag{3.19}$$

where e_i denotes the i-th column of E_r and the λ_i's are the eigenvalues, sorted in decreasing order of magnitude. The (deterministic) vector e_i is known as the i-th *principal component* of Z, and the (random) variable $\sqrt{\lambda_i}X_i$ as the i-th *principal factor*. With (3.19), we have $\operatorname{tr}(\operatorname{Cov}(Z,Z)) = \mathrm{E}(Z^\top Z) = \sum_{i=1}^p \lambda_i$ and $\operatorname{tr}(\operatorname{Cov}(\widetilde{Z},\widetilde{Z})) = \mathrm{E}(\widetilde{Z}^\top\widetilde{Z}) = \sum_{i=1}^r \lambda_i$, i.e. the first r terms in the decomposition (3.19) explain a fraction

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

of the sum of the diagonal elements of the covariance matrix of Z. Principal components decomposition will thus result in a loss of total variance, unless the covariance matrix is either rank-deficient (i.e. has eigenvalues that are strictly zero), or we use a full set of principal components ($p = r$). In many cases of interest to us here, the loss of variance can be small, even if r is a modest number, e.g. 2 or 3. We notice that the covariance matrix for Z, as approximated by (3.19), will be *rank-deficient*, as the number r of non-zero eigenvalues is less than p.

While we have used a setting with Gaussian variables to motivate our treatment of principal components analysis (PCA), it is, in fact, a generically useful tool for uncovering the structure of large-dimensional random vectors, and replacing them with more manageable, lower-dimensional variables; see, e.g., Theil [1971] for more details and an application to empirical non-Gaussian data. Also, PCA identifies which directions of a multi-dimensional random variable are "important", potentially allowing us to allocate computational resources in an intelligent manner. One example of this is shown later in this chapter, in Section 3.2.10.

3.2 Generation of Sample Paths

So far, we have assumed that random variables are characterized by a known distribution function. In most of our applications, however, the random variables $g(T)/N(T)$ used in the basic pricing equation (3.1) are specified through an SDE or, more generally, an Ito process. In this section, we shall discuss Monte Carlo simulation of such processes. We start out with a motivating example, set in the Black-Scholes-Merton economy.

3.2.1 Example: Asian Basket Options in Black-Scholes Economy

Consider a dividend-free stock S, with Black-Scholes dynamics

$$dS(t)/S(t) = r\,dt + \sigma\,dW(t), \tag{3.20}$$

where $W(t)$ is a Brownian motion in the risk-neutral measure Q, r is a constant interest rate, and σ is a constant volatility. Let there be given an increasing set of observation times $\{t_1, t_2, , \ldots, t_m\}$, with $t_m = T$, and define the $\mathcal{F}_T$-measurable (discretely observed) stock average

$$A(T) = \frac{1}{m} \sum_{i=1}^{m} S(t_i). \tag{3.21}$$

An *Asian* (or *average rate*) call option with strike K is defined by the terminal payout

$$g(T) = (A(T) - K)^+ ; \tag{3.22}$$

we wish to price this option by Monte Carlo simulation.

As discussed earlier (see (1.39)), the geometric Brownian motion process (3.20) allows us to express S directly in terms of the Brownian motion,

$$S(t) = S(0)e^{rt - \frac{1}{2}\sigma^2 t + \sigma W(t)}, \quad t > 0,$$

whereby, with $\Delta_i \triangleq t_i - t_{i-1}$ and $t_0 = 0$,

$$S(t_i) = S(t_{i-1}) \exp\left(\left[r - \frac{1}{2}\sigma^2\right] \Delta_i + \sigma \left[W(t_i) - W(t_{i-1})\right]\right),$$

$i = 1, \ldots, m$. By the properties of Brownian motion, the increments $W(t_i) - W(t_{i-1})$ are independent Gaussian variables distributed as $\mathcal{N}(0, \Delta_i)$. For the purposes of Monte Carlo simulation, we can therefore write

$$S(t_i) = S(t_{i-1}) \exp\left(\left(r - \frac{1}{2}\sigma^2\right) \Delta_i\right) \exp\left(\sigma\sqrt{\Delta_i} Z_i\right), \quad i = 1, \ldots, m, \tag{3.23}$$

where the Z_i are independent standard $\mathcal{N}(0,1)$ Gaussian random variables. To produce a single sample draw of $g(T)$, we thus

1. Draw independent standard Gaussian samples Z_i, $i = 1, \ldots, m$ (see Section 3.1.1).
2. Starting from $S(0)$, generate $S(t_i)$, $i = 1, \ldots, m$, from the iteration (3.23).
3. Compute $g(T)$ from (3.21)–(3.22).

Repeating this procedure n times (with Gaussian samples independent from one path to the next), we can generate n random samples $g_1, g_2, \ldots, g_n$ of $g(T)$. Our estimate of the time 0 price of the Asian option is then, from (3.3) with $N(t) = e^{rt}$ and non-random,

$$\overline{V}(0) = e^{-rT} \frac{1}{n} \sum_{j=1}^{n} g_j.$$

Asymptotic confidence intervals can be computed from (3.4). The pricing algorithm involves drawing mn Gaussian variables, so the computational cost of the pricing algorithm is $O(mn)$.

Increasing the complexity, let us now consider an Asian option on a p-dimensional basket of stocks $S_1, S_2, \ldots, S_p$, each following geometric Brownian motion,

$$dS_k(t)/S_k(t) = r\,dt + \sigma_k\,dW_k(t), \quad k = 1, \ldots, p.$$

The Brownian motions W_k and W_j are assumed correlated with constant correlation coefficient $\rho_{k,j}$, $j,k = 1,\ldots,p$, $j \neq k$. Define a unit-weighted basket price as

$$B(t) = \sum_{k=1}^{p} S_k(t),$$

and set the terminal Asian option payout to be

$$g(T) = \left(\frac{1}{m} \sum_{i=1}^{m} B(t_i) - K \right)^+, \tag{3.24}$$

where the time line $\{t_i\}$ is as before. Equivalent to (3.23), we draw sample paths for each asset according to the prescription

$$S_k(t_i) = S_k(t_{i-1}) \exp\left(\left(r - \frac{1}{2}\sigma_k^2 \right) \Delta_i + \sigma_k \sqrt{\Delta_i} Z_{k,i} \right), \tag{3.25}$$
$$i = 1,\ldots,m, \qquad k = 1,\ldots,p,$$

where the $Z_{k,i}$ are Gaussian samples, independently drawn at each time step but correlated across k's. Let C be the Cholesky decomposition of the correlation matrix $\{\rho_{k,j}\}$ (see Section 3.1.2.1), in which case we can generate the correlated sample vectors $Z_i = (Z_{1,i}, Z_{2,i}, \ldots, Z_{p,i})^\top$ as

$$Z_i = CX_i$$

for a p-dimensional vector X_i of independent Gaussian samples. Given joint sample paths of all basket component assets S_k, $k = 1,\ldots,p$, pricing of the Asian basket option proceeds as above, substituting (3.24) for (3.22).

Completion of (3.25) requires pm samples to complete a full path of all p assets, making the total computational effort of an n-sample Monte Carlo scheme $O(nmp)$, with the (probabilistic) convergence order $O(n^{-1/2})$ and dependent only on n. As mentioned earlier, the linearity of computational cost on the dimension of the asset vector p compares favorably to the exponential growth in p of finite difference schemes. Notice also the ease with which the Monte Carlo scheme is able to incorporate path-dependence.

3.2.2 Discretization Schemes, Convergence, and Stability

At the heart of the example in Section 3.2.1 was an iterative scheme for the production of a sample path for a vector-valued SDE; see (3.25). For the simple Black-Scholes model, SDE state variables (stock prices) could be expressed analytically in terms of independent increments of a Brownian motion, making path generation straightforward. In practice, however, we are often working with SDEs that do not permit closed-form solution. In such cases, we need to *time-discretize* the SDE, much the same way as we did for the numerical solution of PDEs.

In the next few sections, we shall consider a few important SDE discretization schemes. Before moving on to this, it is useful to discuss the sense in which we consider a discretization scheme to converge to the true SDE solution. For this, consider a vector-valued SDE

$$dX(t) = \mu(t, X(t))\,dt + \sigma(t, X(t))\,dW(t), \tag{3.26}$$

where $X(t)$ is p-dimensional, W is a d-dimensional vector of independent Brownian motions, and $\mu : [0,T] \times \mathbb{R}^p \to \mathbb{R}^p$ and $\sigma : [0,T] \times \mathbb{R}^p \to \mathbb{R}^{p\times d}$ satisfy the usual regularity conditions. Consider an equidistant[6] time grid $\{0, \Delta, 2\Delta, \ldots, m\Delta\}$, the number of references and let $\widehat{X}$ be an approximation to X, based on some kind of time-discretization scheme on the grid $\{i\Delta\}$. For simplicity of notation, set $\widehat{X}_i \triangleq \widehat{X}(i\Delta)$. We say that the underlying approximation is *weakly consistent* if there exists a function $c(\Delta)$ with

$$\lim_{\Delta\downarrow 0} c(\Delta) = 0$$

such that (dropping the measure superscript on the expectation operator)

$$\mathrm{E}\left(\left|\mathrm{E}\left(\Delta^{-1}\left(\widehat{X}_{i+1} - \widehat{X}_i\right)\middle|\,\mathcal{F}_{i\Delta}\right) - \mu\left(i\Delta, \widehat{X}_i\right)\right|^2\right) \le c(\Delta), \tag{3.27}$$

and

$$\mathrm{E}\left(\left|\mathrm{E}\left(\Delta^{-1}\left(\widehat{X}_{i+1} - \widehat{X}_i\right)\left(\widehat{X}_{i+1} - \widehat{X}_i\right)^\top\middle|\,\mathcal{F}_{i\Delta}\right)\right.\right.$$
$$\left.\left. -\sigma\left(i\Delta, \widehat{X}_i\right)\sigma\left(i\Delta, \widehat{X}_i\right)^\top\right|^2\right) \le c(\Delta), \tag{3.28}$$

for all $i = 0, \ldots, m-1$. The notion of weak consistency[7] thus amounts to requiring that the mean and variance of the increments of the approximating process be close to those of the true SDE solution.

A concept related to consistency is the notion of *weak convergence*. We say that an approximate solution converges weakly to X at time $T = m\Delta$ with respect to a class $\mathcal{C}$ of test functions $g : \mathbb{R}^p \to \mathbb{R}$ if

$$\lim_{\Delta\downarrow 0}\left|\mathrm{E}\left(g(X(T)\right) - \mathrm{E}\left(g(\widehat{X}(T)\right)\right| = 0, \tag{3.29}$$

for all $g \in \mathcal{C}$. Notice that the limit necessarily involves $m \to \infty$.

[6] To keep notation manageable, we use a constant time step Δ in most of this chapter. All results are, however, easily extendable to non-equidistant grids.

[7] *Strong consistency* (which is of little use to us in this book) requires that (3.27) is satisfied, and that the variance of the difference between increments of the true process and the approximation vanish. The second requirement is stronger than (3.28).

The class of test functions used in (3.29) is normally always in the set C_P^l of functions with polynomially bounded[8] derivatives of order $0, 1, \ldots, l$ with maximum power l. We say that a scheme converges with *weak order* β if, for all $g \in C_P^{2(\beta+1)}$, (3.29) can be strengthened to

$$\left|\mathrm{E}\left(g(X(T)\right) - \mathrm{E}\left(g(\widehat{X}(T)\right)\right| \leq c\Delta^{\beta}, \tag{3.30}$$

for all $\Delta \in (0, \Delta_0)$, where Δ_0 and c are constants and c does not depend on Δ (but may depend on g).

One would generally expect that a weakly consistent scheme is weakly convergent. Indeed, this can be established to be the case under certain additional regularity conditions. We will not list the exact result here, but refer to Kloeden and Platen [2000], Theorem 9.7.4.

Finally, a brief word on stability of a time-discretized SDE. A commonly used definition of A-stability focuses on the behavior of a discretized test SDE of the type

$$dX(t) = \lambda X(t)\, dt + dW(t), \tag{3.31}$$

where λ is a complex-valued constant with real part $\mathrm{Re}\,(\lambda) < 0$. We suppose that a discretization scheme can be represented as

$$\widehat{X}_{i+1} = \widehat{X}_i G\,(\lambda\Delta) + Z_i^{\Delta}, \quad i = 0, 1, \ldots, m-1, \tag{3.32}$$

where G is a mapping of the complex plane onto itself and the Z_i^{Δ}'s are random variables independent of the $\widehat{X}_i$'s. In this case, the *region of stability* for a scheme is the set of $\lambda\Delta$ for which $\mathrm{Re}\,(\lambda) < 0$ and

$$|G\,(\lambda\Delta)| < 1. \tag{3.33}$$

Similar to the definition used for finite difference scheme discretizations, we say that an SDE time-discretization scheme is *A-stable*, if the region of stability includes all values of λ with $\mathrm{Re}\,(\lambda) < 0$ and all $\Delta > 0$.

3.2.3 The Euler Scheme

An obvious first scheme to discretize (3.26) treats both dt and $dW(t)$ fully explicitly, evaluating all SDE coefficients on time step $[i\Delta, i\Delta + \Delta]$ at the left interval point $i\Delta$. In other words, we write, starting from $\widehat{X}_0 = X(0)$,

$$\begin{aligned}\widehat{X}_{i+1} &= \widehat{X}_i + \mu\left(i\Delta, \widehat{X}_i\right)\Delta + \sigma\left(i\Delta, \widehat{X}_i\right)\left(W\left(i\Delta + \Delta\right) - W\left(i\Delta\right)\right), \\ i &= 0, 1, \ldots, m-1.\end{aligned} \tag{3.34}$$

[8] A function $f : \mathbb{R}^p \to \mathbb{R}$ is polynomially bounded if $|f(x)| \leq k(1 + |x|^q)$, $x \in \mathbb{R}^p$, for constants k and q.

With this scheme, Monte Carlo generation of paths is straightforward and involves, as in Section 3.2.1, replacing the increments $W(i\Delta + \Delta) - W(i\Delta)$ with $Z_i\sqrt{\Delta}$, for a d-dimensional vector of independent standard Gaussian samples Z_i.

The discretization scheme (3.34) is known as the *Euler scheme*, sometimes also called the *Euler-Maruyama* scheme. The Euler scheme is easy to implement and is a true workhorse that we will often use in this book. We note that the scheme is weakly consistent, as

$$\mathrm{E}\left(\left|\mathrm{E}\left(\Delta^{-1}\left(\widehat{X}_{i+1} - \widehat{X}_i\right)\middle|\mathcal{F}_{i\Delta}\right) - \mu\left(i\Delta, \widehat{X}_i\right)\right|^2\right) = 0,$$

and

$$\mathrm{E}\left(\left|\mathrm{E}\left(\Delta^{-1}\left(\widehat{X}_{i+1} - \widehat{X}_i\right)\left(\widehat{X}_{i+1} - \widehat{X}_i\right)^{\top}\middle|\mathcal{F}_{i\Delta}\right)\right.\right.$$
$$\left.\left. -\sigma\left(i\Delta, \widehat{X}_i\right)\sigma\left(i\Delta, \widehat{X}_i\right)^{\top}\right|^2\right) = O\left(\Delta^2\right).$$

While one might believe that the explicit discretization of the diffusion term — which is only accurate to order $O(\sqrt{\Delta})$ — would give the scheme weak convergence order[9] 1/2, in fact we typically have that the Euler scheme has *weak convergence order* $\beta = 1$. We note that for this result to hold, however, regularity conditions on μ and σ stronger than those of the existence and uniqueness results (Theorem 1.6.1) are needed. For instance, in the case where μ and σ are functions of X alone, Theorem 9.7.6 in Kloeden and Platen [2000] requires that μ and σ be four times continuously differentiable with polynomial growth and uniformly bounded derivatives. See also their Theorem 15.4.2 for a more general result.

Given that the Euler scheme is fully explicit, our experience from finite difference methods suggests that the scheme may have stability problems. To investigate, we follow Section 3.2.2 and consider the test SDE

$$dX(t) = \lambda X(t)dt + dW(t),$$

which is discretized as

$$\widehat{X}_{i+1} = \widehat{X}_i\left(1 + \lambda\Delta\right) + \sqrt{\Delta}Z_i, \tag{3.35}$$

where Z_i's are standard Gaussian. Comparison to (3.32) and (3.33) shows that the region of stability for the Euler scheme is

$$|(1 + \lambda\Delta)| < 1, \quad \mathrm{Re}\,(\lambda) < 0,$$

which is the unit disc in the complex plane centered at $\lambda\Delta = -1$. For a given λ, there are thus restrictions on how big a time step Δ can be used.

[9] The so-called *strong convergence order* of the Euler scheme is in fact only 1/2. The concept of strong convergence order is defined in Kloeden and Platen [2000] and is of little importance to applications in this book.

3.2.3.1 Linear-Drift SDEs

The restricted stability region of the Euler scheme can be a practical concern. For instance, SDEs of the important type

$$dX(t) = \varkappa\left(\theta(t) - X(t)\right)dt + \sigma\left(t, X(t)\right)dW(t) \tag{3.36}$$

arise quite frequently in fixed income modeling, and in cases where $\varkappa$ is big (which is often the case for, say, stochastic volatility models such as those covered in Chapters 8, 9 and 13) the Euler scheme can become unstable and return meaningless results. One way to solve the problem is to switch to an implicit scheme (see next section), but in the case (3.36) we can use the fact that the drift term can be removed by a simple change of variable. For instance, for the case where $X(t)$ is scalar we can set

$$Y(t) = e^{\varkappa t}X(t) - \varkappa\int_0^t e^{\varkappa u}\theta(u)\,du,$$

such that, from Ito's lemma,

$$\begin{aligned} dY(t) &= e^{\varkappa t}\sigma\left(t, X(t)\right)dW(t) \\ &= e^{\varkappa t}\sigma\left(t, e^{-\varkappa t}\left(Y(t) + \varkappa\int_0^t e^{\varkappa u}\theta(u)\,du\right)\right)dW(t). \end{aligned}$$

Euler simulation of the process for $Y(t)$, rather than for $X(t)$, will center X around its analytically known mean

$$\mathrm{E}\left(X\left((i+1)\Delta\right)\middle|\, X\left(i\Delta\right)\right) = e^{-\varkappa\Delta}X\left(i\Delta\right) + \varkappa\int_{i\Delta}^{i\Delta+\Delta} e^{-\varkappa((i+1)\Delta-u)}\theta(u)\,du$$

and will often alleviate any stability problems.

3.2.3.2 Log-Euler Scheme

One potential problem with the pure Euler scheme (3.34) is the fact that all increments are locally Gaussian, thereby implying a non-zero probability of $\widehat{X}$ crossing zero and becoming negative. Many SDEs, however, are known to produce only non-negative solutions, and the functions μ and σ may not allow for negative arguments. This, for instance, is the case for the square-root process

$$dX(t) = \sqrt{X(t)}\,dW(t), \quad X(0) > 0,$$

where the Euler scheme cannot be directly applied. Some authors (e.g., Kloeden and Platen [2000]) suggest heuristic modifications of the Euler scheme, such as

$$\widehat{X}_{i+1} = \widehat{X}_i + \sqrt{|\widehat{X}_i|}\left(W\left(i\Delta + \Delta\right) - W\left(i\Delta\right)\right),$$

but ultimately this is not very satisfying and the resulting scheme will often have large errors[10]. An alternative is to introduce an invertible transformation $X(t) = f(Y(t))$, with $f : \mathbb{R} \to \mathbb{R}_+$, and then apply the Euler scheme to Y, at each step recovering X as $f(Y)$. In finance applications, where many processes are based on SDEs that bear some resemblance to geometric Brownian motion, an often-used choice for f is $f(y) = e^y$. The resulting scheme is known as the *log-Euler scheme.*

Consider the SDE (3.26) and assume for simplicity that X is scalar (if X is vector valued, the log-transform can be applied to all, or a few selected, components of X). Set $X(t) = \exp(Y(t))$, such that $Y(t) = \ln(X(t))$. The process for Y then follows from Ito's lemma:

$$dY(t) = \left(\frac{\mu\left(t, X(t)\right)}{X(t)} - \frac{1}{2}\frac{\sigma\left(t, X(t)\right)^2}{X(t)^2}\right)dt + \frac{\sigma\left(t, X(t)\right)}{X(t)}dW(t),\ X(t) = e^{Y(t)}.$$

Writing out a standard Euler scheme for Y and making the transformation $\widehat{X}_i = \exp(\widehat{Y}_i)$ gives us the (scalar) log-Euler scheme for X :

$$\widehat{X}_{i+1} = \widehat{X}_i \exp\left(\left(\frac{\mu(t, \widehat{X}_i)}{\widehat{X}_i} - \frac{1}{2}\frac{\sigma(t, \widehat{X}_i)^2}{\widehat{X}_i^2}\right)\Delta + \frac{\sigma(t, \widehat{X}_i)}{\widehat{X}_i}Z_i\sqrt{\Delta}\right),$$

where $Z_i \sim \mathcal{N}(0, 1)$. Generalizations of the technique above to situations where the valid range of X is some general set $\mathcal{C}$ are obvious and involve identifying an invertible mapping function $f : \mathbb{R} \to \mathcal{C}$, preferably one that can be inverted analytically. For instance, if $\mathcal{C} = [a, \infty)$, we could use $f(y) = a + e^y$.

3.2.4 The Implicit Euler Scheme

The implicit Euler scheme for the vector-valued SDE (3.26) takes the form

$$\widehat{X}_{i+1} = \widehat{X}_i + \mu\left(i\Delta + \Delta, \widehat{X}_{i+1}\right)\Delta + \sigma\left(i\Delta, \widehat{X}_i\right)\left(W\left(i\Delta + \Delta\right) - W\left(i\Delta\right)\right), \tag{3.37}$$

for $i = 0, 1, \ldots, m-1$. We highlight the fact that the drift coefficient μ is now evaluated at time $i\Delta + \Delta$, rather than at time $i\Delta$. It is easy to show that the implicit Euler scheme is consistent. Under regularity conditions, it can also be shown that the weak convergence order is $\beta = 1$, just as was the case for the explicit Euler scheme.

The main advantage of the implicit Euler scheme over the explicit Euler scheme is numerical stability. To examine the region of stability for the implicit Euler scheme, consider again the test SDE

[10]For a dedicated treatment of the rather delicate problem of simulating square-root process, see Chapter 9.

$$dX(t) = \lambda X(t)\,dt + dW(t).$$

It will now be discretized as (compare to (3.35))

$$\widehat{X}_{i+1} = \widehat{X}_i + \widehat{X}_{i+1}\lambda\Delta + \sqrt{\Delta}Z_i,$$

or

$$\widehat{X}_{i+1}\left(1 - \lambda\Delta\right) = \widehat{X}_i + \sqrt{\Delta}Z_i.$$

Comparison to (3.32) and (3.33) shows that now

$$G\left(\lambda\Delta\right) = \frac{1}{1 - \lambda\Delta}$$

such that the stability criterion $|G(\lambda\Delta)| < 1$ is satisfied for any value of $\lambda\Delta$ where $\mathrm{Re}\left(\lambda\right) < 0$. In other words, the implicit scheme is A-stable.

3.2.4.1 Implicit Diffusion Term

The reader may at this point wonder why the implicit scheme (3.37) only discretized the drift term (μ) implicitly, and not the diffusion term (σ). The answer lies in the differences between a regular Riemann integral and the stochastic integral. Recall in particular that the stochastic Ito integral is defined to be non-anticipative, in the sense that the integrand is always evaluated "to the left" on any partitions of the Brownian motion. As a consequence, if $\sigma(i\Delta, \widehat{X}_i)$ were replaced with $\sigma(i\Delta + \Delta, \widehat{X}_{i+1})$ in (3.37), the resulting scheme would not be weakly consistent, in the sense defined earlier. To illustrate this point, just consider the simple scalar process

$$dX(t) = \sigma X(t)\,dW(t),$$

which we contemplate discretizing as

$$\widehat{X}_{i+1} = \widehat{X}_i + \sigma\widehat{X}_{i+1}\left(W\left(i\Delta + \Delta\right) - W\left(i\Delta\right)\right),$$

or

$$\widehat{X}_{i+1}\left(1 - \sigma Z_i\sqrt{\Delta}\right) = \widehat{X}_i, \quad i = 0, \ldots, m-1. \tag{3.38}$$

Here, a first difficulty arises: the term $(1 - \sigma Z_i\sqrt{\Delta})$ may become 0 (or very close to zero) if Z_i is an (unbounded) Gaussian variable. For fully implicit discretization schemes, it becomes necessary to use a bounded approximation to the Brownian motion. As discussed in Kloeden and Platen [2000], weak convergence order is preserved if in (3.38) we set the Z_i to be independent binomial variables with

$$\mathrm{P}\left(Z_i = 1\right) = \mathrm{P}\left(Z_i = -1\right) = \frac{1}{2}.$$

We assume that $1 - \sigma\sqrt{\Delta} > 0$. Rearranging and Taylor-expanding, we get

$$\begin{aligned}\frac{\widehat{X}_{i+1}-\widehat{X}_i}{\Delta} &= \frac{\widehat{X}_i}{\Delta}\left(\frac{1}{1-\sigma Z_i\sqrt{\Delta}}-1\right)\\ &= \frac{\widehat{X}_i}{\Delta}\left(1+\sigma Z_i\sqrt{\Delta}+\sigma^2 Z_i^2\Delta+O\left(\sigma^3 Z_i^3\Delta^{3/2}\right)-1\right)\\ &= \widehat{X}_i\left(\sigma Z_i\Delta^{-1/2}+\sigma^2 Z_i^2+O\left(\sigma^3 Z_i^3\Delta^{1/2}\right)\right)\end{aligned}$$

such that

$$\mathrm{E}\left(\frac{\widehat{X}_{i+1}-\widehat{X}_i}{\Delta}\middle|\widehat{X}_i\right)=\widehat{X}_i\left(\sigma^2+O\left(\Delta^{1/2}\right)\right).$$

Clearly, this will cause a violation of the consistency condition (3.27).

In the example above, we notice that consistency can be restored if the drift of the original SDE is changed from 0 to $-\sigma^2 X(t)$ before the "doubly" implicit Euler discretization is employed. More generally, it is not difficult to show that (3.37) can be modified to treat the diffusion term implicitly, provided that the drift of the original vector-valued SDE (3.26) is first changed from μ to

$$\overline{\mu}=\mu-\sum_{j=1}^{d}\sum_{k=1}^{p}\left(\sigma_{X_k}\right)_{\cdot,j}\sigma_{k,j}$$

where the p-dimensional vector $(\sigma_{X_k})_{\cdot,j}$ is the j-th column of the $(p\times d)$-dimensional matrix $\sigma_{X_k}=\{\partial\sigma_{i,j}/\partial X_k\}$. Inspired by the theta methods of Chapter 2, we can, in fact, introduce a family of discretizations

$$\begin{aligned}\widehat{X}_{i+1}=\widehat{X}_i&+\left[(1-\theta)\overline{\mu}_\eta\left(i\Delta,\widehat{X}_i\right)+\theta\overline{\mu}_\eta\left(i\Delta+\Delta,\widehat{X}_{i+1}\right)\right]\Delta\\ &+\left[(1-\eta)\sigma\left(i\Delta,\widehat{X}_i\right)+\eta\sigma\left(i\Delta+\Delta,\widehat{X}_{i+1}\right)\right]Z_i\sqrt{\Delta},\end{aligned}\tag{3.39}$$

where the Z_i are binomially distributed variables, $\theta,\eta\in[0,1]$ are parameters, and

$$\overline{\mu}_\eta=\mu-\eta\sum_{j=1}^{d}\sum_{k=1}^{p}\left(\sigma_{X_k}\right)_{\cdot,j}\sigma_{k,j}.\tag{3.40}$$

As it turns out, all these schemes theoretically have identical convergence order $\beta=1$, but in practice some choices of θ,η may turn out to work better than others. We shall discuss methods to raise the theoretical convergence order in Section 3.2.6. The scheme (3.39) can be verified to be A-stable for $\theta\in[1/2,1]$.

3.2.5 Predictor-Corrector Schemes

A closer examination of the implicit Euler scheme (3.37) demonstrates the need to recover $\widehat{X}(i\Delta+\Delta)$ as the vector-valued root of a possibly nonlinear equation. In general, this must be done numerically (using, say, the

Newton-Raphson algorithm), causing a severe deterioration of computational performance. An alternative is to use the explicit Euler scheme as a *predictor* and the implicit scheme as a *corrector*, much the same way we used explicit finite difference approximations as predictors in the Craig-Sneyd algorithm of Section 2.11. Moving straight to the general implicit discretization family (3.39), we write the predictor-corrector as

$$\overline{X}_{i+1} = \widehat{X}_i + \mu\left(i\Delta, \widehat{X}_i\right)\Delta + \sigma\left(i\Delta, \widehat{X}_i\right)\left(W\left(i\Delta + \Delta\right) - W\left(i\Delta\right)\right), \quad (3.41)$$

$$\begin{aligned}\widehat{X}_{i+1} &= \widehat{X}_i + \left[(1-\theta)\overline{\mu}_\eta\left(i\Delta, \widehat{X}_i\right) + \theta\overline{\mu}_\eta\left(i\Delta + \Delta, \overline{X}_{i+1}\right)\right]\Delta \\ &\quad + \left[(1-\eta)\sigma\left(i\Delta, \widehat{X}_i\right) + \eta\sigma\left(i\Delta + \Delta, \overline{X}_{i+1}\right)\right]\left(W\left(i\Delta + \Delta\right) - W\left(i\Delta\right)\right),\end{aligned} \quad (3.42)$$

where $\theta, \eta \in [0,1]$, and $\overline{\mu}_\eta$ is as given in (3.40). It is understood that the Brownian motion increments in (3.41) and (3.42) are to be identical.

For sufficiently smooth coefficients, it can be shown that the predictor-corrector scheme (3.41)–(3.42) converges weakly with order $\beta = 1$, independent of the choice of θ and η. As for stability, discretization of (3.31) leads to

$$\begin{aligned}\overline{X}_{i+1} &= \widehat{X}_i\left(1 + \lambda\Delta\right) + W\left(i\Delta + \Delta\right) - W\left(i\Delta\right), \\ \widehat{X}_{i+1} &= \widehat{X}_i + \left[(1-\theta)\lambda\widehat{X}_i + \theta\lambda\overline{X}_{i+1}\right]\Delta + W\left(i\Delta + \Delta\right) - W\left(i\Delta\right) \\ &= \widehat{X}_i\left(1 + \lambda\Delta\left(1 + \theta\lambda\Delta\right)\right) + \left(W\left(i\Delta + \Delta\right) - W\left(i\Delta\right)\right)\left(1 + \theta\lambda\Delta\right).\end{aligned}$$

The region of stability can be verified to be

$$\left|1 + \lambda\Delta\left(1 + \theta\lambda\Delta\right)\right| < 1, \quad \mathrm{Re}\left(\lambda\right) < 0.$$

For $\theta = \frac{1}{2}$, the stability criterion above is identical to that of the classical *Heun scheme* (or *modified trapezoidal scheme*) used for ordinary differential equations. Indeed, the predictor-corrector scheme above can be seen as an adaptation of this scheme for SDEs. We note that SDE adaptations of more sophisticated ODE solvers (such as Runge-Kutta) are also possible, but this goes beyond the scope of this text.

3.2.6 Ito-Taylor Expansions and Higher-Order Schemes

Despite our various efforts at centering derivatives, none of the schemes listed above theoretically attain second-order weak convergence. To develop such schemes, we need to delve further into adapting classical Taylor expansions to the rules of stochastic (Ito) calculus. As we shall ultimately not have much use for higher-order schemes, we keep the treatment informal and limit ourselves to the scalar case where $p = d = 1$ in (3.26).

3.2.6.1 Ordinary Taylor Expansion of ODEs

To gain intuition, start by setting $\sigma = 0$ in (3.26), such that we first deal with an ordinary ODE

$$dX(t) = \mu\left(t, X(t)\right) dt. \tag{3.43}$$

For a given value of t, we can use Taylor's theorem to write

$$X\left(t+\Delta\right) = X(t) + \frac{dX(t)}{dt}\Delta + \frac{1}{2}\frac{d^2X(t)}{dt^2}\Delta^2 + O\left(\Delta^3\right),$$

where we stop at order $O(\Delta^3)$. We notice that

$$\frac{dX(t)}{dt} = \mu\left(t, X(t)\right)$$

and

$$\begin{aligned}\frac{d^2X(t)}{dt^2} &= \frac{\partial}{\partial t}\mu\left(t, X(t)\right) + \frac{\partial}{\partial x}\mu\left(t, X(t)\right)\cdot\frac{dX(t)}{dt} \\ &= \left(\frac{\partial}{\partial t} + \mu\left(t, X(t)\right)\frac{\partial}{\partial x}\right)\mu\left(t, X(t)\right).\end{aligned}$$

Setting

$$\mathcal{L} \triangleq \frac{\partial}{\partial t} + \mu\frac{\partial}{\partial x},$$

we thus have

$$X\left(t+\Delta\right) = X(t) + \mu\left(t, X(t)\right)\Delta + \frac{1}{2}\mathcal{L}\mu\left(t, X(t)\right)\Delta^2 + O\left(\Delta^3\right). \tag{3.44}$$

Another way to develop (3.44) proceeds by iteration on the integral representation

$$X(t+\Delta) = X(t) + \int_t^{t+\Delta}\mu\left(u, X(u)\right) du. \tag{3.45}$$

First we recognize that (as seen above)

$$d\mu\left(t, X(t)\right) = \mathcal{L}\mu\left(t, X(t)\right) dt$$

such that

$$\mu\left(u, X(u)\right) = \mu\left(t, X(t)\right) + \int_t^u \mathcal{L}\mu\left(s, X(s)\right) ds, \quad u > t. \tag{3.46}$$

Inserting this into (3.45) gives

$$X\left(t+\Delta\right) = X(t) + \mu\left(t, X(t)\right)\int_t^{t+\Delta} du + \int_t^{t+\Delta}\int_t^u \mathcal{L}\mu\left(s, X(s)\right) ds\, du.$$

Applied to $\mathcal{L}\mu(s, X(s))$ the steps that lead to (3.46) yield

$$\mathcal{L}\mu\left(s, X(s)\right) = \mathcal{L}\mu\left(t, X(t)\right) + \int_t^s \mathcal{L}^2\mu\left(v, X(v)\right)\, dv, \quad s > t,$$

such that

$$\begin{aligned} X\left(t+\Delta\right) &= X(t) + \mu\left(t, X(t)\right) \int_t^{t+\Delta} du + \mathcal{L}\mu\left(t, X(t)\right) \int_t^{t+\Delta} \int_t^u ds\, du \\ &\quad + \int_t^{t+\Delta} \int_t^u \int_t^s \mathcal{L}^2\mu\left(v, X(v)\right)\, dv\, ds\, du \\ &= X(t) + \mu\left(t, X(t)\right)\Delta + \frac{1}{2}\mathcal{L}\mu\left(t, X(t)\right)\Delta^2 + O\left(\Delta^3\right), \end{aligned} \tag{3.47}$$

which is just (3.44). We can continue the iteration to arbitrary high order.

3.2.6.2 Ito-Taylor Expansions

One may wonder why in the previous section we bothered with the integral representation of Taylor's theorem when the usual (differential) Taylor expansion lead to the correct result. The reason is that the integral approach can be extended to SDEs, leading to stochastic *Ito-Taylor expansions.* To give a flavor of these, reintroduce a diffusion term to (3.43), and start out with the integral representation

$$X(t+\Delta) = X(t) + \int_t^{t+\Delta} \mu\left(u, X(u)\right) du + \int_t^{t+\Delta} \sigma\left(u, X(u)\right)\, dW(u). \tag{3.48}$$

Applying Ito's lemma to μ gives (compare to (3.46))

$$\mu\left(u, X(u)\right) = \mu\left(t, X(t)\right) + \int_t^u \mathcal{L}_0\mu\left(s, X(s)\right)\, ds + \int_t^u \mathcal{L}_1\mu\left(s, X(s)\right)\, dW(s), \tag{3.49}$$

where

$$\mathcal{L}_0 \triangleq \frac{\partial}{\partial t} + \mu\frac{\partial}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2}{\partial x^2}, \quad \mathcal{L}_1 \triangleq \sigma\frac{\partial}{\partial x}.$$

Similarly,

$$\sigma\left(u, X(u)\right) = \sigma\left(t, X(t)\right) + \int_t^u \mathcal{L}_0\sigma\left(s, X(s)\right)\, ds + \int_t^u \mathcal{L}_1\sigma\left(s, X(s)\right)\, dW(s). \tag{3.50}$$

Plugging (3.49) and (3.50) into (3.48) yields

$$\begin{aligned} X(t+\Delta) &= X(t) + \mu\left(t, X(t)\right) \int_t^{t+\Delta} du + \sigma\left(t, X(t)\right) \int_t^{t+\Delta} dW(u) + R_1 \\ &= X(t) + \mu\left(t, X(t)\right)\Delta + \sigma\left(t, X(t)\right)\left(W(t+\Delta) - W(t)\right) + R_1, \end{aligned} \tag{3.51}$$

where the remainder R_1 is

$$\begin{aligned} R_1 = &\int_t^{t+\Delta}\int_t^u \mathcal{L}_0\mu\left(s, X(s)\right)\, ds\, du \\ &+ \int_t^{t+\Delta}\int_t^u \mathcal{L}_1\mu\left(s, X(s)\right)\, dW(s)\, du \\ &+ \int_t^{t+\Delta}\int_t^u \mathcal{L}_0\sigma\left(s, X(s)\right)\, ds\, dW(u) \\ &+ \int_t^{t+\Delta}\int_t^u \mathcal{L}_1\sigma\left(s, X(s)\right)\, dW(s)\, dW(u). \end{aligned}$$

As for the ODE example above, we can repeat this procedure arbitrarily many times. Going just one step further, we arrive at

$$\begin{aligned} X(t+\Delta) = X(t) &+ \mu\left(t, X(t)\right)\Delta + \sigma\left(t, X(t)\right)\left(W(t+\Delta) - W(t)\right) \\ &+ \mathcal{L}_0\mu\left(t, X(t)\right)\frac{1}{2}\Delta^2 \\ &+ \mathcal{L}_1\mu\left(t, X(t)\right)\int_t^{t+\Delta}\int_t^u dW(s)\, du \\ &+ \mathcal{L}_0\sigma\left(t, X(t)\right)\int_t^{t+\Delta}\int_t^u ds\, dW(u) \\ &+ \mathcal{L}_1\sigma\left(t, X(t)\right)\int_t^{t+\Delta}\int_t^u dW(s)\, dW(u) + R_2, \end{aligned} \tag{3.52}$$

where R_2 contains triple integrals over t and W.

Stochastic Taylor expansions can be continued to arbitrary order, but we shall not go any further.

3.2.6.3 Milstein Second-Order Discretization Scheme

Discarding the remainder R_1 in the one-step iteration (3.51) is seen to lead to the Euler scheme (see Section 3.2.3), known to have weak convergence order $\beta = 1$. Under additional regularity (see Talay [1984]) of μ and σ, discarding the remainder R_2 in the higher-order expansion (3.52) can form the basis of a discretization scheme with weak order $\beta = 2$. For us to implement such a scheme, however, we need to concern ourselves with the simulation of the three stochastic double integrals figuring in (3.52). We go through the integrals in order below.

First,

$$\begin{aligned} I_{(1,1)} &\triangleq \int_t^{t+\Delta}\int_t^u dW(s)\, dW(u) \\ &= \int_t^{t+\Delta}\left(W(u) - W(t)\right) dW(u) = \frac{1}{2}\left(W(t+\Delta) - W(t)\right)^2 - \frac{1}{2}\Delta, \end{aligned} \tag{3.53}$$

where we have used the fact that

$$\int_0^t W(u)\,dW(u) = \frac{1}{2}W(t)^2 - \frac{1}{2}t,$$

as can be verified by Ito's lemma. Second,

$$\begin{aligned} I_{(0,1)} &\triangleq \int_t^{t+\Delta}\int_t^u ds\,dW(u) = \int_t^{t+\Delta}(u-t)\,dW(u) \\ &= \Delta\,(W(t+\Delta)-W(t)) - \int_t^{t+\Delta}(W(u)-W(t))\,du \\ &\triangleq \Delta\,(W(t+\Delta)-W(t)) - I_{(1,0)}, \end{aligned} \tag{3.54}$$

where we have used the integration-by-parts formula

$$\int_0^t u\,dW(u) = tW(t) - \int_0^t W(u)\,du,$$

which follows from applying Ito's lemma to $tW(t)$. In (3.54), the remaining integral $I_{(1,0)}$ on the right-hand-side is the same as the final double integral in (3.52), namely

$$I_{(1,0)} \triangleq \int_t^{t+\Delta}\int_t^u dW(s)\,du = \int_t^{t+\Delta}(W(u)-W(t))\,du.$$

Reversing the order of integration, we get

$$\begin{aligned} I_{(1,0)} &= \int_t^{t+\Delta}\int_t^u dW(s)\,du \\ &= \int_t^{t+\Delta}\int_u^{t+\Delta} ds\,dW(u) = \int_t^{t+\Delta}(t+\Delta-u)\,dW(u) \end{aligned}$$

so we see, from Theorem 1.1.3 and the discussion in Section 1.6 on linear SDEs, that $I_{(1,0)}$ is Gaussian with mean 0 and variance

$$\mathrm{Var}\left(I_{(1,0)}\right) = \int_t^{t+\Delta}(t+\Delta-u)^2\,du = \frac{1}{3}\Delta^3.$$

The covariance between $I_{(1,0)}$ and $W(t+\Delta)-W(t)$ can be computed as

$$\mathrm{Cov}\left(I_{(1,0)}, W(t+\Delta)-W(t)\right) = \int_t^{t+\Delta}(t+\Delta-u)\,du = \frac{1}{2}\Delta^2.$$

With the results above, we can cast the Taylor expansion (3.52) in the form of a simulation scheme (μ, σ, and their derivatives are to be evaluated at $t = i\Delta$ and $X = \widehat{X}(i\Delta)$),

$$\widehat{X}_{i+1} = \widehat{X}_i + \mu\Delta + \sigma Z_{i,1}\sqrt{\Delta} + \mathcal{L}_0\mu\frac{1}{2}\Delta^2 + \mathcal{L}_1\mu Z_{i,2}\sqrt{\frac{1}{3}\Delta^3}$$
$$+ \mathcal{L}_0\sigma\left[\Delta Z_{i,1}\sqrt{\Delta} - Z_{i,2}\sqrt{\frac{1}{3}\Delta^3}\right] + \mathcal{L}_1\sigma\left(\frac{1}{2}Z_{i,1}^2\Delta - \frac{1}{2}\Delta\right), \quad (3.55)$$

where $Z_{i,1}$ and $Z_{i,2}$ are sequences of $\mathcal{N}(0,1)$ Gaussian variables with pairwise correlation

$$\rho\left(Z_{i,1}, Z_{i,2}\right) = \frac{\frac{1}{2}\Delta^2}{\sqrt{\frac{1}{3}\Delta^3}\sqrt{\Delta}} = \sqrt{\frac{3}{4}}.$$

The scheme above is known as the *Milstein scheme*. As mentioned earlier, the scheme has weak convergence order 2 under fairly strong regularity assumptions on μ and σ. We note that in the literature on SDE simulation, the Milstein scheme is often presented in a simplified form with the integral $I_{(1,0)}$ simulated as

$$\mathrm{E}\left(\int_t^{t+\Delta}\int_t^u dW(s)\,du \middle| W(t), W(t+\Delta)\right) = \frac{1}{2}\Delta\left(W(t+\Delta) - W(t)\right),$$

which corresponds to replacing $Z_{i,2}\sqrt{\Delta^3/3}$ with $\frac{1}{2}\Delta Z_{i,1}$ in (3.55). See Kloeden and Platen [2000] for a discussion of why this type of simplification does not affect the weak convergence order. The same source also contains a full discussion of how to extend the Milstein scheme to multi-dimensional SDEs.

3.2.7 Other Second-Order Schemes

The need to explicitly compute derivatives of the functions μ and σ often makes the Milstein scheme inconvenient to apply. High-order simulation schemes that substitute finite difference approximations for derivatives exist, and retain second (or higher) order weak convergence, are surveyed in Kloeden and Platen [2000]. To give an example of such a scheme, consider the scalar case $d = p = 1$ and assume that SDE coefficient functions μ and σ are function of x only. A derivative-free scheme that achieves second-order weak convergence is (from Kloeden and Platen [2000], Chapter 15)

$$\begin{aligned}\widehat{X}_{i+1} = \widehat{X}_i &+ \frac{1}{2}\left(\mu\left(\overline{X}\right) + \mu\left(\widehat{X}_i\right)\right)\Delta \\ &+ \frac{1}{4}\left(\sigma\left(\overline{X}^+\right) + \sigma\left(\overline{X}^-\right) + 2\sigma\left(\widehat{X}_i\right)\right)Z_i\sqrt{\Delta} \\ &+ \frac{1}{4}\left(\sigma\left(\overline{X}^+\right) - \sigma\left(\overline{X}^-\right)\right)\left(Z_i^2\Delta - \Delta\right)\Delta^{-1/2}, \end{aligned} \quad (3.56)$$

where the Z_i's are a sequence of $\mathcal{N}(0,1)$ Gaussian variables, and

$$\overline{X} = \widehat{X}_i + \mu\left(\widehat{X}_i\right)\Delta + \sigma\left(\widehat{X}_i\right) Z_i\sqrt{\Delta},$$
$$\overline{X}^{\pm} = \widehat{X}_i + \mu\left(\widehat{X}_i\right)\Delta \pm \sigma\left(\widehat{X}_i\right)\sqrt{\Delta}.$$

Comparison of (3.56) with the simplified Milstein scheme in the previous section shows that (3.56) avoids derivatives by using additional supporting values $\overline{X}$ and $\overline{X}^{\pm}$.

Another, quite different, approach to avoid explicit derivatives applies the classical idea of *Richardson extrapolation* to the Euler scheme. This idea was proposed by Talay and Tubaro [1990] and takes advantage of the fact that, under additional regularity conditions, the error of the Euler scheme can be sharpened beyond (3.30) (with $\beta = 1$) to

$$\mathrm{E}\left(g\left(\widehat{X}(T)\right)\right) = \mathrm{E}\left(g\left(X(T)\right)\right) + c\Delta + O\left(\Delta^2\right), \tag{3.57}$$

for a constant c. Let $\widehat{X}_\Delta$ and $\widehat{X}_{2\Delta}$ be estimates of X based on Euler discretizations with time steps of Δ and 2Δ, respectively. Provided that (3.57) holds, we can write

$$2\mathrm{E}\left(g\left(\widehat{X}_\Delta(T)\right)\right) - \mathrm{E}\left(g\left(\widehat{X}_{2\Delta}(T)\right)\right) = \mathrm{E}\left(g\left(X(T)\right)\right) + O\left(\Delta^2\right), \tag{3.58}$$

which is our second-order extrapolation formula. As the Euler scheme is simple to set up, the extrapolation scheme is an attractive alternative to other second-order techniques. In practice, however, the convergence of the Euler scheme may not always be smooth enough to make (3.58) work well. Numerical experiments will nearly always be necessary (as is also the case of the Ito-Taylor schemes, for that matter).

A final word about generation of $\widehat{X}_\Delta$ and $\widehat{X}_{2\Delta}$ in the Richardson extrapolation scheme. To avoid duplication of work, we discretize time in buckets of Δ and generate both $\widehat{X}_\Delta$ and $\widehat{X}_{2\Delta}$ simultaneously, combining time steps in pairs for the purpose of generating $\widehat{X}_{2\Delta}$. That is, if we use Gaussian increments of $Z_1\sqrt{\Delta}$, $Z_2\sqrt{\Delta}$,... for $\widehat{X}_\Delta$, we use $(Z_1 + Z_2)\sqrt{\Delta}$, $(Z_3 + Z_4)\sqrt{\Delta}$,... for $\widehat{X}_{2\Delta}$. Not only do we save work by re-using Gaussian draws, we most likely also reduce the statistical error of our Monte Carlo estimate of the difference $2\mathrm{E}(g(\widehat{X}_\Delta(T))) - \mathrm{E}(g(\widehat{X}_{2\Delta}(T)))$ by raising correlation between $g(\widehat{X}_\Delta(T))$ and $g(\widehat{X}_{2\Delta}(T))$. We shall return to this idea in Section 3.3.1.

3.2.8 Bias vs. Monte Carlo Error

When we use an m-step discretization scheme in an n-path Monte Carlo run, we are exposed to two types of errors on the expectation we are trying to evaluate: i) the statistical Monte Carlo error e_s (the standard error); and ii) a bias e_b, originating from the discretization scheme. Raising n will reduce the standard error, but will not affect the bias which can only be

reduced by increasing the number of steps m in the time discretization scheme. Raising m and/or n obviously involves a computational cost, so let us briefly consider explicitly the trade-offs involved in simultaneously reducing bias and standard error.

Assume first that the discretization scheme has weak order β. Proceeding informally, we interpret this as implying

$$e_b = c_b \Delta^\beta,$$

for some constant c_b. Also, we know that the variance of e_s is

$$\mathrm{Var}(e_s) = \frac{c_s}{n},$$

for a constant c_s. The total computing time τ is reasonably assumed to be proportional to nm or, using the fact that $\Delta = T/m$,

$$\tau = n\frac{c_\tau}{\Delta} \tag{3.59}$$

for some constant c_τ. For a given computing budget τ, consider minimizing the total mean-square error (MSE) $c_b^2\Delta^{2\beta} + \frac{c_s}{n}$. Using (3.59) to eliminate a variable, the optimization problem is

$$\min_\Delta \left(c_b^2 \Delta^{2\beta} + \frac{c_s c_\tau}{\tau \Delta} \right).$$

Let Δ^* be the value of Δ at which the minimum MSE is attained. Δ^* is seen to satisfy

$$\Delta^* = C\tau^{-\frac{1}{2\beta+1}}, \quad C \triangleq \left(\frac{c_s c_\tau}{2\beta c_b^2} \right)^{\frac{1}{2\beta+1}}, \tag{3.60}$$

such that the minimum MSE becomes

$$C'\tau^{-\frac{2\beta}{2\beta+1}} \tag{3.61}$$

for yet another constant C'.

Equations (3.60) and (3.61) reveal a number of structural characteristics of Monte Carlo simulation of discretized SDEs. For instance, according to (3.61), the optimal root-mean-square (RMS) error behaves with the computing time τ as

$$\mathrm{RMS} \propto \tau^{-\frac{\beta}{2\beta+1}}. \tag{3.62}$$

The computational cost of working with SDEs that are not explicitly solvable are quantified by (3.62). For an unbiased (that is, exact) SDE simulation scheme, $\beta = \infty$ and the optimal RMS error converges at the rate of $\tau^{-\frac{1}{2}}$, consistent with the results of Section 3.1. However, for an Euler scheme ($\beta = 1$) the RMS error convergence rate is lowered to $\tau^{-\frac{1}{3}}$.

Equation (3.60) in principle tells us how to optimally allocate resources between the competing objectives of a lower bias and a lower standard

error. Let m^* be defined through $\Delta^* = T/m^*$ and let n^* be defined through $\tau = c_\tau n^* m^*/T$. After a few rearrangements, we find the intuitive result

$$\sqrt{n^*} = C''(m^*)^\beta,$$

where C'' is a constant independent of τ. When we increase or decrease our computing budget, it is thus reasonable to allocate resources in such a way that we keep the factor $n^{1/2}m^{-\beta}$ constant. More detailed discussion, as well as asymptotic limit results, can be found in Duffie and Glynn [1995].

3.2.9 Sampling of Continuous Process Extremes

We round off our discussion of path simulation schemes by considering the pricing of options that depend on continuously or high-frequency sampled extremes of an SDE. We focus on the scalar case, with our SDE given as

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

where both X and W are 1-dimensional (i.e., $p = d = 1$). We also assume that the SDE is Euler-discretized according to

$$\widehat{X}_{i+1} = \widehat{X}_i + \mu\left(i\Delta, \widehat{X}_i\right)\Delta + \sigma\left(i\Delta, \widehat{X}_i\right)\sqrt{\Delta}Z_i, \quad i = 0, 1, \ldots, m-1,$$

with $m\Delta = T$.

On the interval $[0, T]$, let the maximum and minimum values of $X(t)$ be denoted $M_{[0,T]}$ and $m_{[0,T]}$, respectively. That is,

$$M_{[0,T]} \triangleq \max_{0 \le t \le T} X(t); \quad m_{[0,T]} \triangleq \min_{0 \le t \le T} X(t).$$

To give examples of options that depend on $M_{[0,T]}$ and $m_{[0,T]}$, consider for instance the up-and-out call option we encountered in Section 2.7. With a knock-out barrier of H and a terminal strike of K, the terminal maturity payout can be written as[11]

$$g(T) = 1_{\{M_{[0,T]} < H\}}\,(X(T) - K)^+.$$

A *double-barrier knock-out call option* with an upper barrier of H and a lower barrier of h pays

$$g(T) = 1_{\{m_{[0,T]} > h\}} 1_{\{M_{[0,T]} < H\}}\,(X(T) - K)^+.$$

Finally, a so-called *lookback call option* (see Section 2.7) pays

$$g(T) = \left(M_{[0,T]} - K\right)^+.$$

[11] If $X(t)$ is the logarithm of the asset price, we replace this expression by $g(T) = 1_{\{M_{[0,T]} \le e^H\}}(e^{X(T)} - K)^+$, and similarly for the other payouts considered.

To price options such as those above, we must provide pathwise estimates of $M_{[0,T]}$ and $m_{[0,T]}$. Given our discretization schemes, natural estimators are

$$\widehat{M}_{[0,T]} = \max\left(X(0), \widehat{X}_1, \ldots, \widehat{X}_m\right), \tag{3.63}$$

$$\widehat{m}_{[0,T]} = \min\left(X(0), \widehat{X}_1, \ldots, \widehat{X}_m\right). \tag{3.64}$$

Even in cases where the discretization scheme itself is perfectly unbiased, it is clear that these estimators will understate the range of the extremes of $X(t)$, by consistently failing to account for the movement (the "overshoot" and "undershoot") of X between sample points $i\Delta$, $i = 0, 1, \ldots, m$. As a consequence, for each simulated path in an otherwise unbiased discretization scheme, almost surely

$$\widehat{M}_{[0,T]} < M_{[0,T]}, \quad \widehat{m}_{[0,T]} > m_{[0,T]}.$$

As shown in Andersen and Brotherton-Ratcliffe [1996], the bias introduced can be very significant, even if Δ is quite small. For instance, for a 1 year lookback option, Andersen and Brotherton-Ratcliffe [1996] report that even daily sampling produces a 6% price error.

To improve the price estimates of options that depend on continuously sampled extremes, we should alter (3.63) and (3.64) to take into consideration movements between sample dates. This can be accomplished by the Brownian bridge technique introduced in Andersen and Brotherton-Ratcliffe [1996] (see also Broadie et al. [1997]). Let us focus on a particular bucket $[i\Delta, (i+1)\Delta]$ and assume, consistent with the Euler scheme, that $\widehat{X}(t)$, $t \in [i\Delta, (i+1)\Delta]$, is a Gaussian process with conditional moments

$$\mathrm{E}\left(\widehat{X}(t) - \widehat{X}_i \middle| \widehat{X}_i\right) = \mu\left(i\Delta, \widehat{X}_i\right)(t - i\Delta), \quad t \in [i\Delta, (i+1)\Delta);$$

$$\mathrm{Var}\left(\widehat{X}(t) - \widehat{X}_i \middle| \widehat{X}_i\right) = \sigma\left(i\Delta, \widehat{X}_i\right)^2 (t - i\Delta), \quad t \in [i\Delta, (i+1)\Delta).$$

Assume that we have already simulated $\widehat{X}_i$ and $\widehat{X}((i+1)\Delta)$ by the Euler scheme above. Conditional on *both* $\widehat{X}_i$ and $\widehat{X}_{i+1}$, the process for $\widehat{X}(t)$, $t \in [i\Delta, (i+1)\Delta]$ is a Gaussian process "pinned" at the levels $\widehat{X}_i$ and $\widehat{X}_{i+1}$. The resulting process is known as a *Brownian bridge* with diffusion coefficient $\sigma(i\Delta, \widehat{X}_i)$. Let $\widehat{M}_i^c$ ($\widehat{m}_i^c$) be defined as the continuously sampled maximum (minimum) of $\widehat{X}(t)$ on $[i\Delta, (i+1)\Delta]$. The following lemma is a special case of a result in Andersen and Brotherton-Ratcliffe [1996]:

Lemma 3.2.1.

$$\mathrm{P}\left(\widehat{M}_i^c \le s \middle| \widehat{X}_i, \widehat{X}_{i+1}\right) = 1 - \xi_i(s), \quad s > \max\left(\widehat{X}_i, \widehat{X}_{i+1}\right),$$

$$\mathrm{P}\left(\widehat{m}_i^c \le s \middle| \widehat{X}_i, \widehat{X}_{i+1}\right) = \xi_i(s), \quad s < \min\left(\widehat{X}_i, \widehat{X}_{i+1}\right),$$

where

$$\xi_i(s) \triangleq \exp\left(\frac{2\left(s-\widehat{X}_i\right)\left(\widehat{X}_{i+1}-s\right)}{\sigma\left(i\Delta,\widehat{X}_i\right)^2\Delta}\right).$$

We can use the result of the lemma in a number of ways. Most obviously, we can apply it to sample $\widehat{M}_i^c$ and $\widehat{m}_i^c$ directly, by the inverse transform method (see Section 3.1.1.1). To illustrate, consider for instance sampling $\widehat{M}_i^c$. Having first drawn $\widehat{X}_i$ and $\widehat{X}_{i+1}$ by usual means, we draw an additional independent $\mathcal{U}(0,1)$ uniform variable U_i, and set

$$1-\xi_i(\widehat{M}_i^c) = U_i$$

or, after a few rearrangements,

$$\widehat{M}_i^c = \frac{1}{2}\left(\widehat{X}_{i+1}+\widehat{X}_i\right)+\frac{1}{2}\sqrt{\left(\widehat{X}_{i+1}-\widehat{X}_i\right)^2-2\sigma\left(i\Delta,\widehat{X}_i\right)^2\Delta\ln\left(1-U_i\right)}.$$

This procedure can be repeated for $i=0,1,\ldots,m-1$, giving us the improved estimator for the maximum of X over $[0,T]$,

$$\widehat{M}_{[0,T]}^c = \max\left(\widehat{M}_0^c,\ldots,\widehat{M}_{m-1}^c\right).$$

For options depending on both the minimum and maximum (such as double barrier options and the double lookback options in He et al. [1998]), the necessary extensions required for joint sampling of minimum and maximum are developed in Andersen [1998].

For barrier options, we note that locating $\widehat{M}_i^c$ and $\widehat{m}_i^c$ directly is typically not necessary, as it suffices to check locally whether the barrier is breached. For an up-and-out knock-out option with barrier H, for each interval it thus suffices to check whether $\widehat{M}_i^c > H$ which, conditional on $\widehat{X}_i$ and $\widehat{X}_{i+1}$, happens with likelihood $\xi_i(H)$. So, provided that $\widehat{X}_i$ and $\widehat{X}_{i+1}$ are both below H, determining whether the barrier was nevertheless breached in $[i\Delta,(i+1)\Delta]$ is a matter of drawing a uniform variable U_i and setting

$$1_{\left\{\widehat{M}_i^c \geq H\right\}} = 1_{\{U_i \leq \xi_i(H)\}}. \tag{3.65}$$

This scheme is easily extended to time-dependent barriers and to cases where there are rebates[12]. As pointed out in Glasserman and Staum [2001], for

[12] To get the timing of rebate payments right, the exact time that the barrier is breached must, in principle, be located. Andersen and Brotherton-Ratcliffe [1996] list analytical Brownian bridge hitting time results that can be used for this purpose. For reasonably fine discretizations, it will often suffice to set the hitting time to, say, the mid-point of the time bucket where the barrier is known to be breached.

Markov processes and the special case of barrier options with no rebates, one can in fact avoid drawing U_i's altogether, as

$$\begin{aligned}
&\mathrm{E}\left(g\left(\widehat{X}_m\right) 1_{\left\{\widehat{M}^c_{[0,T]}<H\right\}} \middle| \widehat{X}_0, \widehat{X}_1, \ldots, \widehat{X}_m \right) \\
&\qquad = \mathrm{E}\left(g\left(\widehat{X}_m\right) \prod_{i=0}^{m-1} 1_{\left\{\widehat{M}^c_i<H\right\}} \middle| \widehat{X}_0, \widehat{X}_1, \ldots, \widehat{X}_m \right) \\
&\qquad = g\left(\widehat{X}_m\right) \prod_{i=0}^{m-1} \mathrm{E}\left(1_{\left\{\widehat{M}^c_i<H\right\}} \middle| \widehat{X}_i, \widehat{X}_{i+1} \right).
\end{aligned}$$

Here,

$$\mathrm{E}\left(1_{\left\{\widehat{M}^c_i<H\right\}} \middle| \widehat{X}_i, \widehat{X}_{i+1} \right) = \begin{cases} 0, & \widehat{X}_i \geq H \text{ or } \widehat{X}_{i+1} \geq H, \\ 1-\xi_i(H), & \widehat{X}_i < H \text{ and } \widehat{X}_{i+1} < H. \end{cases}$$

In other words, rather than explicitly simulating the indicator functions (3.65), it suffices to adjust the terminal payout by the product of conditional survival probabilities along the path. This scheme is an example of *conditional Monte Carlo*, a variance-reduction scheme discussed in more detail in Section 25.2 and in Boyle et al. [1997]. One potential drawback of the scheme is the fact that we typically need to continue the paths for a longer period of time before a barrier crossing is detected and the path can be stopped.

We round off this section with a few comments. First, we note that the schemes above assume that X is well approximated by a Gaussian process. In some applications the geometric Brownian motion, say, may be a more appropriate model. In Lemma 3.2.1, we can easily accommodate this by simply replacing $\widehat{M}^c_i$, $\widehat{m}^c_i$, s, $\widehat{X}_i$, and $\widehat{X}_{i+1}$ with $\ln \widehat{M}^c_i$, $\ln \widehat{m}^c_i$, $\ln s$, $\ln \widehat{X}_i$, and $\ln \widehat{X}_{i+1}$. Other transformations are handled the same way.

Secondly, it should be pointed out that many real options are, in fact, not sampled continuously but rather at some finite but high frequency, often daily. Running an Euler scheme with daily discretization is obviously computationally inefficient. Fortunately, we can often use our scheme above as part of a Richardson-type interpolation idea. Indeed, it can often be established (see Andersen and Brotherton-Ratcliffe [1996]) that options on process extremes converge as $O(\sqrt{\Delta})$. If we first compute a price estimate $\widehat{V}^{\Delta}$ based on a relatively coarse value of Δ, we can then write

$$\widehat{V}^{\Delta^*} \approx \widehat{V}^c + (\widehat{V}^{\Delta} - \widehat{V}^c)\Delta^*/\Delta, \quad \Delta^* < \Delta,$$

where $\widehat{V}^c$ is the continuously monitored price computed by the scheme outlined above. We have here implicitly made the assumption that the regular Euler bias is small relative to the bias induced by using the wrong sampling frequency. The idea above is developed further in Chapter V of Andersen [1996], where a number of numerical results can also be found.

And finally, one may wonder whether it is possible to deal with a continuously monitored barrier option in a discrete-time simulation by adjusting the *barrier*, rather than the underlying *process*. As it turns out, this is indeed possible. Specifically, in the Black-Scholes-Merton model with volatility σ, let $V^c(H)$, $V^\Delta(H)$ be the values of a continuously and discretely sampled barrier options with barrier H, respectively. Assuming that the discrete sampling happens on a time grid with spacing Δ, we have the following result from Broadie et al. [1997]:

Theorem 3.2.2. *The following holds,*

$$V^\Delta(H) = V\left(He^{\pm\beta\sigma\sqrt{\Delta}}\right) + o\left(\sqrt{\Delta}\right),$$

where $+$ *applies if* $H > X(0)$, $-$ *applies if* $H < X(0)$, *and* $\beta = \zeta(1/2)/\sqrt{2\pi} \approx 0.5826$, *with* $\zeta(\cdot)$ *being the Riemann zeta function.*

According to this result we can price a continuous barrier option by evaluating a discrete barrier option instead (e.g., one where the barrier monitoring takes place only on the simulation dates of the Monte Carlo scheme used), but with the discrete barrier level shifted according to the theorem. Theorem 3.2.2 can also be to save computation time by, say, turning a barrier option with daily observations into an option with quarterly observations, as the theorem shows that

$$V^{\Delta^*}(H) \approx V^\Delta\left(He^{\pm\beta\sigma(\sqrt{\Delta^*}-\sqrt{\Delta})}\right).$$

While the result of Theorem 3.2.2 is only proved for the log-normal process, practical experience shows that it is robust across a wide variety of models. A similar approach exists for lookback options, see Broadie et al. [1999].

3.2.10 PCA and Bridge Construction of Brownian Motion Paths

3.2.10.1 Brownian Bridge and Quasi-Random Sequences

To close out the section on sample path simulation, let us address alternative ways of generating sample paths of Brownian motion. So far, to produce a sample of the vector $\mathbf{W} = (W(\Delta), W(2\Delta), \ldots, W(m\Delta))^\top$, we have relied exclusively on the forward recursion

$$W(i\Delta + \Delta) = W(i\Delta) + Z_i\sqrt{\Delta}, \quad W(0) = 0, \tag{3.66}$$

where $Z_0, Z_1, \ldots, Z_{m-1}$ is a sequence of independent standard Gaussian variables. Rather than filling out the elements of $\mathbf{W}$ in order, we may, for instance, rely on a *Brownian bridge (BB) construction* where we first

sample the end-point $W(m\Delta)$, then sample the mid-point[13] $W(\lfloor m/2 \rfloor \Delta)$ *conditional* on $W(m\Delta)$, and so forth. In executing this scheme, we can use the easily proven result below.

Lemma 3.2.3. *Let $\underline{t} < t < \overline{t}$. Conditional on $W(\underline{t})$ and $W(\overline{t})$, $W(t)$ is Gaussian with moments*

$$\mathrm{E}\left(W(t)|W(\underline{t}) = \underline{w}, W(\overline{t}) = \overline{w}\right) = \underline{w} \times \frac{\overline{t} - t}{\overline{t} - \underline{t}} + \overline{w} \times \frac{t - \underline{t}}{\overline{t} - \underline{t}},$$

$$\mathrm{Var}\left(W(t)|W(\underline{t}) = \underline{w}, W(\overline{t}) = \overline{w}\right) = \frac{\left(\overline{t} - t\right)\left(t - \underline{t}\right)}{\overline{t} - \underline{t}}.$$

The BB scheme for construction of $\mathbf{W}$ relies on repeated application of the result in Lemma 3.2.3 to progressively fill in $\mathbf{W}$ in the "bisection" manner described above; consult any Monte Carlo textbook (e.g. Jäckel [2002] or Glasserman [2004]) if further details are required. As is the case for the standard scheme (3.66), a total of m standard Gaussian random variables are needed to construct a single sample of $\mathbf{W}$ by the Brownian bridge scheme, so the latter offers no computational advantage over the former. Why then use the Brownian bridge construction?

One important distinction between the BB construction and (3.66) is the fact that the Brownian bridge assigns different importance to the random numbers used to produce $\mathbf{W}$. For instance, the very first Gaussian number drawn in the BB technique *alone* determines the end-point $W(m\Delta)$ of the Brownian motion — and thereby establishes a significant part of the overall coarse structure of the path of W. Subsequent random number draws contribute by filling in the details of the W-path, with late draws adding only to the fine-structure of the path. In contrast, with (3.66) the end-point $W(m\Delta)$ is affected equally by the m random numbers $Z_0, Z_1, \ldots, Z_{m-1}$. In most financial problems the coarse shape of the path of W is more critical than finer details, so ultimately the BB technique allows us to identify and isolate the important features of the Brownian motion path. In some variance reduction techniques this can be important, as it allows us to focus computational effort on the random numbers that matter the most. Also, some variance reduction techniques that are known to work particularly well on low-dimensional problems can now be applied to the (low-dimensional) random numbers that contribute most to the sample path.

One relevant technique is *quasi-random sequences* (also known as *low-discrepancy sequences*), a method of generating points on the hypercube that are as "dispersed" as possible. A good survey of the underlying ideas and theory can be found in Jäckel [2002] or Glasserman [2004], with source code available (for the special case of *Sobol sequences*) in Press et al. [1992]; suffice to say that quasi-random sequences can, under some circumstances,

[13] $\lfloor x \rfloor$ denotes the integer part of a real variable x.

accelerate Monte Carlo convergence substantially[14]. It is well-known, however, that the efficacy of quasi-random sequences depend strongly on the problem dimension (here: m, the number of random numbers needed per path), and that the sequences deteriorate in higher dimensions. When quasi-random sequences are combined with BB simulation of the path, however, the (low-dimensional) points of the sequences that are well-distributed can be applied to generate — by the methods in Section 3.1.1 — the Gaussian samples that determine the coarse structure of the paths, whereas the poorly distributed (high-dimensional) parts of the sequence can be relegated to the generation of less important fine-structure details[15]. A full account of this idea can be found in Moskowitz and Caflisch [1996].

3.2.10.2 PC Construction

With the Brownian Bridge (BB) construction of Brownian motion, much of the variance of the sample paths $\mathbf{W}$ is explained by the values of the first few (Gaussian) random variables drawn in the path simulation. We recall from Section 3.1.3, however, that the *optimal* way to project the variation of a Gaussian vector onto a low-dimensional set of random variables is done through a principal component (PC) construction, rather than the Brownian bridge. To demonstrate how a PC construction of $\mathbf{W} = (W(\Delta), W(2\Delta), \ldots, W(m\Delta))^\top$ would proceed, we first notice that the $m \times m$ variance-covariance matrix Σ of $\mathbf{W}$ has elements

$$\Sigma_{i,j} = \mathrm{E}\left(W(i\Delta)\, W(j\Delta)\right) = \sqrt{\frac{\min(i,j)}{\max(i,j)}}, \quad i,j = 1,2,\ldots,m.$$

As shown in Åkesson and Lehoczy [1998], the eigenvalues of Σ can be found analytically to be

$$\lambda_i = \frac{\Delta}{4} \sin\left(\frac{\pi}{2} \cdot \frac{2i-1}{2m+1}\right)^{-2}, \quad i = 1,\ldots,m,$$

where $\lambda_1 > \lambda_2 > \ldots > \lambda_m$. Let e_i be the eigenvector associated with λ_i, then it is also known that $e_i = (e_{i,1}, e_{i,2}, \ldots, e_{i,m})^\top$, where

$$e_{i,j} = \frac{2}{\sqrt{2m+1}} \sin\left(j\pi \cdot \frac{2i-1}{2m+1}\right), \quad j = 1,\ldots,m.$$

From the results in Section 3.1.3, we know that we can write

[14] Theoretically from $O(1/\sqrt{N})$ to (nearly) $O(1/N)$. Comparative tests on actual finance problems can be found in, for instance, Brotherton-Ratcliffe [1994], Paskov and Traub [1995], and Joy et al. [1996].

[15] Alternatively we can use regular *pseudo-random numbers* for this.

$$\mathbf{W} = \sum_{i=1}^{m} Z_i \sqrt{\lambda_i} e_i, \tag{3.67}$$

where $Z_1, Z_2, \ldots, Z_m$ is a sequence of independent standard Gaussian random variables. This equation constitutes the principal components construction of the Brownian path, and it is characterized by the fact that for any $k \leq m$, the first k terms of (3.67) (that is, $\sum_{i=1}^{k} Z_i \sqrt{\lambda_i} e_i$) explain as much of the variance of $\mathbf{W}$ as is possible with k Gaussian variables. Even more so than for the Brownian bridge, the PC construction of a Brownian motion thus connects the overall shape of the Brownian path to a few of the Gaussian random variables Z_i, with the remaining random variables contributing only high-frequency details. As explained above, this can be useful in certain variance reduction techniques by allowing us to focus our attention and resources on just a few of the m random variables needed to simulate $\mathbf{W}$. We note that the PC construction is more expensive to compute than the BB technique (the latter is $O(m)$ whereas the former can be seen from (3.67) to be $O(m^2)$), so the optimality of the PC approach may, in some applications, be outweighed by its lack of speed.

While we developed the BB and PC constructions exclusively on an equidistant time grid, they easily extend to non-equidistant grids. When the grid is non-equidistant, the variance-covariance matrix of $\mathbf{W}$ has elements

$$\Sigma_{i,j} = \mathrm{E}\left(W(t_i)W(t_j)\right) = \sqrt{\frac{\min\left(t_i, t_j\right)}{\max\left(t_i, t_j\right)}}, \quad i, j = 1, 2, \ldots, m,$$

and eigenvectors and eigenvalues must then be found numerically, rather than through the analytical results listed earlier. Also, both the BB and PC techniques can easily be extended to the case of multi-variate Brownian motions, see Jäckel [2002].

Finally, for those interested in such matters, we note that in the limit $m \to \infty$, the PC construction of Brownian motion is known as the *Karhunen-Loeve decomposition.* In the continuous-time limit, the BB representation is sometimes known as a *Haar function* decomposition of Brownian motion.

3.3 Sensitivity Computations

In most finance applications, the fact that options must be dynamically hedged and risk managed requires us not only to produce an estimate of an option price, but also to compute reliable estimates of the sensitivity of the price with respect to the underlying state variables, as well as various other model parameters. In this section, we will present a number of methods for sensitivity computations by Monte Carlo methods. For each method, we use the problem of estimating the stock price delta of options in the Black-Scholes economy as a motivating example. We shall spend much more

time on sensitivity computations (by PDE and Monte Carlo methods) in Part V of this book, often in the context of particular interest rate products. Here we just give a flavor of things to come.

3.3.1 Finite Difference Estimates

3.3.1.1 Black-Scholes Delta

Consider a T-maturity European option on a dividend-free stock S in the Black-Scholes economy. Let the payout function be $g(S(T))$, and assume that the continuously compounded interest rate is a constant r. With $V(S_0)$ denoting the time 0 price of the option given $S(0) = S_0$, we are interested in computing

$$\frac{dV}{dS_0} = \lim_{h \to 0} \frac{V(S_0 + h) - V(S_0)}{h}. \tag{3.68}$$

In a Monte Carlo setting, we can approximate this derivative ("delta") by finite difference techniques as follows. First, for some fixed number ε draw random standard Gaussian variables Z and Z_ε, and set (see Section 3.2.1)

$$S(T) = S_0 \exp\left(\left(r - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z\right),$$

$$S_\varepsilon(T) = (S_0 + \varepsilon)\exp\left(\left(r - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z_\varepsilon\right),$$

where σ as always denotes the constant volatility of the stock. We then form the difference

$$\delta = e^{-rT}\varepsilon^{-1}\left(g\left(S_\varepsilon(T)\right) - g(S(T))\right),$$

such that δ constitutes a single-sample estimate for dV/dS_0. By generating n independent replications of δ and forming the sample average, we will obtain in the limit $n \to \infty$ the finite difference ratio

$$\frac{V(S_0 + \varepsilon) - V(S_0)}{\varepsilon}. \tag{3.69}$$

We know from Chapter 2 that this estimate will be biased relative to the true derivative dV/dS_0 by an amount of order $O(\varepsilon^2)$.

We have so far not mentioned whether the standard Gaussian variables Z and Z_ε should be independent or not. To analyze this, we need to consider the variance of the Monte Carlo estimator of (3.69). From Theorem 3.1.2, we know that for a finite number of trials n, the variance of our sample average will decrease as v_ε/n, where

$$\begin{aligned} v_\varepsilon &= \varepsilon^{-2}e^{-2rT}\mathrm{Var}\left(g\left(S_\varepsilon(T)\right) - g(S(T))\right) \\ &= \varepsilon^{-2}e^{-2rT}\left[\mathrm{Var}\left(g\left(S(T)\right)\right) + \mathrm{Var}\left(g\left(S_\varepsilon(T)\right)\right)\right. \\ &\quad \left. -2\mathrm{Cov}\left(g\left(S_\varepsilon(T)\right), g\left(S(T)\right)\right)\right]. \end{aligned}$$

If the random numbers Z and Z_ε are independent, $\mathrm{Cov}(g(S_\varepsilon(T)), g(S(T)))$ will be zero and

$$v_\varepsilon \approx 2\varepsilon^{-2} e^{-2rT} \mathrm{Var}\left(g\left(S(T)\right)\right).$$

Making ε approach zero — as is needed to reduce the bias of the finite difference approximation (3.69) — will cause v_ε grow at a rate of $O(\varepsilon^{-2})$. This is obviously not ideal as our Monte Carlo estimate will be swamped by noise if ε is picked too small. On the other hand, if we set Z and Z_ε to be *identical*, we see that

$$S_\varepsilon(T) = (S_0 + \varepsilon)\, S(T)/S_0$$

and would expect

$$\mathrm{Cov}\left(g\left(S_\varepsilon(T)\right), g(S(T))\right) > 0,$$

which would reduce v_ε relative to the independent case. For smooth g, a Taylor expansion in ε shows that

$$\begin{aligned} g\left(S_\varepsilon(T)\right) &= g\left(S(T)\right) + \left(S_\varepsilon(T) - S(T)\right) g'\left(S(T)\right) + \ldots \\ &= g\left(S(T)\right) + \varepsilon S(T)/S_0 \cdot g'\left(S(T)\right) + \ldots. \end{aligned}$$

If derivatives of g are bounded

$$\mathrm{Cov}\left(g\left(S_\varepsilon(T)\right), g\left(S(T)\right)\right) = \mathrm{Var}\left(g\left(S(T)\right)\right) + O\left(\varepsilon^2\right),$$

and similarly for $\mathrm{Var}(g(S_\varepsilon(T)))$. In other words,

$$v_\varepsilon = e^{-2rT} O(1) \tag{3.70}$$

which is a clear improvement over the earlier $O(\varepsilon^{-2})$ result.

The result (3.70) hinged on the payout function having bounded derivatives. We can, in fact, relax this considerably, to functions that are essentially just continuous in the stock price; see Section 3.3.2.2 for a discussion. For discontinuous payouts, however, (3.70) will not hold. To demonstrate, consider a digital option paying

$$g\left(S(T)\right) = 1_{\{S(T)>K\}},$$

for some strike K. With $Z = Z_\varepsilon$ we get (assuming $\varepsilon > 0$ and that the probability measure is P)

$$\begin{aligned} \mathrm{E}\left(\left[g\left(S_\varepsilon(T)\right) - g\left(S(T)\right)\right]^2\right) &= \mathrm{P}\left(S(T) \le K < S_\varepsilon(T)\right) \\ &= \mathrm{P}\left(S(T) \le K < \left(1 + \varepsilon/S_0\right) S(T)\right) = O(\varepsilon), \end{aligned}$$

compared with the $O(\varepsilon^2)$ result for smooth g.

3.3.1.2 General Case

To generalize the problem considered in the previous section, we consider a setting where a random variable Y depends on a parameter $\alpha \in \mathbb{R}$, in the sense that each value of α uniquely determines a scheme for the generation of Y. The random variable Y will typically represent a (discounted) option payout, and α is typically an initial value of an asset price (as in Section 3.3.1.1) or a parameter in the (vector) equations determining the dynamics of the underlying model. Let

$$V(\alpha) = \mathrm{E}\left(Y(\alpha)\right),$$

and consider the problem of determining $dV/d\alpha$.

In the basic finite difference Monte Carlo approximation to $dV/d\alpha$, we use the sample average of one-sided difference coefficients,

$$\overline{\delta}_n = \frac{\overline{Y}_n(\alpha+\varepsilon) - \overline{Y}_n(\alpha)}{\varepsilon},$$

where $\overline{Y}_n(\alpha)$ is the sample average of n realizations of $Y(\alpha)$. In the limit,

$$\lim_{n\to\infty} \overline{\delta}_n = \frac{V(\alpha+\varepsilon) - V(\alpha)}{\varepsilon} = dV/d\alpha + O(\varepsilon^2).$$

If we instead wish to use a central estimator

$$\overline{\delta}_n^c = \frac{\overline{Y}_n(\alpha+\varepsilon) - \overline{Y}_n(\alpha-\varepsilon)}{2\varepsilon},$$

we get

$$\lim_{n\to\infty} \overline{\delta}_n^c = \frac{V(\alpha+\varepsilon) - V(\alpha-\varepsilon)}{2\varepsilon} = dV/d\alpha + O(\varepsilon^3),$$

but now need to simulate an extra random variable (that is, $Y(\alpha-\varepsilon)$), increasing the computational cost.

In the generation of $\overline{Y}_n(\alpha \pm \varepsilon)$ and $\overline{Y}_n(\alpha)$, the individual samples of $Y(\alpha+\varepsilon)$, $Y(\alpha-\varepsilon)$, and $Y(\alpha)$ would typically be based on a series of draws of vector-valued support variables Z, with $Y(\alpha) = Y(Z;\alpha)$, and so forth. For instance, in an m-step Euler simulation of an SDE with d Brownian motions, each SDE path (and each outcome of Y) would involve $d \cdot m$ i.i.d. standard Gaussian variables $Z_1, \ldots, Z_{d\cdot m}$. The observations in the previous section tell us that to minimize variance we should use the same Z for $Y(\alpha+\varepsilon)$, $Y(\alpha-\varepsilon)$, and $Y(\alpha)$. In practice, this is often easiest to accomplish by simply using the same random number seed (see Section 3.1.1) in otherwise separate computations of each of the quantities $\overline{Y}_n(\alpha+\varepsilon)$, $\overline{Y}_n(\alpha-\varepsilon)$, and $\overline{Y}_n(\alpha)$. Assuming this so-called *common random number scheme* is followed, the variance analysis in Section 3.3.1.1 can be generalized to our setting, and we would expect that either i) $\mathrm{Var}(\overline{\delta}_n) = \mathrm{Var}(\overline{\delta}_n^c) = O(\varepsilon^{-1}n^{-1})$; or ii)

$\mathrm{Var}(\overline{\delta}_n) = \mathrm{Var}(\overline{\delta}_n^c) = O(n^{-1})$. Case ii) essentially requires a.s. continuity[16] of $Y(\alpha)$ with respect to α, as would be the case when Y represents a continuous option payout function. Case i) generally applies when Y represents a discontinuous option payout, such as the digital option considered in Section 3.3.1.1.

If case ii) above applies, the estimator variance is independent of ε, and ε should be picked as small as possible (a matter of machine precision) to minimize the $O(\varepsilon^2)$ and $O(\varepsilon^3)$ biases of $\overline{\delta}_n$ and $\overline{\delta}_n^c$. If the overhead of evaluating $Y(Z;\alpha)$ for given Z is small relative to the cost of generating Z, the central estimator $\overline{\delta}_n^c$ will dominate. For complicated payout functions, however, there may be situations when $\overline{\delta}_n$ is preferable, despite its slower convergence rate in ε. For case i), we must weigh bias against variance in a manner quite similar to the discussion in Section 3.2.8: if ε is small the difference coefficient bias will be small, but the variance of the estimators $\overline{\delta}_n$ and $\overline{\delta}_n^c$ will be high. An RMS minimization similar to the one in Section 3.2.8 is possible; in the interest of brevity, we leave this as an exercise to the reader (see also Glasserman [2004], Chapter 7).

3.3.2 Pathwise Estimate

3.3.2.1 Black-Scholes Delta

Reverting back to the setting of Section 3.3.1.1, let us take another look at the delta definition (3.68):

$$\begin{aligned}\frac{dV}{dS_0} &= \lim_{h\to 0}\frac{V(S_0+h)-V(S_0)}{h}\\ &= e^{-rT}\lim_{h\to 0}\mathrm{E}\left(\frac{g\left(S_h(T)\right)-g\left(S(T)\right)}{h}\right),\end{aligned} \tag{3.71}$$

where we have used the same notation as earlier:

$$S_h(T) = (S_0+h)\,S(T)/S_0. \tag{3.72}$$

Under sufficient regularity on g, we can interchange expectation and limit in (3.71), such that simply

$$\begin{aligned}\frac{dV}{dS_0} &= e^{-rT}\mathrm{E}\left(\lim_{h\to 0}\frac{g\left(S_h(T)\right)-g\left(S(T)\right)}{h}\right)\\ &= e^{-rT}\mathrm{E}\left(g'\left(S(T)\right)\frac{dS(T)}{dS_0}\right)\\ &= e^{-rT}\mathrm{E}\left(g'\left(S(T)\right)\frac{S(T)}{S_0}\right),\end{aligned} \tag{3.73}$$

[16] More precisely, we need uniform integrability in the difference coefficients $[Y(\alpha+\varepsilon)-Y(\alpha)]\varepsilon^{-1}$ and $\frac{1}{2}[Y(\alpha+\varepsilon)-Y(\alpha-\varepsilon)]\varepsilon^{-1}$. See Section 3.3.2.2 for more precise conditions.

where $g'(x) \triangleq dg/dx$, and the last equality follows by the linearity of (3.72).

We can implement the result (3.73) directly in a Monte Carlo trial, by generating samples of $S(T)$ and recording the sample averages of $g'(S(T))S(T)/S_0$. The resulting estimate for dV/dS_0 is a direct and unbiased estimate of the true derivative; it is known as a *pathwise estimate.*

For (3.73) to hold, g should be continuous, but does not necessarily need to be differentiable everywhere (it suffices that g is Lipschitz continuous, as discussed below). A regular call option payout $g(x) = (x-K)^+$, for instance, is non-differentiable at $x = K$, but we simply write

$$g'(x) = 1_{\{x>K\}}$$

and proceed directly with (3.73). For discontinuous payouts[17], however, care must be taken as a direct application of (3.73) will introduce a bias. To demonstrate, consider the case $g(x) = 1_{\{x>K\}}$. Proceeding informally, a literal application of (3.73) results in

$$\frac{dV}{dS_0} = e^{-rT}\mathrm{E}\left(\delta\left(S(T)-K\right)\frac{S(T)}{S_0}\right),$$

where $\delta(x)$ is the Dirac delta function. While correct, this result is unsuited for Monte Carlo simulation: no matter how many samples n we draw of $S(T)$, the likelihood of $\delta(S(T)-K)$ being non-zero is zero, and the derivative would almost surely be estimated as 0. The correct result, however, is

$$\begin{aligned} e^{-rT}\mathrm{E}\left(\delta\left(S(T)-K\right)\frac{S(T)}{S_0}\right) &= e^{-rT}\mathrm{E}\left(\delta\left(S(T)-K\right)\frac{K}{S_0}\right) \\ &= e^{-rT}\frac{K}{S_0}\mathrm{E}\left(\delta\left(S(T)-K\right)\right) \\ &= e^{-rT}\frac{K}{S_0}\varphi_S(K), \end{aligned}$$

where $\varphi_S(\cdot)$ is the density of $S(T)$.

3.3.2.2 General Case

In a general setting, the technique employed in Section 3.3.2.1 above is known as *infinitesimal perturbation analysis.* A broad overview of the technique can be found in Glasserman [2004], with applications to finance covered in Broadie and Glasserman [1996]. Our treatment follows the latter closely.

Borrowing the notation of Section 3.3.1.2, we again consider estimating $dV/d\alpha$, where $V(\alpha) = \mathrm{E}(Y(\alpha))$. The basic idea of the pathwise derivative estimate is to write

[17] Or for the evaluation of, say, the second derivative (gamma) of a call payout.

$$\frac{dV}{d\alpha} = \frac{d}{d\alpha}\mathrm{E}\left(Y(\alpha)\right) = \mathrm{E}\left(\frac{d}{d\alpha}Y(\alpha)\right). \tag{3.74}$$

The exchange of expectation and differentiation requires certain regularity conditions to be valid. In practice, the most interesting situation arises when Y represents a (discounted) payout function, such that

$$Y(\alpha) = g\left(X(\alpha)\right),$$

where $X(\alpha) = (X_1(\alpha), \ldots, X_q(\alpha))^\top$ is a q-dimensional random vector of observations (possibly at different dates) of asset prices. In this case, we have the following result, from Broadie and Glasserman [1996]:

Proposition 3.3.1. *For all α in some open interval $\mathcal{A}$ assume that $dX_i/d\alpha$ exists almost surely for all $i = 1, \ldots, q$. Suppose that the function g is almost surely differentiable*[18] *and is Lipschitz, such that*

$$|g(x) - g(y)| \leq k|x - y|$$

for some constant k. Finally, assume that there exists finite-mean random variables β_i, $i = 1, \ldots, q$, such that for all $\alpha_1, \alpha_2 \in \mathcal{A}$

$$|X_i(\alpha_2) - X_i(\alpha_1)| \leq \beta_i \left|\alpha_2 - \alpha_1\right|.$$

In this case, (3.74) holds.

The first two assumptions of Proposition ensure that the random variable $dY(\alpha)/d\alpha$ exists almost surely, with its value given by the chain rule

$$\frac{d}{d\alpha}Y(\alpha) = \sum_{i=1}^{q} \frac{\partial g}{\partial X_i}\frac{dX_i}{d\alpha}. \tag{3.75}$$

As we saw earlier, in Section 3.3.2.1, almost sure existence of $dY(\alpha)/d\alpha$ is not sufficient for the pathwise method to yield an unbiased estimator, we also need, roughly speaking, for g to be continuous at the points at which differentiability fails. The last two conditions ensure this, and together imply that Y is almost surely Lipschitz in α:

$$|Y(\alpha_2) - Y(\alpha_1)| \leq \beta_Y \left|\alpha_2 - \alpha_1\right|, \quad \beta_Y = k\sum_{i=1}^{q}\beta_i.$$

As $\alpha^{-1}|Y(h+\alpha) - Y(h)|$ is then bounded by β_Y, where $\mathrm{E}(\beta_Y) < \infty$, the result of Proposition 3.3.1 follows from the dominated convergence theorem. See Broadie and Glasserman [1996] for further details.

[18] That is, differentiable everywhere except on some set $\mathcal{X}$ where $\mathrm{P}(X(\alpha) \in \mathcal{X}) = 0$.

Remark 3.3.2. If in Proposition 3.3.1 we further assume that $\mathrm{E}(\beta_i^2) < \infty$, it follows that $\mathrm{E}(Y(h+\varepsilon) - Y(h)) \leq \beta_Y^2 \varepsilon^2$, such that

$$\mathrm{Var}\,(Y(h+\varepsilon) - Y(h)) = O(\varepsilon^2).$$

We recognize this as case ii) from Section 3.3.1.2 on finite difference estimates, for which we have now made the regularity conditions more precise. In practice, the Lipschitz continuity of g is the critical condition.

Remark 3.3.3. For discontinuous payouts, the pathwise method will yield a biased estimator. As we saw in Section 3.3.2.1, however, for a simple process where the transition density is known, the bias can often be accounted for. We shall see an example of this in Chapter 24. Notice that if the transition density is known, another method for sensitivity simulation — the likelihood ratio method — also applies. See Section 3.3.3 below for details about this method.

3.3.2.3 Sensitivity Path Generation

In Section 3.3.2.1, simulation of $dY(\alpha)/d\alpha$ was straightforward due to the simplicity of the Black-Scholes dynamics. In general, we see from (3.75) that generation of the random variables $dY(\alpha)/d\alpha$ will require us to compute the partial derivatives of the payout with respect to the underlying assets $(\partial g/\partial X_i)$, as well as the sensitivities of the assets with respect to the perturbation parameter α $(dX_i/d\alpha)$. The latter is normally the most difficult, and we shall outline a general approach here.

For illustration, consider a scalar SDE of the usual form

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

where X and W are one-dimensional. Let α be a parameter on which $X(t)$ depends (such as $X(0)$ or some parameter of μ or σ). Let $D_\alpha(t)$ denote $dX(t)/d\alpha$. Formally differentiating the SDE with respect to α, we get

$$dD_\alpha(t) = \mu'\,(t, X(t))\,D_\alpha(t)\,dt + \sigma'\,(t, X(t))\,D_\alpha(t)\,dW(t), \tag{3.76}$$

where $\mu'(t,x) = \partial\mu(t,x)/\partial x$, and similar for σ'. This SDE can be discretized and simulated in parallel with the simulation of the SDE for $X(t)$ itself. In general, the work associated with this will obviously be more substantial than for the Black-Scholes delta, where we saw that $D_\alpha(t)$ could be recovered as the simple fraction $X(t)/X(0)$ (see equation (3.73)).

A few notes on the technique above. First, some regularity is obviously needed for (3.76) to be meaningful; see Kunita [1990] for some relevant results. Second, extensions to multi-dimensional SDEs are straightforward, although the dimension of the total scheme can be large. For instance, if X is p-dimensional and we wish to compute sensitivities with respect to $X_i(0)$,

$i = 1, \ldots, p$, a $p \times p$ system of SDEs for quantities $dX_i(t)/dX_j(0)$ will be required. We shall discuss approximative methods to improve efficiency of such high-dimensional matrix SDEs later, in Chapter 24 (see also Glasserman and Zhao [1999]).

3.3.3 Likelihood Ratio Method

As discussed above, the pathwise derivative method typically applies only to options with sufficiently smooth payouts[19] and can be cumbersome for multi-dimensional SDEs. For processes with explicitly known transition densities, the alternative *likelihood ratio method* can be used. This method applies to discontinuous payout functions and, unlike the pathwise method, requires little knowledge of the payout function and its derivatives, making it convenient for general implementation on a computer. When both methods apply, however, the pathwise derivative method generally is more efficient.

3.3.3.1 Black-Scholes Delta

In the notation of Sections 3.3.1.1 and 3.3.2.1, the Black-Scholes price of a call option can be written

$$\begin{aligned} V(S_0) &= e^{-rT}\mathrm{E}\left((S(T)-K)^+ \middle| S(0) = S_0 \right) \\ &= e^{-rT}\mathrm{E}\left(\left(e^{\ln S_0 + \left(r - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z} - K \right)^+ \right) \\ &= e^{-rT}\mathrm{E}\left(\left(e^{Y(T)} - K \right)^+ \right), \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ and

$$Y(T) \sim \mathcal{N}\left(\ln S_0 + \left(r - \frac{1}{2}\sigma^2 \right) T, \sigma^2 T \right).$$

The density of Y is thereby a function of S_0:

$$\begin{aligned} P(Y \in dy) &= \frac{dy}{\sqrt{2\pi}\sigma\sqrt{T}} \exp\left(-\frac{1}{2} \left(\frac{y - \ln S_0 - \left(r - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} \right)^2 \right) \\ &\triangleq \varphi(y; S_0)\, dy. \end{aligned}$$

Thereby

$$V(S_0) = e^{-rT} \int_{-\infty}^{\infty} \left(e^y - K\right)^+ \varphi(y; S_0)\, dy, \tag{3.77}$$

such that

[19] But see Remark 3.3.3.

$$\begin{aligned}\frac{dV(S_0)}{dS_0} &= e^{-rT}\int_{-\infty}^{\infty}(e^y-K)^+\frac{\partial\varphi(y;S_0)}{\partial S_0}\,dy\\ &= e^{-rT}\int_{-\infty}^{\infty}(e^y-K)^+\frac{\partial\varphi(y;S_0)}{\partial S_0}\frac{\varphi(y;S_0)}{\varphi(y;S_0)}\,dy\\ &= e^{-rT}\int_{-\infty}^{\infty}(e^y-K)^+\frac{\partial\ln\varphi(y;S_0)}{\partial S_0}\varphi(y;S_0)\,dy. \qquad (3.78)\end{aligned}$$

Comparison of (3.78) with (3.77) demonstrates that we can effectively compute the Black-Scholes delta as the price of a security that pays out at time T the amount

$$\left(e^{Y(T)}-K\right)^+ l(Y(T)), \qquad (3.79)$$

where l is the so-called *log-likelihood ratio* (also known as the *score function*)

$$l(Y(T)) = \frac{\partial\ln\varphi(Y(T);S_0)}{\partial S_0} = \frac{Y(T)-\ln S_0-\left(r-\frac{1}{2}\sigma^2\right)T}{S_0\sigma^2 T} = \frac{Z}{S_0\sigma\sqrt{T}}. \qquad (3.80)$$

By differentiating the density, rather than the payout itself, the likelihood ratio technique applies to even discontinuous payouts, requiring only that the density is smooth (which is clearly the case here). Notice in particular that the log-likelihood ratio is independent of the payout, allowing us to use the same function $l(Y(T))$ for all European style payout functions $V(T) = g(S(T))$.

3.3.3.2 General Case

As in Section 3.3.2.2, consider now the general case where a random variable $Y(\alpha)$ represents a (deflated) payout function g applied to a vector of random variables $X(\alpha) = (X_1(\alpha),\ldots,X_q(\alpha))^\top$. Again, α is a parameter with respect to which we wish to compute sensitivities. Let the joint density of $X(\alpha)$ be denoted $f(x;\alpha)$, $x\in\mathbb{R}^q$. We then have

$$V(\alpha) = \int_{\mathbb{R}^q} g(x)f(x;\alpha)\,dx.$$

Making the reasonable assumption that density $f(x;\alpha)$ is a smooth function of α, we interchange integration and differentiation, such that

$$\frac{\partial V(\alpha)}{\partial\alpha} = \int_{\mathbb{R}^q} g(x)\frac{\partial f(x;\alpha)}{\partial\alpha}\,dx = \int_{\mathbb{R}^q} g(x)\frac{\partial\ln f(x;\alpha)}{\partial\alpha}f(x;\alpha)\,dx.$$

As for the Black-Scholes case above, the derivative $\partial V(\alpha)/\partial\alpha$ can thus be computed as the expectation of the payout modified by a log-likelihood ratio:

$$g(x)l(x), \quad l(x) = \frac{\partial\ln f(x;\alpha)}{\partial\alpha}.$$

3.3.3.3 Euler Schemes

In practice, the reliance on explicit knowledge of a transition density can be a considerable obstacle, and may rule out the application of the likelihood ratio method for many complex models. In cases where process dynamics are simulated through a simple time-discretization scheme, the situation is, however, salvageable, as we shall now demonstrate.

For illustration, consider an asset $X(t)$ that follows an SDE of the type

$$dX(t) = \mu\left(t, X(t); \alpha\right) dt + \sigma\left(t, X(t); \alpha\right) dW(t),$$

where μ and σ are smooth functions, and where α is a parameter. In general, we do not know the exact transition density for $X(t)$. However, suppose now that we use an Euler scheme to simulate on some grid $\{t_i\}_{i=1}^m$, i.e.

$$\widehat{X}(t_{i+1}) = \widehat{X}(t_i) + \mu\left(t_i, \widehat{X}(t_i); \alpha\right)(t_{i+1} - t_i) + \sigma\left(t_i, \widehat{X}(t_i); \alpha\right)\sqrt{t_{i+1} - t_i}\, Z_i,$$

where $t_0 = 0$ and $Z_0, Z_1, \ldots, Z_{m-1}$ is a sequence of i.i.d. standard Gaussian random variables. Clearly, the transition density for the $\widehat{X}(t_{i+1})$ is now Gaussian,

$$\widehat{X}(t_{i+1}) | \widehat{X}(t_i) \sim \mathcal{N}\left(m_i\left(\widehat{X}(t_i); \alpha\right), s_i\left(\widehat{X}(t_i); \alpha\right)\right),$$

where, for $i = 0, \ldots, m-1$,

$$m_i\left(\widehat{X}(t_i); \alpha\right) = \widehat{X}(t_i) + \mu\left(t_i, \widehat{X}(t_i); \alpha\right)(t_{i+1} - t_i),$$
$$s_i\left(\widehat{X}(t_i); \alpha\right) = \sigma\left(t_i, \widehat{X}(t_i); \alpha\right)^2 (t_{i+1} - t_i).$$

Set $\widehat{X} = (\widehat{X}(t_1), \ldots, \widehat{X}(t_m))^\top$; the density of this vector is (where $x_0 = X(0)$)

$$f(x_1, \ldots x_m; \alpha) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sqrt{s_{i-1}\left(x_{i-1}; \alpha\right)}} \exp\left(-\frac{\left(x_i - m_{i-1}\left(x_{i-1}; \alpha\right)\right)^2}{2 s_{i-1}\left(x_{i-1}; \alpha\right)}\right). \tag{3.81}$$

Consider some (potentially path-dependent) security V with payout function $g(\widehat{X})$ and time 0 price of $V(\alpha)$. Equipped with (3.81), we can estimate the parameter sensitivity by the likelihood ratio method as

$$\frac{\partial V(\alpha)}{\partial \alpha} = \mathrm{E}\left(g(\widehat{X}) \frac{d}{d\alpha} \ln\left(f(\widehat{X}; \alpha)\right)\right),$$

where

$$\frac{d}{d\alpha} \ln\left(f(\widehat{X}; \alpha)\right)$$
$$= \frac{d}{d\alpha} \sum_{i=1}^m \left(-\frac{1}{2} \ln\left(s_{i-1}\left(x_{i-1}; \alpha\right)\right) - \frac{\left(x_i - m_{i-1}\left(x_{i-1}; \alpha\right)\right)^2}{2 s_{i-1}\left(x_{i-1}; \alpha\right)}\right). \tag{3.82}$$

This derivative can typically be computed in closed form; if not, one can estimate it by finite differences (as in Su and Randall [2008]). Notice that when $\alpha = X(0)$ (i.e. we are trying to estimate the delta), only s_0 and m_0 will depend on α, simplifying computations.

The idea used above extends easily to vector-valued $X(t)$. In principle, higher-order schemes (e.g. the Milstein scheme) can also be used, although the complexity increases considerably.

3.3.3.4 Some Remarks

The main advantage of the likelihood ratio is the fact that it applies to classes of payouts for which other methods (pathwise methods, finite difference method) do not work well. Moreover, the method is easy and efficient to implement, as the log-likelihood ratio l is independent of the payout and does not — unlike the general pathwise method — require simulation of any quantities other than the vector X itself. As discussed above, the primary drawback of the method is its reliance on explicit knowledge of the process density, ruling out many of the more advanced models (although, as Section 3.3.3.3 demonstrated, there may sometimes be ways around this). Further, the variance of the likelihood ratio method can often be quite big, particularly if the parameter α simultaneously affects multiple stochastic variables. A fuller discussion of this issue, as well as the related issue of absolute continuity, can be found in Glasserman [2004]. We note in passing that the likelihood ratio method is a special case of a body of methods that have emerged from the so-called *Malliavin calculus*; see Fournie et al. [1999] for a survey. Most Malliavin methods other than the basic likelihood ratio method are, however, not particularly attractive due to computational issues[20].

We round off this section by noting that the various methods for derivative estimates can often be successfully combined. For instance, while it is common to use either the pathwise method or the likelihood ratio method to compute first order sensitivities (such as delta), second-order sensitivities (such as gamma) are often done by the finite difference method applied to first-order sensitivities, often using fairly sizable shifts of the underlying variables. By combining the pathwise method with the likelihood ratio method, we can also address the fact that the first derivative of many kinked option payouts is discontinuous, allowing us to produce a bias-free estimate of the second derivative. See Fournie et al. [1999] for other examples of combining the pathwise method with the likelihood ratio method.

[20] Besides, the Malliavin calculus itself, a very technical area of mathematics even for specialists, can be avoided altogether, as Chen and Glasserman [2007b] demonstrate.

3.4 Variance Reduction Techniques

As discussed earlier, the convergence of the Monte Carlo method is quite slow, of order $O(n^{-1/2})$ where n is the number of Monte Carlo samples. While there is little that can be done to improve[21] the $n^{-1/2}$ order itself, the constant multiplying $n^{-1/2}$ can be affected by a careful choice of the Monte Carlo estimator. Methods to improve numerical efficiency this way are known collectively as *variance reduction techniques*, and constitute a major area of research in the theory of Monte Carlo methods. Our introduction of the topic is limited to a few basic examples. More details are provided later in the book for concrete models and products (see e.g. Chapter 25), and more information can be found in the standard Monte Carlo literature, including Hammersley and Handscomb [1965] and the survey article Boyle et al. [1997].

3.4.1 Variance Reduction and Efficiency

We recall that the goal of the Monte Carlo method is to estimate some quantity μ (e.g., the price of a financial contract) as the sample mean of n i.i.d. random variables $Y_1, \ldots, Y_n$, where each Y_i has expectation $\mathrm{E}(Y_i) = \mu$ and variance $\mathrm{Var}(Y_i) = \sigma^2$. From Section 3.1, we know that for large n the standard error of the sample mean $n^{-1} \sum Y_i$ is $\sigma n^{-1/2}$, with the probabilistic error bounds on the estimate of μ being proportional to the standard error.

Suppose now that we have available two sets of i.i.d. sequences $Y_{1,i}$ and $Y_{2,i}$, $i = 1, \ldots, n$, where $\mathrm{E}(Y_{1,i}) = \mathrm{E}(Y_{2,i}) = \mu$, but $\mathrm{Var}(Y_{1,i}) = \sigma_1^2$ and $\mathrm{Var}(Y_{2,i}) = \sigma_2^2$, with $\sigma_1 \neq \sigma_2$. Also suppose that the time it takes on a computer to generate individual samples $Y_{1,i}$ and $Y_{2,i}$ is τ_1 and τ_2, respectively. Which of the two estimators $n^{-1} \sum Y_{1,i}$ and $n^{-1} \sum Y_{2,i}$ is preferable? To answer this question, assume that we have a large fixed computing time budget τ. The number of replications of $Y_{1,i}$ and $Y_{2,i}$ that can be executed are thus (the integer parts of) τ/τ_1 and τ/τ_2, respectively. To this correspond sample mean standard errors of

$$\frac{\sigma_1}{\sqrt{\tau/\tau_1}} \text{ and } \frac{\sigma_2}{\sqrt{\tau/\tau_2}},$$

respectively. It follows that, for large τ, the estimator based on the sequence $Y_{1,i}$, $i = 1, \ldots, n$, is preferable, if

$$\frac{\sigma_1}{\sqrt{\tau/\tau_1}} < \frac{\sigma_2}{\sqrt{\tau/\tau_2}},$$

or equivalently

[21] As we discussed earlier, the quasi-random Monte Carlo method can theoretically achieve better convergence order than $O(n^{-1/2})$.

$$\sigma_1^2 \tau_1 < \sigma_2^2 \tau_2.$$

For obvious reasons, the product of variance and per-sample computing time is known as the *efficiency* of a Monte Carlo estimator. In devising methods to improve Monte Carlo performance, efficiency should always constitute the measure of comparison. For instance, a high-variance estimator may, in fact, be preferable to a low-variance estimator, provided that the former takes less time to compute than the latter.

Duffie and Glynn [1995] discuss Monte Carlo efficiency in more depth, with additional analysis of the effects of bias (see also Section 3.2.8) and cases where τ_1 and τ_2 are random.

3.4.2 Antithetic Variates

3.4.2.1 The Gaussian Case

A simple and easily implemented variance reduction technique is the method of *antithetic variates*. Assume that we are interested in estimating the expected value of a random variable $Y = G(Z)$, where G is a real-valued function and Z is a q-dimensional vector of independent standard Gaussian random variables. This problem routinely arises in determining the expected value of a function of assets driven by a vector SDE; Z then represents the aggregation of all independent standard Gaussian variables used to produce Brownian motion increments; see for instance Section 3.2.3. For n independent realizations of Z, $Z_1, \ldots, Z_n$, rather than using the regular sample average estimator for $\mathrm{E}(Y)$, consider instead using

$$\overline{Y}_n^a = n^{-1} \sum_{i=1}^{n} \frac{G(Z_i) + G(-Z_i)}{2}.$$

In other words, in addition to the set $Z_1, \ldots, Z_n$ of Gaussian samples, we also effectively include the set $-Z_1, \ldots, -Z_n$ in the Monte Carlo trial. As $-Z_1, \ldots, -Z_n$ itself is a sequence of n independent Gaussian samples, we still must have

$$\mathrm{E}\left(\overline{Y}_n^a\right) = \mathrm{E}(Y),$$

so the antithetic estimator is unbiased. Also, as $G(Z)$ and $G(-Z)$ have identical variance,

$$\mathrm{Var}\left(\overline{Y}_n^a\right) = n^{-1}\left[\frac{1}{2}\mathrm{Var}(Y) + \frac{1}{2}\mathrm{Cov}\left(G(Z), G(-Z)\right)\right] = \frac{\mathrm{Var}(Y)}{n}\frac{(1+\rho)}{2},$$

where ρ is the correlation between $G(Z)$ and $G(-Z)$. Recalling that the regular sample average has variance

$$\mathrm{Var}\left(\overline{Y}_n\right) = \frac{\mathrm{Var}(Y)}{n},$$

we conclude that $\mathrm{Var}(\overline{Y}_n^a) < \mathrm{Var}(\overline{Y}_n)$ as long as $\rho < 1$ (which is obviously likely).

While use of antithetic variates can always be expected to lower the standard error, it is not necessarily more efficient than regular Monte Carlo, in the sense defined in Section 3.4.1 . For instance, if generation of the Z_i's is of negligible cost relative to the evaluation of $G(Z)$, computation of $\overline{Y}_n^a$ will take about twice as long as the regular sample average $\overline{Y}_n$. For this case, the results in Section 3.4.1 show that for antithetic variates to constitute an improvement in computational efficiency, we must require that

$$\mathrm{Var}\left(\overline{Y}_n^a\right) \leq \frac{1}{2}\mathrm{Var}\left(\overline{Y}_n\right),$$

or

$$\rho \leq 0.$$

A sufficient condition for $\rho < 0$ is that G be monotone in all q elements of Z. Given this, we would expect antithetic variates to be most suitable for option payouts that depend monotonically on prices.

3.4.2.2 General Case

While the method of antithetic variates is primarily associated with the idea of changing signs on Gaussian variables, the method can, in fact, be extended to other distributions. At the most basic level, most simulation trials involve a series of uniform draws that are translated to other random variables, using techniques described in Section 3.1.1. In this case, we can focus our attention on estimating the mean of a random variable $Y = H(U)$, where H is a function and U is a vector of independent uniformly distributed random variables. We notice that if $U = (U_1, \ldots, U_q)^\top$ is a vector of independent uniform random variables on $[0,1]$, then so is $\widetilde{U} = (1 - U_1, \ldots, 1 - U_q)^\top$. The pair $\{U, \widetilde{U}\}$ is thereby antithetic (negatively dependent) in the same way as the Gaussian pair $\{Z, -Z\}$ above, and we can estimate the mean of Y as the average of independent samples of the form

$$\frac{H(U) + H(\widetilde{U})}{2}.$$

From the discussion above, it follows that if H is monotonic in U, the resulting scheme will exhibit better computational efficiency than regular Monte Carlo.

As an aside, we note that the simple "reflection" of a vector of uniforms advocated above is, as should be obvious, not the only possible way of generating an antithetic sample — for instance, we could have chosen to reflect only select dimensions of the U-vector. A similar observation holds for the case of vector-valued Gaussian variables. The general idea of applying deterministic transformations to a vector-valued sample of random numbers as a way to reduce variance is sometimes known as *systematic sampling*.

3.4.3 Control Variates

3.4.3.1 Basic Idea

While we may need to use Monte Carlo simulation to estimate the unknown mean of a random variable Y, there may be random variables "close" to Y with means that can be computed analytically. It seems reasonable that the additional information about Y revealed by these random variables could be useful in improving our estimate of $\mathrm{E}(Y)$. While a number of strategies are possible[22], we shall here focus on the so-called *control variate* method.

Formally, let

$$Y^c = \left(Y_1^c, \ldots, Y_q^c\right)^\top$$

be a vector of *control variates* (or just *controls*), ideally with strong negative or positive correlation to a variable Y. The mean of Y^c is known to be

$$\mathrm{E}(Y^c) = \mu^c = \left(\mu_1^c, \ldots, \mu_q^c\right)^\top.$$

Now, introduce an arbitrary constant vector

$$\beta = (\beta_1, \ldots, \beta_q)^\top$$

and consider forming the linear combination

$$X = Y - \beta^\top \left(Y^c - \mu^c\right). \tag{3.83}$$

Clearly

$$\mathrm{E}(X) = \mathrm{E}(Y) - \beta^\top \left(\mathrm{E}(Y^c) - \mu^c\right) = \mathrm{E}(Y),$$

so using Monte Carlo sampling to estimate the mean of X will provide an unbiased estimate of $\mathrm{E}(Y)$, regardless of the choice of β.

To analyze the variance of the new variable X, let Σ_{Y^c} be the $q \times q$ covariance matrix of the vector Y^c, and let Σ_{Y,Y^c} be the q-dimensional vector of covariances between Y and the components of Y^c. The variance of X can then be shown to be

$$\mathrm{Var}(X) = \mathrm{Var}(Y) - 2\beta^\top \Sigma_{Y,Y^c} + \beta^\top \Sigma_{Y^c} \beta. \tag{3.84}$$

Whether or not this constitutes an improvement (in the sense that $\mathrm{Var}(X) < \mathrm{Var}(Y)$) is largely a matter of what β is chosen to be. We have the following easily proven lemma.

Lemma 3.4.1. *The function* $\mathrm{Var}(X) = \mathrm{Var}(Y) - 2\beta^\top \Sigma_{Y,Y^c} + \beta^\top \Sigma_{Y^c} \beta$ *is minimized at*

$$\beta^* = \Sigma_{Y^c}^{-1} \Sigma_{Y,Y^c}$$

[22] Other methods include *moment matching* and *importance sampling*. We shall cover the latter strategy shortly; the former is discussed in Boyle et al. [1997], where it is concluded that control variates are superior, at least asymptotically.

with minimum value

$$\min_{\beta} \mathrm{Var}(X) = (1 - R^2)\mathrm{Var}(Y), \quad R^2 \triangleq \frac{\Sigma_{Y,Y^c}^{\top} \Sigma_{Y^c}^{-1} \Sigma_{Y,Y^c}}{\mathrm{Var}(Y)} \geq 0. \tag{3.85}$$

In the lemma, we recognize the scalar R^2 as the R-squared of a multi-dimensional regression of Y against Y^c. Similarly, the components of the optimal vector β^* are the regression coefficients (the slopes) on the vector Y^c. In practice, we may not know Σ_{Y^c} and Σ_{Y,Y^c} explicitly, in which case we simply replace these with empirical estimates, as obtained by an n-sample Monte Carlo trial. We note that if the random samples used to estimate β^* are the same as those used to estimate $\mathrm{E}(X)$, a small bias is typically introduced. This can be circumvented by using separate random numbers for the estimates of β^* and $\mathrm{E}(X)$, but in practice this is rarely worth the effort. Nelson [1990], among others, analyzes this issue in more detail.

While the usage of control variates will always lower variance (unless Y and Y^c are perfectly uncorrelated), an improvement of computational efficiency over standard Monte Carlo is, of course, not guaranteed. Consider, for instance, the case where the computational effort involved in generating a single sample of X is $q + 1$ times that of generating Y itself. This will be the case, if i) the effort of drawing random numbers is small relative to computing Y itself; and ii) each of the components of Y^q take about the same time to compute as Y. According to the result in Section 3.4.1, for this special case the control variate method will only entail an increase in efficiency, if

$$(1 - R^2)\mathrm{Var}(Y)(q + 1) < \mathrm{Var}(Y)$$

or

$$1 - R^2 < \frac{1}{q + 1}.$$

As q grows large, this requirement obviously becomes increasingly difficult to satisfy. Rather than indiscriminately adding multiple controls, it is therefore normally best to properly analyze a given problem and use only a few well-chosen variables with strong (negative or positive) correlation to the variable in question.

3.4.3.2 Non-Linear Controls

Our discussion of the control variate method has so far only considered linear controls (3.83), where the modified estimator involves a linear combination of control variates. The resulting estimate of $\mathrm{E}(Y)$ are n-point sample averages of the type

$$\overline{Y}_n - \beta^{\top}\left(\overline{Y}_n^c - \mu^c\right).$$

A more general formulation than (3.83) approximates $\mathrm{E}(Y)$ with

$$f\left(\overline{Y}_n^c, \overline{Y}_n\right) \tag{3.86}$$

for some function f satisfying

$$f\left(\mu^c, y\right) = y. \tag{3.87}$$

The requirement (3.87) ensures that $f(\overline{Y}_n^c, \overline{Y}_n)$ approaches $\mathrm{E}(Y)$ in the large-sample limit; unlike the regular control variate formulation, however, (3.86) may involve a bias for finite sample sizes.

If f is smooth, a result by Glynn and Whitt [1989] demonstrates that for sufficiently large samples, any non-linear control variate estimator of the type (3.86) is equivalent to an ordinary linear control variate estimator. Still, there may be situations where a non-linear control variate estimator is appropriate, either because i) the sample size is not large enough to justify the result in Glynn and Whitt [1989]; or ii) because the "effective" β weighting of $\overline{Y}_n^c$ implied by f is close to optimal, allowing us to skip the estimation of β^*.

To give an example of non-linear control variates, let us consider the "delta" method of Clewlow and Carverhill [1994]. To state the basic idea, consider the estimate of

$$V(0) = \mathrm{E}\left(g\left(X(T)\right)\right),$$

where $X(t)$ is a p-dimensional vector process and $g : \mathbb{R}^p \to \mathbb{R}$ is a smooth function. Assume that all components of X are martingales, as is the case when X represents assets deflated by a numeraire. We recall from Section 1.7 that, under certain regularity conditions, we have

$$V(T) = V(0) + \int_0^T \sum_{i=1}^p V_{x_i}(t)\, dX_i(t),$$

where we use the notation $V_{x_i}(t)$ from Section 1.7 to denote, informally, $V_{x_i}(t) = \partial V(t)/\partial X_i(t)$. On a simulation time line $\{t_j\}_{j=1}^m$, we can write, in the style of an Euler scheme,

$$V(T) \approx V(0) + \sum_{j=1}^m \sum_{i=1}^p V_{x_i}(t_{j-1})\left(X_i(t_j) - X_i(t_{j-1})\right).$$

As the zero-mean quantity

$$\sum_{j=1}^m \sum_{i=1}^p V_{x_i}(t_{j-1})\left(X_i(t_j) - X_i(t_{j-1})\right)$$

is likely to have high correlation to $V(T)$, we can consider using it as a control variate. One obstacle is the fact that the derivatives $V_{x_i}(t)$ are likely

to be unknown (as the function $V(t)$ is unknown). Often, however, we can provide an inspired guess for these derivatives, based on perhaps a simpler model or on regression information. The former idea is outlined in Clewlow and Carverhill [1994], and the latter shall be discussed further in Chapter 25. In any case, the resulting scheme ends up effectively using the increments $X_i(t_j) - X_i(t_{j-1})$ as controls, with non-constant weights $V_{x_i}(t_{j-1})$ being functions of the X_i themselves.

3.4.4 Importance Sampling

3.4.4.1 Basic Idea

The basic idea of the *importance sampling method* is to use a measure shift to reduce variance. For a given measure P, consider estimating

$$\mu = \mathrm{E}^{\mathrm{P}}(Y), \tag{3.88}$$

where Y is a scalar random variable. Let $\widehat{\mathrm{P}}$ be a measure equivalent to P. From the Radon-Nikodym theorem in Chapter 1, we have

$$\mu = \mathrm{E}^{\widehat{\mathrm{P}}}\left(Y/R\right), \tag{3.89}$$

where R is the Radon-Nikodym derivative

$$R = d\widehat{\mathrm{P}}/d\mathrm{P}, \quad \mathrm{E}^{\mathrm{P}}(R) = 1.$$

While (3.88) and (3.89) are both valid expressions for μ, it is possible that the variance of Y/R under measure $\widehat{\mathrm{P}}$ is lower than the variance of Y under P, making (3.89) potentially more efficient for Monte Carlo purposes. As an extreme case, consider setting (assuming $Y > 0$ a.s.)

$$R = \frac{Y}{\mathrm{E}^{\mathrm{P}}\left(Y\right)} = \frac{Y}{\mu}. \tag{3.90}$$

In this case

$$Y/R = \mu$$

and non-random, implying that the measure shift from P to $\widehat{\mathrm{P}}$ has removed *all* variance. The problem with the "perfect" choice (3.90) is obviously that we do not know μ — if we did, there would be no need to estimate it by Monte Carlo methods. Nevertheless, we may be able to provide a good guess for μ, allowing us to use (3.90) in an approximate sense.

3.4.4.2 Density Formulation

Importance sampling methods are often most conveniently (and most intuitively) treated in terms of probability densities, so let us cast the description

of Section 3.4.4.1 in such terms. Specifically, let us assume that Y can be represented as $g(X)$, where $g : \mathbb{R}^p \to \mathbb{R}$ is a well-behaved function and X is p-dimensional with probability density $f : \mathbb{R}^p \to \mathbb{R}$. We then write

$$\mu = \mathrm{E}^{\mathrm{P}}\left(g\left(X\right)\right) = \int_{\mathbb{R}^p} g(x) f(x)\, dx,$$

to which corresponds a regular Monte Carlo estimator

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i),$$

where the X_i are independent samples of X, drawn from the density f. Let $h : \mathbb{R}^p \to \mathbb{R}$ be another density, satisfying the continuity requirement that $h(x) > 0$ whenever $f(x) > 0$. We can then also represent μ as

$$\mu = \int_{\mathbb{R}^p} g(x)\frac{f(x)}{h(x)} h(x)\, dx,$$

which we can interpret as

$$\mu = \mathrm{E}^{\widehat{\mathrm{P}}}\left(g\left(X\right)\frac{f\left(X\right)}{h\left(X\right)}\right),$$

where $\widehat{\mathrm{P}}$ is a measure under which X has density $h(x)$. Comparison to the results above identifies the so-called *likelihood ratio* $l(x) = f(x)/h(x)$ as the Radon-Nikodym derivative $d\mathrm{P}/d\widehat{\mathrm{P}}$ (or $1/R$) governing the shift from P to $\widehat{\mathrm{P}}$. If now $X_1, \ldots, X_n$ are independent draws from h (and *not* f), the importance sampling Monte Carlo estimator for μ takes the form

$$\overline{\mu}_n^h = \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{f(X_i)}{h(X_i)}.$$

Let us investigate under which circumstances importance sampling will lead to an improvement in variance. We have

$$\begin{aligned}\mathrm{Var}\left(\overline{\mu}_n^h\right) &= \frac{1}{n}\left[\mathrm{E}^{\widehat{\mathrm{P}}}\left(g\left(X\right)^2\frac{f\left(X\right)^2}{h\left(X\right)^2}\right) - \mu^2\right] \\ &= \frac{1}{n}\left[\mathrm{E}^{\mathrm{P}}\left(g\left(X\right)^2\frac{f\left(X\right)}{h\left(X\right)}\right) - \mu^2\right],\end{aligned}$$

and

$$\mathrm{Var}\left(\overline{\mu}_n\right) = \frac{1}{n}\left[\mathrm{E}^{\mathrm{P}}\left(g\left(X\right)^2\right) - \mu^2\right].$$

Hence, importance sampling will lower variance, provided that

$$\mathrm{E}^{\mathrm{P}}\left(g(X)^2 \frac{f(X)}{h(X)}\right) < \mathrm{E}^{\mathrm{P}}\left(g(X)^2\right).$$

Choosing the importance sampling density $h(x)$ wisely is key to the efficiency of the importance sampling. As an extreme, suppose we could set

$$h(x) = Cf(x)g(x), \tag{3.91}$$

where the constant C is dictated by the need for $h(x)$ to integrate to 1:

$$C^{-1} = \int_{\mathbb{R}^p} g(x)f(x)\,dx = \mu.$$

In this case,

$$\mathrm{E}^{\mathrm{P}}\left(g(X)^2 \frac{f(X)}{h(X)}\right) = C^{-1}\mathrm{E}^{\mathrm{P}}\left(g(X)\right) = \mu^2$$

and

$$\mathrm{Var}\left(\overline{\mu}_n^h\right) = 0.$$

This replicates a similar argument in Section 3.4.4.1 (see equation (3.90)), and is equally useless in practice: to compute (3.91) we need to normalize by the constant $1/\mu$, where μ is the quantity that we are trying to estimate in the first place. Nevertheless, (3.91) provides some useful practical guidance: a good choice of likelihood density will sample in proportion to f and g. That is, values of X where both the density $f(X)$ and the payout $g(X)$ are high should be assigned a high value of $h(X)$ (high "importance"), and values of X where either $f(X)$ or $g(X)$ (or both) are low should be assigned a low value of $h(X)$ (low "importance"). This rule is often particularly easy and efficient to apply to situations where $g(X)$ is significant only for a set $X \in \mathcal{A}$, where $P(X \in \mathcal{A})$ is small. Such rare-event problems are a classical application of importance sampling; we give a simple example in Section 3.4.4.5. Related applications to barrier options can be found later, in Chapter 25, with more such examples in Boyle et al. [1997].

3.4.4.3 Importance Sampling and SDEs

Consider now a dynamic setting where we are given a P-measure SDE

$$dX(t) = \mu(t, X(t))\,dt + \sigma(t, X(t))\,dW(t), \tag{3.92}$$

where X is p-dimensional and W is d-dimensional. We wish to evaluate

$$\mathrm{E}^{\mathrm{P}}\left(g(X(T))\right)$$

for a real-valued function g. To shift measure, we introduce the density process

$$d\varsigma(t) = -\varsigma(t)\theta(t)^\top dW(t), \quad \varsigma(0) = 1, \tag{3.93}$$

for some adapted d-dimensional process $\theta(t)$, sufficiently regular to make $\varsigma(\cdot)$ a martingale (see Chapter 1). Let

$$\varsigma(t) = \mathrm{E}_t^{\mathrm{P}}\left(\frac{d\widehat{\mathrm{P}}}{d\mathrm{P}}\right),$$

for a new measure $\widehat{\mathrm{P}}$. By the Girsanov theorem, under $\widehat{\mathrm{P}}$,

$$dX(t) = [\mu(t, X(t)) - \sigma(t, X(t))\,\theta(t)]\,dt + \sigma(t, X(t))\,d\widehat{W}(t), \tag{3.94}$$

where $\widehat{W}$ is a Brownian motion in $\widehat{\mathrm{P}}$. Also, by the Radon-Nikodym theorem,

$$\mathrm{E}^{\mathrm{P}}(g(X(T))) = \mathrm{E}^{\widehat{\mathrm{P}}}\left(\frac{g(X(T))}{\varsigma(T)}\right). \tag{3.95}$$

In a Monte Carlo setting, rather than simulating (3.92) (using methods from Section 3.2) and computing the sample mean of $g(X(T))$, we can instead jointly simulate (3.93) and (3.94) and compute the sample mean of $g(X(T))/\varsigma(T)$. The validity of this approach is independent of the choice of θ in (3.93), and we can use θ as a parameter to minimize the variance of $g(X(T))/\varsigma(T)$ under $\widehat{\mathrm{P}}$.

To find the optimal choice for θ, define

$$u(t, X(t)) = \mathrm{E}_t^{\mathrm{P}}(g(X(T))), \quad t \le T,$$

and consider setting

$$\varsigma(t)u(0, X(0)) = u(t, X(t)).$$

By Ito's lemma,

$$d\varsigma(t) = -\varsigma(t)\theta(t)^\top dW(t), \quad \varsigma(0) = 1,$$

where

$$\theta(t) = -u(t, X(t))^{-1}\,\sigma(t, X(t))^\top \frac{\partial u(t, X(t))}{\partial x}, \tag{3.96}$$

with $\partial u(t, X(t))/\partial x$ being a p-dimensional vector of partial derivatives $\{\partial u(t, X(t))/\partial x_i\}$. The choice for θ in (3.96) is optimal as we have

$$g(X(T))/\varsigma(T) = u(0, X(0)) = \mathrm{E}^{\mathrm{P}}(g(X(T))),$$

which is non-random with zero variance. As in earlier examples, the optimal choice for $\theta(t)$ cannot be applied directly as it requires knowledge of $\mathrm{E}_t^{\mathrm{P}}(g(X(T)))$ for all t, knowledge which we never possess in practice. In many applications, however, we can often make an educated guess for u, based perhaps on either a simpler SDE than (3.92) or on a simpler payout function than g. We shall see an example of this in Chapter 25; another application can be found in Schoenmakers and Heemink [1997].

3.4.4.4 More on SDE Path Simulation

Let us consider an alternative point of view about SDE simulations, where we assume that the SDE (3.92) is simulated by an m–dimensional Euler scheme (or similar), such that we can write (see also Section 3.4.2.1) for some function $G : \mathbb{R}^{p\times m} \to \mathbb{R}$,

$$g\left(X(T)\right) = G\left(Z_1, \ldots, Z_m\right),$$

where the Z_i are independent p-dimensional Gaussian vectors. With the Gaussian density of Z_i being denoted $\phi(z)$, $z \in \mathbb{R}^p$, the independence of the Z_i's allows us to write

$$\mathrm{E}^{\mathrm{P}}\left(g\left(X(T)\right)\right) = \int_{\mathbb{R}^{p\times m}} G\left(z_1, \ldots, z_m\right) \prod_{i=1}^{m} \phi(z_i)\, dz, \quad z \triangleq \left(z_1, \ldots, z_m\right).$$

If we apply a change of measure that preserves independence of Z_i but alters the common marginal density from $\phi(z)$ to $h(z)$, the likelihood ratio is easily seen to be

$$l(z) = \prod_{i=1}^{m} \frac{\phi(z_i)}{h(z_i)},$$

such that

$$\mathrm{E}^{\mathrm{P}}\left(g\left(X(T)\right)\right) = \mathrm{E}^{\widehat{\mathrm{P}}}\left(G\left(Z_1, \ldots, Z_m\right) \prod_{i=1}^{m} \frac{\phi(Z_i)}{h(Z_i)}\right).$$

It is understood that the Z_i used to advance the SDE simulation under $\widehat{\mathrm{P}}$ are drawn from the density h, rather than ϕ.

To give a concrete example of a measure shift, assume for simplicity that $p = 1$ and consider shifting the means of the Z_i from zero to some scalar[23] μ, but retaining unit variance. For this, we must set

$$h(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_i - \mu)^2\right),$$

whereby

$$l\left(z\right) = l(z;\mu) = \exp\left(-\mu \sum_{i=1}^{m} z_i + \frac{m}{2}\mu^2\right). \tag{3.97}$$

Here μ is a free variable, which can be set to minimize the variance of the term

$$G(Z)l(Z;\mu), \quad Z \triangleq \left(Z_1, \ldots, Z_m\right)$$

[23]It is also straightforward to introduce a measure shift that moves the means of the Z_i to *different* means μ_i, $i = 1, \ldots, m$.

under $\widehat{\mathrm{P}}$. Sometimes this minimization problem can be handled analytically (see Section 3.4.4.5), but most often numerical methods are required. Examples of how to perform this minimization by Monte Carlo simulation can be found in, for example, Su and Fu [2002] and Capriotti [2007]. The approach in Capriotti [2007] (called *least-squares importance sampling*) is particularly straightforward, as the optimization problem is here cast as a least-squares regression problem for which well-known numerical schemes exist such as, e.g., the Levenberg-Marquardt routine in Press et al. [1992]. Both Su and Fu [2002] and Capriotti [2007] point out that, when computing variance, it is advantageous to cast the problem back into the original probability measure P by using

$$\mathrm{E}^{\widehat{\mathrm{P}}}\left(G(Z)^2 l(Z;\mu)^2\right) = \mathrm{E}^{\mathrm{P}}\left(G(Z)^2 l(Z;\mu)\right).$$

Let us finally note that the measure transformation employed above is a special case of so-called *exponential twisting* (also known as *Esscher transform*), under which a density $f(x)$, $x \in \mathbb{R}$, is transformed into

$$f_\theta(x) = e^{\theta x - \gamma(\theta)} f(x),$$

where θ is a twisting parameter and γ is the *cumulant-generating function*

$$\gamma(\theta) = \ln\left(\int_{\mathbb{R}} e^{\theta x}\, dx\right).$$

For a standard Gaussian variable, $\gamma(\theta) = \theta^2/2$, demonstrating that the shift of mean employed above is indeed a special case of exponential twisting. We notice that exponential twisting is often a very convenient starting point when working with parametric families of Radon-Nikodym derivatives.

3.4.4.5 Rare Event Simulation and Linearization

For illustrative purposes, consider finally the problem of estimating by Monte Carlo

$$\mathrm{P}\left(Z > c\right), \tag{3.98}$$

where $Z \sim \mathcal{N}(0,1)$ is standard Gaussian under the measure P, and c is a big number. In ordinary Monte Carlo, we write

$$\mathrm{P}\left(Z > c\right) = \mathrm{E}^{\mathrm{P}}\left(1_{\{Z>c\}}\right)$$

and use the sample mean estimator

$$\mathrm{P}\left(Z > c\right) \approx \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i>c\}},$$

where $Z_1, \ldots, Z_n$ are independent standard Gaussian samples. We notice that

$$\begin{aligned}\mathrm{Var}^{\mathrm{P}}\left(1_{\{Z>c\}}\right) &= \mathrm{E}^{\mathrm{P}}\left(\left(1_{\{Z>c\}}\right)^2\right) - \mathrm{E}^{\mathrm{P}}\left(1_{\{Z>c\}}\right)^2 \\ &= \mathrm{E}^{\mathrm{P}}\left(1_{\{Z>c\}}\right) - \mathrm{E}^{\mathrm{P}}\left(1_{\{Z>c\}}\right)^2 \\ &= \mathrm{P}(Z>c)\left(1-\mathrm{P}(Z>c)\right),\end{aligned}$$

with sample mean estimator variance being n times smaller. Consider now introducing a probability measure that shifts the mean of Z from 0 to μ. The likelihood ratio is seen from (3.97) to be

$$l(z) = e^{-\mu z+\mu^2/2},$$

such that

$$\mathrm{P}\left(Z>c\right) = \mathrm{E}^{\widehat{\mathrm{P}}}\left(e^{-\mu Z+\mu^2/2}\,1_{\{Z>c\}}\right),$$

where $Z \sim \mathcal{N}(\mu,1)$ in the measure $\widehat{\mathrm{P}}$. A Monte Carlo estimator for this is then

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\mu(Z_i+\mu)+\mu^2/2}\,1_{\{Z_i+\mu>c\}},$$

where $Z_1,\ldots,Z_n$ are again independent standard Gaussian samples. Notice that we have added to the Z_i the mean μ to reflect the shift of measure from P to $\widehat{\mathrm{P}}$. As for variance, we have

$$\begin{aligned}\mathrm{Var}^{\widehat{\mathrm{P}}}\left(e^{-\mu Z+\mu^2/2}\,1_{\{Z>c\}}\right) &= \mathrm{E}^{\widehat{\mathrm{P}}}\left(e^{-2\mu Z+\mu^2}\left(1_{\{Z>c\}}\right)^2\right) - \mathrm{P}\left(Z>c\right)^2 \\ &= \mathrm{E}^{\widehat{\mathrm{P}}}\left(e^{-2\mu Z+\mu^2}\,1_{\{Z>c\}}\right) - \mathrm{P}\left(Z>c\right)^2 \\ &= \mathrm{E}^{\mathrm{P}}\left(e^{-\mu Z+\mu^2/2}\,1_{\{Z>c\}}\right) - \mathrm{P}\left(Z>c\right)^2 \\ &= e^{\mu^2}\,\mathrm{P}\left(Z>c+\mu\right) - \mathrm{P}\left(Z>c\right)^2, \qquad (3.99)\end{aligned}$$

where the last equation follows from the properties of the standard Gaussian density. The choice of μ that minimizes the variance under $\widehat{\mathrm{P}}$ is the solution to

$$\min_{\mu} e^{\mu^2}\,\mathrm{P}\left(Z>c+\mu\right).$$

Differentiating with respect to μ and setting the resulting expression to zero shows that the variance is minimized at μ^*, where

$$2\mu^*\left[1-\Phi\left(c+\mu^*\right)\right] - \phi\left(c+\mu^*\right) = 0, \qquad (3.100)$$

with Φ and ϕ being the standard Gaussian distribution function and density, respectively. This expression can be solved for μ^* with the aid of a numerical root solver. Alternatively, we can use the fact that c is large to rely on the asymptotic approximation

$$1-\Phi\left(c+\mu^*\right) \approx \frac{\phi\left(c+\mu^*\right)}{c+\mu^*},$$

which leads to

$$2\mu^* \frac{\phi(c+\mu^*)}{c+\mu^*} \approx \phi(c+\mu^*) \Rightarrow \mu^* \approx c. \tag{3.101}$$

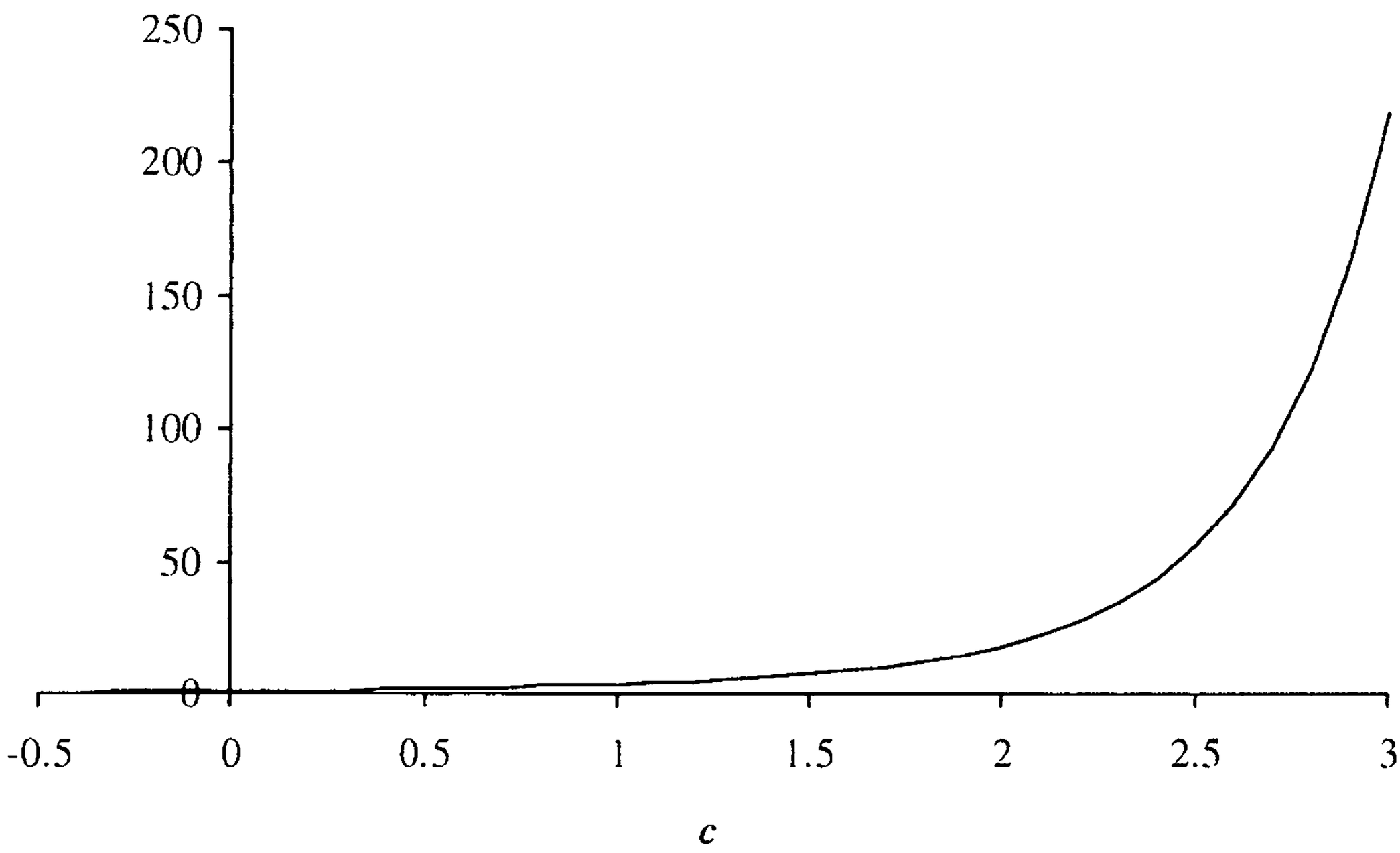
Note that this implies that the probability of Z exceeding c in measure $\widehat{\mathrm{P}}$ is approximately $\frac{1}{2}$. This is an intuitive result[24], consistent with the discussion at the end of Section 3.4.4.2.

To measure the efficacy of importance sampling, we can use (3.99) to define a variance efficiency ratio as

$$\frac{\mathrm{P}(Z>c)\,(1-\mathrm{P}(Z>c))}{e^{\mu^2}\mathrm{P}\,(Z>c+\mu)-\mathrm{P}\,(Z>c)^2}. \tag{3.102}$$

Figure 3.1 graphs this ratio when μ is set to c, as prescribed in (3.101). For large c, the improvements to variance associated with using importance sampling can be seen to be extremely significant.

Fig. 3.1. Variance Ratio



Notes: The figure graphs the ratio (3.102), with μ set according to (3.101).

It is also illustrative to consider the multi-variate extension to the problem above. Here, we are interested in estimating

$$\mathrm{E}^{\mathrm{P}}\left(1_{\{X>c\}}\right),$$

[24] For a somewhat more accurate approximation to μ^*, see Jäckel [2004].

where c is a p-dimensional constant and X is a p-dimensional vector of Gaussian random variables with mean 0 and covariance matrix Σ. Let C be the Cholesky decomposition of Σ, such that

$$\mathrm{E}^{\mathrm{P}}\left(1_{\{X>c\}}\right) = \mathrm{E}^{\mathrm{P}}\left(1_{\{CZ>c\}}\right) = \mathrm{E}^{\mathrm{P}}\left(1_{\{Z>c'\}}\right), \quad c' \triangleq C^{-1}c,$$

where Z is a p-dimensional vector of independent standard Gaussian variables. Let us introduce a measure $\widehat{\mathrm{P}}$ where the mean of Z has been shifted to μ, a p-dimensional vector. Following the same steps as for the univariate case, we have

$$\mathrm{E}^{\mathrm{P}}\left(1_{\{Z>c'\}}\right) = \mathrm{E}^{\widehat{\mathrm{P}}}\left(\exp\left(-\mu^{\top}Z + \frac{1}{2}\mu^{\top}\mu\right)1_{\{Z>c'\}}\right),$$

with variance

$$\begin{aligned}\mathrm{Var}^{\widehat{\mathrm{P}}}&\left(\exp\left(-\mu^{\top}Z + \frac{1}{2}\mu^{\top}\mu\right)1_{\{Z>c'\}}\right)\\ &= e^{\mu^{\top}\mu}\mathrm{E}^{\mathrm{P}}\left(1_{\{Z>c'+\mu\}}\right) - \mathrm{P}\left(X>c\right)^2.\end{aligned}$$

A direct optimization of this expression in μ involves multi-dimensional Gaussian integrals, so we wish to resort to approximations. We can use the arguments of Section 3.4.4.2 to argue that the optimal importance sampling density should be proportional to

$$1_{\{z>c'\}}\exp\left(-\frac{1}{2}z^{\top}z\right), \tag{3.103}$$

since $\exp(-z^{\top}z/2)$ is proportional to the normal density. Following the idea in Glasserman et al. [1999], we can choose μ such that the location of the peak of an $\mathcal{N}(\mu, I)$ distribution coincides with the peak of (3.103). In other words, we approximate the optimal μ as the value μ^* of z that solves

$$\max_z\left\{1_{\{z>c'\}}\exp\left(-\frac{1}{2}z^{\top}z\right)\right\} = \min_{z>c'}\left\{z^{\top}z\right\}. \tag{3.104}$$

If we assume, say, that all components of c' are larger than 0, then obviously

$$\mu^* = c' = C^{-1}c,$$

consistent with the approximative univariate result (3.101).

We note that the idea behind (3.104) is not limited to situations where we evaluate expectations of an indicator function. For instance, suppose we, as in Section 3.4.4.4, wish to estimate

$$\mathrm{E}^{\mathrm{P}}\left(G(Z)\right),$$

for a smooth function $G : \mathbb{R}^p \to \mathbb{R}$. Restricting ourselves again to the class of measure shifts that only move the mean of Z, the approximately optimal mean shift μ solves

$$\max_z \left\{ G(z) \exp\left(-\frac{1}{2} z^\top z \right) \right\}$$

or, if $G(Z)$ is strictly positive,

$$\max_z \left\{ w(z) - \frac{1}{2} z^\top z \right\}, \quad w(z) \triangleq \ln(G(z)).$$

The first-order condition for the optimum is

$$\nabla w(\mu^*) = (\mu^*)^\top, \tag{3.105}$$

where ∇ is the gradient operator, $\nabla = (\partial/\partial z_1, \ldots, \partial/\partial z_p)$ (row vector). This is a fixed-point condition that can be solved by numerical methods. The result (3.105) is exact if w is linear in its argument; the method above can thus be seen as a linearization through a first-order Taylor approximation. Glasserman et al. [1999] demonstrate that, under some conditions, (3.105) satisfies a certain asymptotic optimality property.

3.5 Some Notes on Bermudan Security Pricing

As alluded to in the beginning of this chapter, one drawback of Monte Carlo methods is the difficulty associated with the pricing of securities with early exercise rights. We demonstrated earlier in the chapter that Monte Carlo path generation runs *forward* in time, making direct application of dynamic programming and backward induction (see Chapters 1 and 2) impossible. Indeed, until the early 1990's, it was generally believed that Monte Carlo techniques were inherently incompatible with the pricing of early exercise rights. In the last decade, however, this belief has been overturned, with the advent of several different techniques for Monte Carlo pricing of options with early exercise rights. Most of these techniques are rather advanced and a detailed description will be postponed until later in this book, when the interest rate modeling foundation has been properly laid and the details of callable interest rate securities have been covered. For now, we only provide a brief discussion of certain generic principles, with additional details to be filled in later, in Chapters 18 and 19, among others. We start by establishing some notation and reminding the reader of some basic results from Chapter 1.

3.5.1 Basic Idea

For the remainder of this section, we consider the pricing of a Bermudan security C, with a payout function[25] $U(t) = U(t, x(t))$, where $x(t)$ is a

[25] For many exotic interest rate options, the function $U(t, x(t))$ may actually not be known in closed form. We deal with this complication in Chapter 18.

p-dimensional vector of Markovian state variables. The allowed discrete set of exercise dates is denoted $\mathcal{D} = \{T_1, T_2, \ldots, T_B\}$, with $T_B = T$ being the terminal maturity of C. We fix a numeraire N, assumed to be a function of $x(t)$, $N(t) = N(t, x(t))$. From Section 1.10, we recall that

$$C(0) = N(0) \sup_{\tau \in \mathcal{T}} \mathrm{E}^N \left(\frac{U(\tau)}{N(\tau)} \right),$$

where E^N denotes expectation in the measure Q^N induced by the numeraire N, and $\mathcal{T}$ is the set of stopping time strategies taking values in $\mathcal{D}$. More generally, we write

$$C(t) = N(t) \sup_{\tau \in \mathcal{T}(t)} \mathrm{E}_t^N \left(\frac{U(\tau)}{N(\tau)} \right), \tag{3.106}$$

where $\mathcal{T}(t)$ is the set of stopping time strategies in $\mathcal{D}$ for which $\tau \geq t$. We also recall that when $t \in (T_{i-1}, T_i]$, we have

$$C(t) = N(t) \mathrm{E}_t^N \left(N(T_i)^{-1} \max \left(H_i(T_i), U(T_i) \right) \right), \tag{3.107}$$

where the hold value (see Section 1.10) $H_i(T_i)$ is defined as

$$H_i(T_i) = N(T_i) \mathrm{E}_{T_i}^N \left(N(T_{i+1})^{-1} C(T_{i+1}) \right).$$

Notice that (3.107) establishes that the optimal exercise strategy, as seen from time t, is

$$\tau^* = \inf \{ T_i \geq t : U(T_i) \geq H_i(T_i) \}. \tag{3.108}$$

3.5.2 Parametric Lower Bound Methods

Assuming that we are able to simulate the p-dimensional vector $x(t)$ through time, it follows from (3.106) that a lower bound for $C(0)$ can be computed by Monte Carlo method through any exogenous guess for the optimal exercise strategy τ^*. One fairly intuitive approach involves a user-supplied specification of a parametric stopping rule, $\tau(\alpha) \in \mathcal{T}$, where $\alpha \in A \subset \mathbb{R}^m$ is an m-dimensional parameter vector. Defining $x = (x(T_1), \ldots, x(T_B))$, for a given value of α, we have the following algorithm.

1. Generate n independent paths $x^{(k)}$, $k = 1, \ldots, n$. For path k, let $\tau^{(k)}(\alpha)$ be the exercise date suggested by the parametric stopping rule.
2. For each path k, set $U^{(k)} = U(\tau^{(k)}(\alpha), x(\tau^{(k)}(\alpha)))$ and $C_\alpha^{(k)} = N(\tau^{(k)}(\alpha), x(\tau^{(k)}(\alpha)))^{-1} U^{(k)}$.
3. Return $\overline{C}_\alpha(0) = N(0) n^{-1} \sum_{k=1}^n C_\alpha^{(k)}$ as our estimate for the Bermudan option value $C(0)$.

Let $C_\alpha(0) \triangleq \mathrm{E}^N(\overline{C}_\alpha(0))$. As $\tau(\alpha)$ in general will be sub-optimal, it is clear that

$$C_\alpha(0) \leq C(0). \tag{3.109}$$

To get as close as possible to $C(0)$, it is, of course, preferable to use the value $\alpha^* \in A$ for which $C_\alpha(0)$ is optimized, i.e.

$$\alpha^* = \operatorname*{argsup}_{\alpha \in A} C_\alpha(0).$$

A tempting way of estimating $C_{\alpha^*}(0)$ would be to modify step 3 in the algorithm above to

3a. Return $\overline{C}_{\alpha^*}(0) = \sup_{\alpha \in A} \overline{C}_\alpha(0)$.

Leaving aside the question of how one might execute the optimization in Step 3a, we notice that the estimator $\overline{C}_{\alpha^*}(0)$ is biased high relative to $C_{\alpha^*}(0)$:

$$\mathrm{E}^N\left(\overline{C}_{\alpha^*}(0)\right) \geq \sup_{\alpha \in A} C_\alpha(0). \tag{3.110}$$

This inequality states that the expected value of the maximum over α must be at least as large as the maximum over α of expected values, a consequence of Jensen's inequality. We may interpret the bias of $\mathrm{E}^N(\overline{C}_{\alpha^*}(0))$ as a *perfect foresight bias*: by using in-sample information to estimate α^*, we effectively "cheat" by making the optimum specific to the same n samples that are also used to determine the option value.

The combination of inequalities (3.109) and (3.110) shows that the quantity $\overline{C}_{\alpha^*}(0)$ from Step 3a has an *indeterminate* bias relative to the true option price. As a bias is generally inevitable when using parametric exercise strategies, in practice it is preferable to at least know its sign. To accomplish this, we can retain the estimated value of α^* found as described above, but draw a *separate* set of Monte Carlo paths when pricing the option. That is, we replace Step 3a with the following two steps:

3b. Set $\widehat{\alpha}^* = \operatorname{argsup}_{\alpha \in A} \overline{C}_\alpha(0)$.
4b. Draw a fresh set of n_2 independent paths for x and N, with α locked at the value $\widehat{\alpha}^*$. Return $\overline{C}_{\widehat{\alpha}^*}(0) = N(0)/n_2 \sum_{k=1}^{n_2} C_{\widehat{\alpha}^*}^{(k)}$, where the $C_{\widehat{\alpha}^*}^{(k)}$ are computed on the new set of paths.

As the parameter $\widehat{\alpha}^*$ will a.s. never equal α^*, it follows that

$$\mathrm{E}^N\left(\overline{C}_{\widehat{\alpha}^*}(0)\right) \leq C_{\alpha^*}(0) \leq C(0),$$

i.e. we are now assured that $\overline{C}_{\widehat{\alpha}^*}(0)$ is low-bound estimator.

3.5.3 Parametric Lower Bound: An Example

What constitutes a good parametric exercise rule is strongly instrument-specific and typically requires case-by-case analysis. Even for simple Markov models and standard option payouts, the topology of exercise and continuation regions can be highly complicated (see e.g. Broadie and Detemple [1997]), so this exercise is by no means straightforward. As a first approximation, however, one can always attempt to use a simple rule based on outright "moneyness" of the underlying option payout, as in Andersen [2000a]. According to this rule, one sets $\alpha = (h_1, h_2, \ldots, h_B)$ (i.e. $m = B$) and writes

$$\tau(\alpha) = \inf\{T_i : U(T_i, x(T_i)\} > h_i). \tag{3.111}$$

That is, exercise of the option takes place when it is sufficiently deep in the money, with the term "sufficiently deep" quantified through unknown trigger thresholds $h_i \geq 0$, $i = 1, \ldots, B$.

While α is B-dimensional, finding its optimal value is not truly a B-dimensional optimization problem. Rather, due to the Markov assumption on $x(t)$, we may decompose it into a series of $B-1$ *one-dimensional* optimization problems. Specifically, working backwards in time, suppose that the optimal values of $h_{j+1}, h_{j+2}, \ldots, h_B$ are known. We then find the optimal value of h_j, by optimizing on $\overline{C}_\alpha(0)$, but subject to the constraint that exercise is *not* allowed to take place before time j. As $h_{j+1}, h_{j+2}, \ldots, h_B$ are assumed known — and $h_1, \ldots, h_{j-1}$ do not come into play — the only variable involved in this optimization is h_j. The algorithm starts with $j = B-1$ and the known[26] boundary condition $h_B = 0$.

A couple of comments on the algorithm above are in order. First, we notice that $U(T_i, x(T_i)) > h_i$ can be replaced with any one-parameter boolean function $g(T_i, x(T_i); h_i)$ without affecting the basic algorithm — see Andersen [2000a] for some examples. Second, if in such a boolean function $g(T_i, x(T_i); h_i)$ each h_i is allowed to be q-dimensional, the optimization problem reduces to $B-1$ q-dimensional optimization problems. And third, for a finite-path simulation, the objective functions in each of the $B-1$ optimization problems will not be smooth; consequently, the optimization is best performed by an iterative search rather than a derivative-based method. Andersen [2000a] uses the golden section search (see Press et al. [1992]), but simpler strategies based on, say, outright sorting are also possible.

3.5.4 Regression-Based Lower Bound

According to (3.108), an approximation for the optimal exercise strategy can always be constructed through an estimate for hold values $H_i(T_i)$ at all $i = 1, 2, \ldots, B-1$. In our Markov setting, we know that

[26] At the last possible exercise date, we would, of course, always exercise the option if it is in-the-money.

$$H_i(T_i) = q_i x(T_i)) = q_i(x_1(T_i), \ldots, x_p(T_i)$$

for a set of $B-1$ functions $q_i : \mathbb{R}^p \to \mathbb{R}$, $i = 1, 2, \ldots, B-1$; the problem of estimating hold values is equivalent to the problem of estimating the functions q_i.

From Section 3.5.1, we know that

$$q_i(x) = N(T_i, x)\mathrm{E}^N \left(C(T_{i+1})/N\left(T_{i+1}, x(T_{i+1})\right) | x(T_i) = x\right), \tag{3.112}$$

which can be interpreted as the *regression* of $C(T_{i+1})/N(T_{i+1})$ on the Markov state variables $x(T_i)$. Several authors — including Carrière [1996], Longstaff and Schwartz [2001], and Tsitsiklis and Roy [2001] — have used this observation to suggest that $q_i(x)$ be estimated by a linear combination of exogenously specified (basis) functions of $x(T_i)$, with least-squares regression on Monte Carlo paths used to determine the best weights for these functions. That is, we fundamentally assume that

$$q_i(x) = \sum_{j=1}^{d} \beta_{i,j}\psi_j(x), \tag{3.113}$$

for a set of d basis-functions $\psi_j : \mathbb{R}^p \to \mathbb{R}$, $j = 1, 2, \ldots, d$. Setting $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,d})^\top$ and $\psi(x) = (\psi_1(x), \ldots, \psi_d(x))^\top$, we can rewrite (3.113) as $q_i(x) = \psi(x)^\top \beta_i$ or, from (3.112),

$$\begin{aligned}\mathrm{E}^N \left(N\left(T_i, x(T_i)\right) C(T_{i+1})/N\left(T_{i+1}, x(T_{i+1})\right) | x(T_i) = x\right) \\ = \mathrm{E}^N \left(\psi\left(x(T_i)\right)^\top | x(T_i) = x\right) \beta_i.\end{aligned}$$

This, in turn, implies that

$$\Omega_i = \Psi_i \beta_i \quad \Rightarrow \quad \beta_i = \Psi_i^{-1}\Omega_i, \tag{3.114}$$

where Ω_i is the d-dimensional vector

$$\Omega_i = \mathrm{E}^N \left(\psi\left(x(T_i)\right) \frac{N\left(T_i, x(T_i)\right)}{N\left(T_{i+1}, x(T_{i+1})\right)} C(T_{i+1})\right),$$

and Ψ_i is the $d \times d$ matrix

$$\Psi_i = \mathrm{E}^N \left(\psi\left(x(T_i)\right) \psi\left(x(T_i)\right)^\top\right).$$

The rationale for rewriting (3.113) into the seemingly more convoluted representation (3.114) is that the latter leads naturally to the algorithm for a least-squares estimation of β_i: one simply replaces the expectations in Ψ_i and Ω_i with sample averages $\overline{\Psi}_i$ and $\overline{\Omega}_i$ computed on a set of Monte Carlo paths. That is, one uses[27]

[27] In practice, a direct solution of linear equations in this fashion can be suboptimal if the matrix $\overline{\Psi}_i$ is ill-conditioned. Instead, one would use either truncated singular value decomposition (TSVD) or Tikhonov regularization to find $\widehat{\beta}_i$. We return to this issue in Chapter 18.

$$\widehat{\beta}_i = \overline{\Psi}_i^{-1}\overline{\Omega}_i \tag{3.115}$$

as the sample estimate.

We shall discuss the details of (and many variations on) the regression approach later, in Chapter 18. For now, let us just notice that computation of $\overline{\Omega}_i$ requires estimation of $C(T_{i+1})$, which naturally encourages running the estimation of the $\widehat{\beta}_i$ backwards in i, starting from $i = B - 1$. We also notice that the success of the regression approach depends critically on the choice and number of basis functions ψ_j. We give specific advice on this topic in Chapters 18 and 19.

3.5.5 Upper Bound Methods

Given a martingale M in measure Q^N, we recall from Section 1.10.2 that an M-specific upper bound $C_M(0)$ for a Bermudan option can always be constructed as

$$C_M(0) = N(0)\left\{M(0) + \mathrm{E}^N\left(\max_{t\in\mathcal{D}}\left(\frac{U(t,x(t))}{N(t,x(t))} - M(t)\right)\right)\right\} \geq C(0). \tag{3.116}$$

Let $\mathbf{M} = (M(T_1), \ldots, M(T_B))^\top$. As long as $M(\cdot)$ can be simulated along with the vector $x(\cdot)$, (3.116) suggests the following Monte Carlo algorithm:

1. Generate n independent paths $x^{(k)}$, $\mathbf{M}^{(k)}$, $k = 1, \ldots, n$. For path k, let $\gamma^{(k)}$ be the maximum value of $U(T_i, x(T_i))/N(T_i, x(T_i)) - M(T_i)$ over $i = 1, 2, \ldots, B$.
2. Return $\overline{C}_M(0) = N(0)\{M(0) + n^{-1}\sum_{k=1}^n \gamma^{(k)}\}$.

For the upper bound method to be practically useful, we would want the gap $\mathrm{E}^N(\overline{C}_M(0)) - C(0)$ to be small. This, in turn, requires that we specify the martingale $M(t)$ to be "reasonable". More specifically, from results in Section 1.10.2, we would like $M(t)$ to represent the martingale component of a good approximation to the supermartingale $C(t)$. For instance, if we happen to think that $C(t)$ is well-approximated by some known function v of time and x,

$$C(t) = v(t, x(t)), \tag{3.117}$$

we would set

$$M(t) = \sum_{j=1}^{p} \frac{\partial v(t,x(t))}{\partial x_j}\left(dx_j(t) - \mathrm{E}_t^N(dx_j(t))\right). \tag{3.118}$$

As an example, suppose that W^N is a (possibly vector-valued) Brownian motion in Q^N and $dx_j(t) = \mu_j(t, x(t))\,dt + \sigma_j(t, x(t))\,dW^N(t)$, in which case we could use

$$M(t) = \sum_{j=1}^{p} \frac{\partial v(t,x(t))}{\partial x_j}\sigma_j(t,x(t))\,dW^N(t). \tag{3.119}$$

Occasionally, a natural analytical guess for the function v may exist, but most often the only estimate for v is given only implicitly, through a low-bound estimator based on an approximation of the optimal exercise strategy. A completely generic algorithm to turn a guess for the optimal exercise strategy into a proxy for $M(t)$ is developed in Andersen and Broadie [2004]; we shall discuss this algorithm in detail in Chapter 18. If the evolution of $x(t)$ is described by an SDE, a regression approach to estimation of the terms multiplying $dW^N(t)$ in (3.119) can be found in Belomestny et al. [2007].

3.5.6 Confidence Intervals

Suppose that we simultaneously apply a lower bound and an upper bound method to provide two sample estimates $\overline{C}_{\text{lo}}$ and $\overline{C}_{\text{up}}$, with

$$\mathrm{E}^N\left(\overline{C}_{\text{lo}}\right) \leq C(0) \leq \mathrm{E}^N\left(\overline{C}_{\text{up}}\right). \tag{3.120}$$

Let us assume that the sample standard errors on $\overline{C}_{\text{lo}}$ and $\overline{C}_{\text{up}}$ have been computed as s_{lo} and s_{up}, respectively. For a sufficiently large number of Monte Carlo trials, we can then use the central limit theorem from Section 3.1 to set up a confidence interval

$$\left[\overline{C}_{\text{lo}} - u_{\gamma/2} \cdot s_{\text{lo}}, \overline{C}_{\text{up}} + u_{\gamma/2} \cdot s_{\text{up}}\right],$$

where $\Phi(u_{\gamma/2}) = 1 - \gamma/2$. It is clear from (3.120) that the likelihood of this interval bracketing the true price $C(0)$ is *at least* $1 - \gamma$. It is also clear that this confidence interval will not shrink to zero — even in the limit of an infinite number of samples where $s_{\text{lo}} \to 0$ and $s_{\text{up}} \to 0$ — unless $\overline{C}_{\text{lo}}$ and $\overline{C}_{\text{up}}$ simultaneously achieve the unlikely feat of being perfectly unbiased estimators for $C(0)$.

Finally, let us note that any number inside the interval $[\overline{C}_{\text{lo}}, \overline{C}_{\text{up}}]$ can reasonably be used as an estimator for $C(0)$. To the extent that we have reason to believe[28] that $\overline{C}_{\text{lo}}$ and $\overline{C}_{\text{up}}$ have roughly opposite biases, a natural estimator is $(\overline{C}_{\text{lo}} + \overline{C}_{\text{up}})/2$.

3.5.7 Other Methods

The methods described so far are those that, in our opinion, are most useful for practical Monte Carlo pricing of interest rate options with early exercise rights. Several other methods, however, have been proposed in the literature, some of which have interesting theoretical properties. We highlight the *random tree* method (Broadie and Glasserman [1997]) which builds a random

[28] There is some evidence that the upper-bound method in Andersen and Broadie [2004] produces a bias that is often roughly opposite of that of the low-bound method from which the martingale M in (3.116) is extracted.

non-recombining lattice by Monte Carlo methods; backward induction arguments are then used to construct high- and low-biased estimators, both of which are convergent to the true price as the number of Monte Carlo paths are increased. The drawback of the method is its computational complexity which increases exponentially in the number of exercise dates (B), ruling out its practical usage for many realistic applications. Broadie and Glasserman [2004] suggest a recombining *stochastic mesh* method that grows only linearly in the number of exercise weights; in its basic form, this method requires explicit knowledge of transition densities as it relies on likelihood ratios to set weights on nodes in the mesh (see Section 3.3.3). As discussed in Glasserman [2004], the concept of stochastic meshes can, however, be broadened to include several other methods, include the regression approach in Section 3.5.4.

3.A Appendix: Constants for Φ^{-1} Algorithm

a_0	2.50662823884	c_0	0.3374754822726147
a_1	-18.61500062529	c_1	0.9761690190917186
a_2	41.39119773534	c_2	0.1607979714918209
a_3	-25.44106049637	c_3	0.0276438810333863
b_0	-8.47351093090	c_4	0.0038405729373609
b_1	23.08336743743	c_5	0.0003951896511919
b_2	-21.06224101826	c_6	0.0000321767881768
b_3	3.13082909833	c_7	0.0000002888167364
		c_8	0.0000003960315187

4

Fundamentals of Interest Rate Modeling

The purpose of this brief chapter is twofold. First, we introduce notations to characterize prices and yields of basic fixed income market securities. In addition to providing the foundation for a more expansive discussion of fixed income markets (which we shall undertake in Chapter 5), this part of the chapter serves to identify and characterize a number of probability measures that are of fundamental importance in models for the term structure of interest rates. A brief discussion of measures used in a two-currency setting is also provided.

In the second part of the chapter, we discuss general characteristics of models with dynamics driven by vector-valued Brownian motions. This analysis leads to the fundamental class of Heath-Jarrow-Morton (HJM) (see Heath et al. [1992]) models of continuously compounded forward rates. Among other special cases, we discuss in some detail tractable HJM models with Gaussian volatility structure, and provide some results for the case where such models are Markovian. These discussions continue in Chapters 10 through 12 where we consider one- and multi-factor short rate models, and in the Chapter 13 where we introduce the important class of quasi-Gaussian HJM models with local and stochastic volatility.

4.1 Fixed Income Notations

4.1.1 Bonds and Forward Rates

As in earlier chapters, let $P(t,T)$ denote the time t price of a zero-coupon bond (also known as a *discount bond*) delivering for certain \$1 at maturity $T \geq t$. Suppose we are interested in purchasing at some future time T a zero-coupon bond maturing at $T+\tau$, $\tau > 0$. At time $t < T$, the price of such a bond can be locked in by i) purchasing at time t one $(T+\tau)$-maturity zero-coupon bond; and ii) selling short ("shorting") $P(t,T+\tau)/P(t,T)$

T-maturity zero-coupon bonds. The time t cost of executing this strategy is zero,

$$-1 \cdot P(t, T+\tau) + P(t, T+\tau)/P(t, T) \cdot P(t, T) = 0,$$

but a flow of

$$-P(t, T+\tau)/P(t, T)$$

will take place at time T as the T-maturity short position matures. This is compensated by an inflow of \$1 at time $T+\tau$. In other words, our trading strategy effectively fixes the time T purchase price of the $(T+\tau)$-maturity bond at

$$P(t, T, T+\tau) \triangleq P(t, T+\tau)/P(t, T), \quad \tau > 0,$$

a quantity known as the time t *forward price* for the zero-coupon bond spanning $[T, T+\tau]$.

It is often convenient to characterize a forward bond price by a discount rate. One such rate is the *continuously compounded forward yield* $y(t, T, T+\tau)$, defined by

$$e^{-y(t,T,T+\tau)\tau} = P(t, T, T+\tau). \tag{4.1}$$

The time between the maturity of the forward bond and the expiry of the forward contract, i.e. τ, is often called the *tenor* of the forward bond or the forward yield. In the definition of the continuously compounded yield lies an implicit, and idealized, assumption of continuous reinvestment of investment proceeds. Most actual market quotes, however, are based on discrete-time compounding of proceeds. Accordingly, we define a *simple forward rate* $L(t, T, T+\tau)$ as

$$1 + \tau L(t, T, T+\tau) = 1/P(t, T, T+\tau). \tag{4.2}$$

Again, τ is the *tenor* of the forward rate. For an arbitrary set of dates $T = T_0 < T_1 < T_2 < \ldots < T_n$, notice that forward bond prices can be recovered from forward rates by simple compounding,

$$P(t, T_n)/P(t, T) = \prod_{i=1}^{n} \frac{1}{1 + (T_i - T_{i-1})\, L(t, T_{i-1}, T_i)}.$$

Unless we state otherwise, throughout this book we shall typically make the assumption that spot rates $L(T, T, T+\tau)$ are the *Libor (London Interbank Offered Rate) rates* quoted in the interbank market. Libor rates are quoted on values of τ ranging from one week ($\tau = 1/52$)[1] to 12 months ($\tau = 1$), and form the basis for a number of floating-rate derivative contracts, such as interest rate swaps and Eurodollar futures. We shall examine these securities in more detail in Chapter 5.

[1]Note that in reality the calculation of year fractions τ are governed by fairly complicated market conventions. A brief discussion of this topic can be found in Appendix 5.A.

In the limit $\tau \downarrow 0$,

$$L(t, T, T+\tau) \to f(t, T),$$

where the quantity $f(t, T)$ is the time t *instantaneous forward rate* to time T. We think of $f(t, T)$ as the forward rate spanning $[T, T+dT]$, observed at time t. The relation between instantaneous forward rates and bond prices is given by the continuous compounding formula

$$P(t, T, T+\tau) = \exp\left(-\int_T^{T+\tau} f(t, u)\, du\right), \tag{4.3}$$

such that

$$f(t, T) = -\frac{\partial \ln P(t, T)}{\partial T}, \tag{4.4}$$

and, from (4.1),

$$y(t, T, T+\tau) = \tau^{-1} \int_T^{T+\tau} f(t, u)\, du, \quad f(t, T) = \lim_{\tau \downarrow 0} y(t, T, T+\tau).$$

We also notice the relationship

$$f(t, T) = \frac{\partial\left(y(t, t, T)(T-t)\right)}{\partial T} = y(t, t, T) + (T-t)\frac{\partial y(t, t, T)}{\partial T}.$$

The quantity

$$r(t) \triangleq f(t, t) \tag{4.5}$$

is an $\mathcal{F}_t$-measurable random variable known as the *short rate* or sometimes the *spot rate*. Loosely speaking, we can think of $r(t)$ as the overnight rate in effect at time t.

Finally, let us note that interest rates of various flavors and related quantities are typically quoted in percentage points or sometimes in *basis points*, where 1 basis point = 1/100 of one percent.

4.1.2 Futures Rates

Through the market for Eurodollar futures (see Chapter 5), investors can enter into securities that will pay at time T an amount of

$$1 - L(T, T, T+\tau). \tag{4.6}$$

At time 0, a Eurodollar futures contract can be entered into at no upfront cost, but with an implicit obligation of the holder to pay at time T per unit of notional

$$1 - F(0, T, T+\tau)$$

in return for the payout (4.6). Here, $F(t,T,T+\tau)$ is the time t *simple futures rate* for the period $[T,T+\tau]$. Importantly, the futures rate is *marked to market* (or *resettled*) each day, with the day's change in the futures rate immediately credited to or debited from the contract holder's account with the futures exchange. Specifically, after holding the contract for a period of $\Delta = 1$ day, the futures contract holder would thus experience a cash flow of

$$(1 - F(\Delta,T,T+\tau)) - (1 - F(0,T,T+\tau)) \\ = -(F(\Delta,T,T+\tau) - F(0,T,T+\tau)).$$

Continuing the mark-to-market process to maturity shows that the total amount of cash flow received by the holder on $[0,T]$ is

$$-(F(T,T,T+\tau) - F(0,T,T+\tau)) = -(L(T,T,T+\tau) - F(0,T,T+\tau)) \tag{4.7}$$

where we have used the fact that $F(T,T,T+\tau)$ must equal $L(T,T,T+\tau)$ to avoid a *delivery arbitrage.*

The fact that the net cash flow payment (4.7) on a Eurodollar futures contract has been made incrementally on a daily basis has important valuation consequences, and causes the futures rate to differ from the forward rate defined earlier. For instance, under a scenario of rising interest rates, the holder of a Eurodollar futures contract must make payments to the futures exchange. As rates are rising, the contract holder will be faced with a high-rate — and thus unfavorable — borrowing environment for funding these payments. Conversely, when interest rates fall, the reinvestment of received funds will take place at increasingly low rates. Due to the adverse behavior of funding costs and reinvestment gains, we would expect the purchaser of a Eurodollar futures contract to pay less for these instruments than for a comparable instrument without daily mark-to-market. Consequently, we would expect the futures rate to be *above* the corresponding forward rate. We shall quantify this effect in Section 4.5.1 and, with more advanced models, in Section 16.8.

We notice that we can define *instantaneous futures rates* $q(t,T)$ in the same fashion as we defined instantaneous forward rates:

$$q(t,T) = \lim_{\tau \downarrow 0} F(t,T,T+\tau).$$

4.1.3 Annuity Factors and Par Rates

Most fixed income securities involve multiple cash flows taking place on a pre-set schedule of dates, often referred to as a *tenor structure*,

$$0 \le T_0 < T_1 < \ldots < T_N.$$

Given a tenor structure, for any two integers k,m satisfying $0 \le k < N$, $m > 0$, and $k+m \le N$, we can define an *annuity factor* $A_{k,m}$ by

$$A_{k,m}(t) = \sum_{n=k}^{k+m-1} P(t, T_{n+1})\tau_n, \quad \tau_n = T_{n+1} - T_n. \tag{4.8}$$

Annuity factors provide for compact notation when pricing coupon-bearing securities. For instance, a security making m coupon payments of $c\tau_n$ at all $T_{n+1}, n = k, \ldots, k+m-1$, is easily seen to have time t value of

$$cA_{k,m}(t), \quad t \leq T_k.$$

If the security also involves a back-end return of notional at time T_{k+m} (as is the case for a regular coupon-bearing bond), the pricing expression is

$$cA_{k,m}(t) + P(t, T_{k+m}), \tag{4.9}$$

where we assume that the bond has been normalized to have a unit notional. The time t forward price to T_k of the security (4.9) is

$$cA_{k,m}(t)/P(t, T_k) + P(t, T_{k+m})/P(t, T_k);$$

the value of the coupon c for which this expression equals 1 is known as the *forward par rate* or, when used in the context of swap pricing, as the *forward swap rate*. With $S_{k,m}(t)$ denoting the time t swap rate, we apparently have

$$S_{k,m}(t) = \frac{P(t, T_k) - P(t, T_{k+m})}{A_{k,m}(t)}, \quad t \leq T_k. \tag{4.10}$$

From the definition of $L(t, T_n, T_{n+1})$ in (4.2), a little thought shows that the numerator of the expression for $S_{k,m}(t)$ can be expanded into a weighted sum of forward rates, leading to the alternative expression

$$S_{k,m}(t) = \frac{\sum_{n=k}^{k+m-1} \tau_n P(t, T_{n+1}) L_n(t)}{A_{k,m}(t)}, \quad t \leq T_k, \tag{4.11}$$

where we have introduced the useful shorthand

$$L_n(t) \triangleq L(t, T_n, T_{n+1}).$$

It follows that the forward swap rate can be loosely interpreted as a weighted average of simple forward rates on the specified tenor structure. We note for the future that the time T_k is sometimes referred to as the *fixing date*, or *expiry*, of the swap rate $S_{k,m}$, while the length of the corresponding swap, $T_{k+m} - T_k$, is sometimes called the *tenor* of the swap rate.

4.2 Fixed Income Probability Measures

As discussed in Chapter 1, selecting an equivalent martingale measure is largely a matter of choosing a *numeraire*, an asset price process used to

re-normalize the prices of other traded securities. For later reference, this section lists and names a number of important numeraires and measures used in fixed income pricing. Throughout the section, we assume that the market is complete, and we use $V(t)$ to denote the time t price of a derivative security making an $\mathcal{F}_T$-measurable payment of $V(T)$.

4.2.1 Risk Neutral Measure

The numeraire defining the risk-neutral measure Q is the continuously compounded money market account $\beta(t)$, satisfying the locally deterministic SDE

$$d\beta(t) = r(t)\beta(t)\,dt, \quad \beta(0) = 1, \tag{4.12}$$

where $r(t)$ is the short rate, $r(t) = f(t,t)$. Solving this equation yields

$$\beta(t) = e^{\int_0^t r(u)\,du}.$$

From the results of Chapter 1, in the absence of arbitrage the numeraire-deflated process $V(t)/\beta(t)$ must be a martingale, implying the derivative security valuation formula

$$V(t)/\beta(t) = \mathrm{E}_t^{\mathrm{Q}}\left(V(T)/\beta(T)\right), \quad t \le T, \tag{4.13}$$

or equivalently

$$V(t) = \mathrm{E}_t^{\mathrm{Q}}\left(e^{-\int_t^T r(u)\,du} V(T)\right). \tag{4.14}$$

If we apply (4.14) to the special case of $V(T) = 1$, we obtain a fundamental bond pricing formula. We highlight the importance of this result by listing it in a lemma.

Lemma 4.2.1. *In the absence of arbitrage, the time t price $P(t,T)$ of a T-maturity zero-coupon bond is*

$$P(t,T) = \mathrm{E}_t^{\mathrm{Q}}\left(e^{-\int_t^T r(u)\,du}\right). \tag{4.15}$$

It follows from Lemma 4.2.1 that specification of the dynamics of $r(t)$ under Q suffices to determine the prices of discount bonds at all times and maturities. Models that are based on such a direct specification of $r(t)$ dynamics are known as *short rate models* and are the subject of Chapters 10 through 12. Notice the resemblance between expressions (4.3) and (4.15); if $r(t)$ is deterministic, the two expressions will agree as $r(u) = f(t,u)$, $u \ge t$. If $r(t)$ is random, one may wonder whether this result will hold in expectation. The answer to this is negative, i.e.

$$f(t,u) \ne \mathrm{E}_t^{\mathrm{Q}}\left(r(u)\right), \tag{4.16}$$

provided r is random. We prove this in the section below. Under certain idealized conditions, however, equality holds in (4.16) provided $f(t,u)$ is replaced by the *futures* rate $q(t,u)$. The exact result is as follows.

Lemma 4.2.2. *Assume that mark-to-market takes place continuously. Under regularity conditions on the short rate $r(\cdot)$ — it suffices that $r(\cdot)$ is positive and bounded — the futures rate $F(\cdot, T, T+\tau)$ is a Q-martingale, and*

$$F(t, T, T+\tau) = \mathrm{E}_t^{\mathrm{Q}}\left(L(T, T, T+\tau)\right). \tag{4.17}$$

Proof. Over a small interval $[t, t+dt]$, we have earlier shown that the cash proceeds from a futures contract are proportional to

$$dF(t, T, T+\tau) = F(t+dt, T, T+\tau) - F(t, T, T+\tau).$$

Suppose that we hold the futures contract up to some arbitrary horizon $t < \mathcal{T} \leq T$ at which point we exit (e.g., by selling the futures contract). Deflating all cash proceeds from this strategy with the numeraire $\beta(t)$ and integrating provides us with the time t value of the futures contract as

$$V_{\mathrm{fut}}(t) = \beta(t)\mathrm{E}_t^{\mathrm{Q}}\left(\int_t^{\mathcal{T}} \beta(s)^{-1}\, dF(s, T, T+\tau) + \beta(\mathcal{T})^{-1} V_{\mathrm{fut}}(\mathcal{T})\right).$$

As it is always costless to enter into a futures contract, $V_{\mathrm{fut}}(t) = V_{\mathrm{fut}}(\mathcal{T}) = 0$ by definition, so for arbitrary $t < \mathcal{T} \leq T$ we must have (since $\beta(t)$ is positive)

$$\mathrm{E}_t^{\mathrm{Q}}\left(\int_t^{\mathcal{T}} \beta(s)^{-1}\, dF(s, T, T+\tau)\right) = 0. \tag{4.18}$$

Provided that $\beta(s)^{-1}$ is almost surely positive (which is the case if r is bounded), the fact that (4.18) holds for arbitrary horizons $t < \mathcal{T} \leq T$ shows that

$$\mathrm{E}_t^{\mathrm{Q}}\left(dF(s, T, T+\tau)\right) = 0, \quad t \leq s \leq T,$$

which demonstrates that F is a Q-martingale. The result (4.17) then immediately follows. □

Equation (4.17) states that the futures rate is the Q-expectation of the time T spot rate $L(T, T, T+\tau)$. A similar relation must then hold for the instantaneous futures rate, i.e.

$$q(t, u) = \mathrm{E}_t^{\mathrm{Q}}\left(r(u)\right) \tag{4.19}$$

as stated earlier[2].

Lemma 4.2.2 was first proven by non-probabilistic methods in Cox et al. [1981], who employed a direct, and quite instructive, hedging argument to

[2]This result should not be confused with the classical *expectations hypothesis* which states that futures (or sometimes forward) rates are unbiased estimators of future spot rates, in the *real-life* probability measure P: $\mathrm{E}_t^{\mathrm{P}}(r(T)) = q(t, T)$. The expectations hypothesis amounts to a strong assumption about the market price of risk (see Chapter 1), whereas equation (4.19) is a preference-free arbitrage relationship.

show the result. The assumption of continuous resettlement in the lemma may appear idealized, but the difference between daily and continuous settlement is quite small, as shall be demonstrated in Chapter 16. Explicit modeling of discrete resettlement is nevertheless quite straightforward, and basically involves shifting measure, from the risk-neutral measure to the so-called spot measure, defined below. We return to this issue in Chapter 16.

4.2.2 T-Forward Measure

The T-forward measure Q^T was introduced in Jamshidian [1991b] (see also Geman et al. [1995]), and uses a T-maturity zero-coupon bond as the numeraire asset. As is customary, we let $\mathrm{E}^T(\cdot)$ denote expectations in measure Q^T, such that

$$V(t)/P(t,T) = \mathrm{E}_t^T\left(V(T)/P(T,T)\right), \quad t \leq T.$$

As obviously $P(T,T) = 1$, this expression simplifies to the convenient form

$$V(t) = P(t,T)\mathrm{E}_t^T\left(V(T)\right). \tag{4.20}$$

Comparison of (4.20) and (4.14) shows that shifting to the T-forward measure in a sense decouples the expectation of the terminal payout $V(T)$ from that of the numeraire. As we shall see, this is often very convenient when we attempt to construct analytical formulas for prices of certain simple interest rate derivatives. From the results of Section 1.3, we note that the explicit connection between the risk-neutral and T-forward measures is given by the density

$$\mathrm{E}_t^{\mathrm{Q}}\left(\frac{d\mathrm{Q}^T}{d\mathrm{Q}}\right) = \frac{P(t,T)/P(0,T)}{\beta(t)}. \tag{4.21}$$

As $P(t,T+\tau)$ is the price of a traded asset, from the definition of the T-forward measure it follows that forward bond prices

$$P(t,T,T+\tau) = P(t,T+\tau)/P(t,T)$$

are martingales in the T-forward measure. We highlight a related result for forward rates below.

Lemma 4.2.3. *In the absence of arbitrage the forward Libor rate $L(t,T,T+\tau)$ is a martingale under $\mathrm{Q}^{T+\tau}$, such that*

$$L(t,T,T+\tau) = \mathrm{E}_t^{T+\tau}\left(L(T,T,T+\tau)\right), \quad t \leq T. \tag{4.22}$$

Proof. By definition (see (4.2))

$$L(t,T,T+\tau) = \tau^{-1}\left(P(t,T)/P(t,T+\tau) - 1\right).$$

As $P(t,T)/P(t,T+\tau)$ is a martingale under $\mathrm{Q}^{T+\tau}$, so is $L(t,T,T+\tau)$. The result follows. □

Taking the limit $\tau \downarrow 0$ and setting $T = u$ yields

$$f(t,u) = \mathrm{E}_t^u\left(f(u,u)\right) = \mathrm{E}_t^u\left(r(u)\right), \tag{4.23}$$

which should be compared to the result (4.19).

4.2.3 Spot Measure

When working with a multitude of forward rates on a tenor structure $0 = T_0 < T_1 < \ldots < T_N$, it is often convenient to introduce a numeraire that can be extended to arbitrary horizons by compounding. While the continuously compounded money market account β would accomplish this, working with a continuously compounded numeraire is inherently awkward in a setting with a discrete tenor structure. As an alternative, we can introduce a discrete-time equivalent of the continuously compounded money market account to be the value of the following trading strategy. At time 0, \$1 is invested in $1/P(0,T_1)$ T_1-maturity discount bonds, returning the amount

$$1/P(0,T_1) = 1 + \tau_0 L(0,0,\tau_0)$$

at time T_1. This amount is then reinvested ("rolled") at time T_1 in T_2-maturity bonds, returning

$$1/P(0,T_1)\cdot 1/P(T_1,T_2) = \left(1+\tau_0 L(0,0,T_1)\right)\left(1+\tau_1 L(T_1,T_1,T_2)\right)$$

at time T_2. Repeating this re-investment strategy at each date in the tenor structure gives rise to an asset price process $B(t)$, where $B(0) = 1$ and

$$B(t) = \prod_{n=0}^{i}\left(1+\tau_n L_n(T_n)\right)P(t,T_{i+1}), \quad T_i < t \le T_{i+1}, \tag{4.24}$$

where we used the already introduced short-hand notation $L_n(t) = L(t,T_n,T_{n+1})$. The process $B(t)$ is effectively a rolling certificate of deposit, and can be interpreted as a discrete-time equivalent of $\beta(t)$. $B(t)$ will approach $\beta(t)$ as the time spacing of the tenor structure is made increasingly fine.

The measure induced by $B(t)$ is known as the *spot measure* (or sometimes *spot Libor measure*), denoted Q^B. With $\mathrm{E}^B(\cdot)$ denoting expectations in this measure we have

$$V(t) = \mathrm{E}_t^B\left(V(T)\frac{B(t)}{B(T)}\right),$$

where

$$\frac{B(t)}{B(T)} = \prod_{n=i+1}^{j}\left(1+\tau_n L_n(T_n)\right)^{-1}\frac{P(t,T_{i+1})}{P(T,T_{j+1})},$$
$$T_i < t \le T_{i+1}, \quad T_j < T \le T_{j+1}.$$

The similarity between the discrete and continuous money market accounts makes the spot Libor measure resemble the risk-neutral measure in many ways. For example, as we recall from Lemma 4.2.2, the risk-neutral measure is characterized by the fact that a continuously resettled futures rate is a martingale. In close parallel, the futures rate that is marked to market (resettled) discretely on dates $T_0, \ldots, T_N$ turns out to be a martingale in the spot Libor measure. We show this in Section 16.8.

4.2.4 Terminal and Hybrid Measures

One advantage of the spot measure over an arbitrary T-forward measure is the fact that the numeraire asset $B(t)$ will remain alive throughout the span of the tenor structure $\{T_n\}_{n=0}^N$. This property of $B(t)$ is necessary for the valuation of securities which may mature (randomly) at any date in the tenor structure. Securities of this type include for instance barrier options (such as range accruals) and options with early exercise rights (such as Bermudan swaptions). On the other hand, if we pick the T-forward measure corresponding to the *last* maturity in the tenor structure, $T = T_N$, this also yields a numeraire asset — the T_N-maturity zero-coupon bond — that is certain to remain alive at all dates in the tenor structure. The measure induced by $P(t, T_N)$ (Q^{T_N}) is often referred to as the *terminal measure*. For a security V maturing at a date $T \leq T_N$ we get, from the usual martingale pricing formula,

$$V(t) = P(t, T_N)\mathrm{E}_t^{T_N}\left(V(T)/P(T, T_N)\right), \quad t \leq T \leq T_N. \tag{4.25}$$

In (4.25) it is useful to notice that $V(T)/P(T, T_N)$ is the time T_N proceeds of rolling at time T the security payout $V(T)$ into a zero-coupon bond maturing at time T_N, effectively aligning the maturity of the numeraire and the cash flow date of the underlying asset. As an alternative, $V(T)$ could be rolled into the spot numeraire asset $B(T)$, leading to a T_N payout of $V(T)/B(T) \cdot B(T_N)$. This gives rise to the equivalent formula

$$V(t) = P(t, T_N)\mathrm{E}_t^{T_N}\left(V(T)B(T_N)/B(T)\right), \quad t \leq T \leq T_N. \tag{4.26}$$

We note that this formula can also be derived from the basic relationship between the measures Q^B and Q^{T_N} by simply noting that

$$P(T, T_N)\mathrm{E}_T^{T_N}\left(B(T_N)/B(T)\right) = B(T)\mathrm{E}_T^B\left(B(T_N)/B(T)/B(T_N)\right) = 1,$$

such that, by iterated conditional expectations[3],

[3] The *law of iterated conditional expectations*, sometimes known as the *tower rule*, states that for an $\mathcal{F}_T$-measurable random variable X, $E(E(X|\mathcal{F}_s)|\mathcal{F}_t) = E(X|\mathcal{F}_t)$, where $t \leq s \leq T$.

$$\begin{aligned} V(t) &= P(t,T_N)\mathrm{E}_t^{T_N}\left(V(T)/P(T,T_N)\right) \\ &= P(t,T_N)\mathrm{E}_t^{T_N}\left(V(T)\mathrm{E}_T^{T_N}\left(B(T_N)/B(T)\right)\right) \\ &= P(t,T_N)\mathrm{E}_t^{T_N}\left(V(T)B(T_N)/B(T)\right), \end{aligned}$$

as before.

As mentioned, equations (4.25) and (4.26) effectively involve reinvestment of the proceeds $V(T)$ to align cash payment with the numeraire $P(t,T_N)$. If the numeraire expires before the derivative security, we can apply the same reinvestment idea, but this time to the numeraire asset. Consider for instance a derivative security maturing at time T_N (paying $V(T_N)$), and suppose we wish to extend the T-forward measure to price this option. For instance, we can define a numeraire asset as follows:

$$\widetilde{P}(t,T) = \begin{cases} P(t,T), & t \le T, \\ B(t)/B(T), & t > T. \end{cases}$$

This asset corresponds to an investment strategy where we i) at time 0 purchase the T-maturity zero-coupon bond; and ii) at time T invest the proceeds from the zero-coupon bond (\$1) in the spot measure numeraire asset (4.24). Letting $\widetilde{\mathrm{Q}}^T$ denote the measure induced by $\widetilde{P}(t,T)$, we can write

$$\begin{aligned} V(t) &= \widetilde{P}(t,T)\widetilde{\mathrm{E}}_t^T\left(V(T_N)/\widetilde{P}(T_N,T)\right) \\ &= \widetilde{P}(t,T)\widetilde{\mathrm{E}}_t^T\left(V(T_N)B(T)/B(T_N)\right), \quad T < T_N,\ t < T_N, \end{aligned}$$

where $\widetilde{\mathrm{E}}^T$ is the expectations operator for the measure $\widetilde{\mathrm{Q}}^T$. If also $t \le T$, this expression becomes (compare to (4.26))

$$V(t) = P(t,T)\widetilde{\mathrm{E}}_t^T\left(V(T_N)B(T)/B(T_N)\right), \quad t \le T < T_N,$$

which, in effect, uses $B(T)/B(T_N)$ to discount $V(T_N)$ back to time T. The equivalent result in the T-forward measure is

$$V(t) = P(t,T)\mathrm{E}_t^T\left(B(T)\mathrm{E}_T^B\left(V(T_N)/B(T_N)\right)\right).$$

Notice that if V matures at time T, rather than at T_N, we simply have

$$V(t) = P(t,T)\widetilde{\mathrm{E}}_t^T\left(V(T)\right) = P(t,T)\mathrm{E}_t^T\left(V(T)\right),$$

which is obvious from the definition of the numeraire $\widetilde{P}(t,T)$.

The measure $\widetilde{\mathrm{Q}}^T$ is by construction a hybrid between the spot measure and the T-forward measure. Obviously, many other such measures exist, corresponding to different reinvestment strategies of expiring numeraire assets.

4.2.5 Swap Measures

Being a linear combination of zero-coupon bonds (see (4.8)), an annuity factor $A_{k,m}(t)$ on a tenor structure qualifies as a numeraire asset. The measure $\mathrm{Q}^{k,m}$ induced by this numeraire is known as a *swap measure* or an *annuity measure*. In the absence of arbitrage we have

$$V(t) = A_{k,m}(t)\mathrm{E}_t^{k,m}\left(V(T)/A_{k,m}(T)\right),$$

where $\mathrm{E}^{k,m}(\cdot)$ denotes expectation under $\mathrm{Q}^{k,m}$.

Lemma 4.2.4. *In the absence of arbitrage, the forward swap rate $S_{k,m}(t)$ is a martingale in measure $\mathrm{Q}^{k,m}$.*

Proof. By definition

$$S_{k,m}(t) = \frac{P(t,T_k) - P(t,T_{k+m})}{A_{k,m}(t)}.$$

As the numeraire deflated assets $P(t,T_k)/A_{k,m}(t)$ and $P(t,T_{k+m})/A_{k,m}(t)$ must both be martingales, so must be their difference. □

As we shall see later, swap measures are very useful for analytical manipulations of price formulas for options on swaps.

4.3 Multi-Currency Markets

While this book is primarily dedicated to the study of single-currency interest rate derivatives, occasionally it will be necessary to consider certain effects associated with trading in a multi-currency economy. For instance, in Chapter 6 we touch upon issues of yield curve constructions in non-domestic currencies, and in Chapter 16 we discuss the important practical case where a derivative pays out in a foreign currency, but has a payout function that depends on one or more domestic interest rate variables. This brief section provides background material and notation required for these and other cross-currency applications.

4.3.1 Notations and FX Forwards

We consider two economies, a "domestic" economy and a "foreign" economy. Let $P_d(t,T)$ and $P_f(t,T)$ denote time t zero-coupon bond prices in the domestic and foreign economies, respectively. So, $P_f(t,T)$ (say) is the time t price, in foreign currency, of one unit of foreign currency delivered for certain at time T. Translation of values in foreign currency to domestic currency takes place at a foreign exchange (FX) rate of $X(t)$, measured in units of domestic currency per unit of foreign currency. In other words, the value $\widetilde{P}_d$ to a domestic investor of one foreign zero-coupon bond is

$$\widetilde{P}_d(t,T) = X(t)P_f(t,T).$$

The quantity

$$X_T(t) = \frac{\widetilde{P}_d(t,T)}{P_d(t,T)} = X(t)\frac{P_f(t,T)}{P_d(t,T)}$$

is known as the *forward FX rate* to time T. The name is motivated by the following arbitrage strategy:

- Buy one foreign zero-coupon bond, at a cost of $\widetilde{P}_d(t,T)$ in domestic currency.
- Finance the purchase by selling short domestic zero-coupon bonds on a notional of $\widetilde{P}_d(t,T)/P_d(t,T)$.

With no outlay at time t, the strategy will generate a net cash flow at time T of one unit of foreign currency and $-X_T(t)$ units of domestic currency, such that the trading strategy in effect has locked in a time t a future time T exchange rate of $X_T(t)$.

4.3.2 Risk Neutral Measures

Let $\beta_d(t)$ and $\beta_f(t)$ be the continuously compounded money market accounts in the domestic and foreign economies, respectively. $\beta_d(t)$ and $\beta_f(t)$ induce two separate risk-neutral measures, denoted Q^d and Q^f; let us investigate how these measures are related. If $g(T)$ is a random payout at time T made in foreign currency, in a complete market the value (in units of foreign currency) of this payout to a foreign investor is, from standard principles,

$$V_f(t) = \beta_f(t)\mathrm{E}_t^f\left(g(T)\beta_f(T)^{-1}\right), \tag{4.27}$$

where E_t^f denotes expectations in the foreign risk-neutral measure Q^f. For a domestic investor, the payout of $g(T)$ must be translated to domestic currency units at a rate of $X(T)$, making the effective domestic payout function $g(T)X(T)$. Thereby,

$$V_d(t) = \beta_d(t)\mathrm{E}_t^d\left(g(T)X(T)\beta_d(T)^{-1}\right), \tag{4.28}$$

where E_t^d denotes expectations in measure Q^d. Importantly, the expressions in (4.27) and (4.28) are linked by the spot exchange rate, as the absence of a cross-currency arbitrage dictates that

$$V_d(t) = X(t)V_f(t),$$

or

$$\beta_d(t)\mathrm{E}_t^d\left(g(T)X(T)\beta_d(T)^{-1}\right) = X(t)\beta_f(t)\mathrm{E}_t^f\left(g(T)\beta_f(T)^{-1}\right). \tag{4.29}$$

We use this result to establish the following lemma:

Lemma 4.3.1. *The domestic and foreign risk-neutral probability measures* Q^d *and* Q^f *are related by the density process*

$$\mathrm{E}_t^d\left(\frac{d\mathrm{Q}^f}{d\mathrm{Q}^d}\right) = \frac{\beta_f(t)X(t)}{\beta_d(t)X(0)}, \quad t \geq 0.$$

Proof. For an $\mathcal{F}_T$-measurable variable $Y(T) = g(T)X(T)\beta_d(T)^{-1}$ satisfying regularity conditions, a rearrangement of the basic relation (4.29) yields

$$\mathrm{E}_t^d\left(Y(T)\right) = X(t)\frac{\beta_f(t)}{\beta_d(t)}\mathrm{E}_t^f\left(\frac{Y(T)}{X(T)}\frac{\beta_d(T)}{\beta_f(T)}\right),$$

From the results of Section 1.3, the density relating measures Q^d and Q^f is then as given in the lemma. □

With $\beta_f(t)X(t)/\beta_d(t)$ being a martingale in the domestic risk-neutral measure, we note that if $X(t)$ is an Ito process, it must take the form

$$dX(t) = X(t)\left(r_d(t) - r_f(t)\right)dt + \sigma_X(t)^\top dW(t),$$

where $r_d(t) - r_f(t)$ is the spread between domestic and foreign short rates, $W(t)$ is a (vector-valued) Q^d-Brownian motion, and $\sigma_X(t)$ is some adapted stochastic process satisfying regularity conditions.

4.3.3 Other Measures

Having established the Radon-Nikodym derivative relating the domestic and foreign risk-neutral measures, relations between various other domestic and foreign probability measures are easy to establish. For instance, the following result is easily proven the same way as Lemma 4.3.1.

Lemma 4.3.2. *Let* $\mathrm{E}_t^{T,d}$ *denote expectations in the domestic* T*-forward probability measure. The domestic and foreign* T*-forward probability measures* $\mathrm{Q}^{T,d}$ *and* $\mathrm{Q}^{T,f}$ *are related by the density process*

$$\mathrm{E}_t^{T,d}\left(\frac{d\mathrm{Q}^{T,f}}{d\mathrm{Q}^{T,d}}\right) = \frac{P_f(t,T)P_d(0,T)X(t)}{P_d(t,T)P_f(0,T)X(0)} = \frac{X_T(t)}{X_T(0)}, \quad t \geq 0.$$

We highlight the fact that the forward FX rate $X_T(t)$ is a $\mathrm{Q}^{T,d}$-martingale satisfying $X_T(T) = X(T)$. For an $\mathcal{F}_T$-measurable variable $Y(T)$, we thereby have the convenient expression

$$\mathrm{E}_t^{T,d}\left(Y(T)\right) = X_T(t)\mathrm{E}_t^{T,f}\left(\frac{Y(T)}{X(T)}\right),$$

where $\mathrm{E}_t^{T,f}$ denotes expectations in the foreign T-forward probability measure.

4.4 The HJM Analysis

Having defined notations and established basic arbitrage relationships, let us turn to assigning dynamics to the many quantities we have introduced so far. We shall here follow the Heath et al. [1992] (Heath-Jarrow-Morton, or HJM) approach, where all information in the economy is assumed to originate with a finite number of Brownian motions. The resulting class of models is quite broad, and in much of the rest of this book we shall deal with ways to reduce the general HJM model to specific, and tractable, special cases. For now, however, we concentrate on a general analysis, although we keep our treatment fairly informal.

4.4.1 Bond Price Dynamics

In the HJM framework, we concern ourselves with the modeling of how an entire continuum of T-indexed bond prices $P(\cdot, T)$ jointly evolves our time, starting from a known condition $P(0, T)$. We consider models of finite horizon, i.e. with $T \in [0, \mathcal{T}]$, $\mathcal{T} < \infty$, and specialize to a filtration generated by a d-dimensional Brownian motion. We assume that a risk-neutral measure Q exists and is unique. Let $W(t)$ be an adapted d-dimensional Q-Brownian motion, and define deflated bond values as $P_\beta(t, T) = P(t, T)/\beta(t)$, where $\beta(t)$ as always is the continuously rolled money market account. In the absence of arbitrage, $P_\beta(t, T)$ is a martingale in the risk-neutral measure, and the martingale representation theorem then implies that

$$dP_\beta(t, T) = -P_\beta(t, T)\sigma_P(t, T)^\top \, dW(t), \quad t \leq T, \tag{4.30}$$

where $\sigma_P(t, T) = \sigma_P(t, T, \omega)$ is a d-dimensional stochastic process adapted to the filtration generated by W. We assume that $\sigma_P(t, T)$ is regular enough for $P_\beta(t, T)$ to be a square-integrable martingale. Also, as the bond $P(t, T)$ must equal \$1 at $t = T$ ("pull to par"), we impose the consistency condition

$$\sigma_P(T, T) = 0.$$

Using (4.12) and Ito's lemma, it follows from (4.30) that

$$dP(t, T)/P(t, T) = r(t)\, dt - \sigma_P(t, T)^\top dW(t), \tag{4.31}$$

where $r(t)$ is the short rate process. Equation (4.31) defines the class of d-dimensional HJM models.

Another application of Ito's lemma shows that forward bond prices $P(t, T, T + \tau) = P(t, T + \tau)/P(t, T)$ must satisfy

$$\begin{aligned} dP(t, T, T + \tau)/P(t, T, T + \tau) = &- [\sigma_P(t, T + \tau) - \sigma_P(t, T)]^\top \sigma_P(t, T)\, dt \\ &- [\sigma_P(t, T + \tau) - \sigma_P(t, T)]^\top dW(t). \end{aligned} \tag{4.32}$$

In the T-forward measure Q^T, $P(t,T,T+\tau)$ is a martingale (see Section 4.2.2), and

$$dP(t,T,T+\tau)/P(t,T,T+\tau) = -\left[\sigma_P(t,T+\tau) - \sigma_P(t,T)\right]^\top dW^T(t), \tag{4.33}$$

where $W^T(t)$ is a Q^T-Brownian motion. Comparison of (4.32) and (4.33) shows that

$$dW^T(t) = dW(t) + \sigma_P(t,T)\,dt \tag{4.34}$$

which by Girsanov's theorem identifies the density process for the measure shift between Q^T and Q in the HJM setting:

$$\varsigma(t) = \mathrm{E}_t^{\mathrm{Q}}\left(\frac{d\mathrm{Q}^T}{d\mathrm{Q}}\right) = \mathcal{E}\left(-\int_0^t \sigma_P(u,T)^\top dW(u)\right), \tag{4.35}$$

or

$$d\varsigma(t)/\varsigma(t) = -\sigma_P(t,T)^\top dW(t).$$

This result could, of course, have been established from the first principles as well — see equation (4.21).

4.4.2 Forward Rate Dynamics

Traditionally, HJM models are stated in terms of instantaneous forward rates, rather than bond prices. Besides eliminating the need to consider the short rate r, this also reveals a number of fundamental properties of the class of HJM models. By Ito's lemma, in measure Q,

$$d\ln P(t,T) = O(dt) - \sigma_P(t,T)^\top dW(t),$$

where for convenience we have omitted writing out the drift term. Differentiating the right- and left-hand sides of this equation with respect to T, we get from equation (4.4),

$$df(t,T) = \mu_f(t,T)\,dt + \sigma_f(t,T)^\top dW(t),$$

where

$$\sigma_f(t,T) = \frac{\partial}{\partial T}\sigma_P(t,T), \tag{4.36}$$

and $\mu_f(t,T)$ is listed below.

Lemma 4.4.1. *The process for $f(t,T)$ in the T-forward measure is*

$$df(t,T) = \sigma_f(t,T)^\top dW^T(t). \tag{4.37}$$

In the risk-neutral measure, the process is

$$\begin{aligned} df(t,T) &= \sigma_f(t,T)^\top \sigma_P(t,T)\,dt + \sigma_f(t,T)^\top dW(t) \\ &= \sigma_f(t,T)^\top \int_t^T \sigma_f(t,u)\,du\,dt + \sigma_f(t,T)^\top dW(t). \end{aligned} \tag{4.38}$$

Proof. The SDE (4.37) follows directly from the martingale relation (4.23). The risk-neutral process (4.38) then can be derived from the relations (4.34) and (4.35), with the second equality following from (4.36). □

The equation (4.38) is often considered to be the main result of Heath et al. [1992]. It demonstrates that an HJM model is fully specified once the forward rate diffusion coefficients $\sigma_f(t,T)$ have been specified for all t and T. Note that HJM models take initial forward rates $f(0,T)$ as exogenous inputs, ensuring that these models are automatically consistent with discount bond prices at time 0. This is true irrespective of the choice of $\sigma_f(t,T)$, which can be set freely (subject to regularity conditions) from either empirical analysis, or from a calibration to market prices of fixed income derivatives.

While it is convenient that HJM models are automatically calibrated to initial bond prices, a number of other features of the general HJM model are less attractive. Particularly problematic is the sheer dimensionality of the model: to describe the time t state of a discount bond curve spanning $[t,T]$, we need to keep track of a continuum of forward rates $\{f(t,u),\ t \le u \le T\}$. By Lemma 4.4.1 the forward rate curve follows an infinite-dimensional diffusion process, leaving us with an infinite number of state variables to diffuse. In practice, the implementation of an HJM model will require either making special assumptions about the σ_f process that permit a finite-dimensional Markovian representation of the forward rate curve; or moving from infinitesimal forward rates to continuously compounded forward rates that span time-buckets of finite length. Chapters 10 through 13 and Section 4.5.2 below give examples of the former idea, and Andersen [1995] discusses the latter approach in a Monte Carlo setting. An idea closely related to the discussion in Andersen [1995] is to built a model around a finite set of simple (Libor) forward rates on a fixed tenor structure. This approach has a number of computational and theoretical advantages, and is the subject of Chapter 14. For now, we note that *any* arbitrage-free interest rate model set in a filtration generated exclusively by Brownian motions must be a special case of an HJM model. In particular, any such model must correspond to a particular choice of $\sigma_f(t,T)$.

4.4.3 Short Rate Process

As discussed earlier, specification of a short rate process is, in principle, sufficient to completely specify a full yield curve model. In the HJM framework, it follows from (4.38) that the short rate $r(t)$ in measure Q is

$$r(t) = f(t,t) = f(0,t) + \int_0^t \sigma_f(u,t)^\top \int_u^t \sigma_f(u,s)\,ds\,du + \int_0^t \sigma_f(u,t)^\top dW(u).$$

The process for $r(t)$ is generally not Markovian, as can be seen by focusing on the path-dependent term

$$D(t) = \int_0^t \sigma_f(u,t)^\top dW(u)$$

for which we must have

$$D(T) = D(t) + \int_t^T \sigma_f(u,T)^\top dW(u) + \left\{ \int_0^t \sigma_f(u,T)^\top dW(u) - \int_0^t \sigma_f(u,t)^\top dW(u) \right\}. \quad (4.39)$$

Thereby

$$\mathrm{E}^Q\left(D(T)|\, D(t)\right) \neq \mathrm{E}_t^Q\left(D(T)\right)$$

unless the bracketed term in (4.39) is either non-random, or a deterministic function of $D(t)$ (which is generally not the case).

An interesting area of investigation concerns the conditions under which either $r(t)$ is outright Markov[4] or, less restrictively, can be written as

$$r(t) = h\left(t, x(t)\right),$$

for a deterministic function h and a finite-dimensional Markovian vector of state variables $x(t)$. Definitive results are given in Björk [2001], building on earlier (and considerably less abstract) work by Jamshidian [1991b], Cheyette [1991], and Ritchken and Sankarasubramanian [1995]. Section 4.5.2 and Chapter 13 list some of the results of these papers.

4.5 Examples of HJM Models

4.5.1 The Gaussian Model

In the HJM bond price dynamics (4.31), we now assume that $\sigma_P(t,T)$ is a bounded (d-dimensional) deterministic function of t and T. It follows from (4.32) and (4.33) that forward bond prices are then log-normally distributed in both Q and Q^T. The forward rate process in Q is

$$df(t,T) = \sigma_f(t,T)^\top \sigma_P(t,T)\, dt + \sigma_f(t,T)^\top dW(t), \quad \sigma_f(t,T) = \frac{\partial}{\partial T}\sigma_P(t,T), \quad (4.40)$$

which implies that $r(T) = f(T,T)$ is Gaussian with Q-moments

$$\mathrm{E}_t^{\mathrm{Q}}\left(f(T,T)\right) = \int_t^T \sigma_f(u,T)^\top \sigma_P(u,T)\, du,$$

$$\mathrm{Var}_t^{\mathrm{Q}}\left(f(T,T)\right) = \int_t^T \sigma_f(u,T)^\top \sigma_f(u,T)\, du.$$

[4] In the sense that the time t expectations of functionals of $r(T)$ only require knowledge of $r(t)$ itself. In this case the process for $r(t)$ must be a diffusion characterized by an SDE $dr(t) = \mu_r(t, r(t))dt + \sigma_r(t, r(t))^\top dW(t)$.

The simple form of the Gaussian HJM model makes it quite tractable, permitting analytical price formulas for a number of European options and futures contracts[5]. While the Gaussian HJM model suffers from the drawback of allowing negative forward and spot rates, analytical results derived in the model are often very useful in gaining a deeper understanding of a given contract, even if ultimately a more realistic model will be required for serious pricing purposes. Indeed, results derived for the Gaussian HJM model can often be used as a starting point for development of closed-form approximations in other models; we shall see many examples of this later in the book.

For illustration, we list a few select analytical results below. More formulas can be found in numerous sources, including Chapter II in Andersen [1996], Jamshidian [1991b], and Jamshidian [1993], to name a few.

Proposition 4.5.1 (Option on Zero-Coupon Bond). *Consider a European call option paying at maturity T the amount*

$$V(T) = (P(T,T^*) - K)^+, \quad T^* > T.$$

In the Gaussian HJM model (4.40), we have

$$V(t) = P(t,T^*)\Phi(d_+) - P(t,T)K\Phi(d_-), \tag{4.41}$$

where

$$d_\pm = \frac{\ln\left(P(t,T^*)/\left(KP\left(t,T\right)\right)\right) \pm v/2}{\sqrt{v}},$$

$$v = \int_t^T \left|\sigma_P(u,T^*) - \sigma_P(u,T)\right|^2 du.$$

Proof. In the T-forward measure Q^T we have, from (4.20),

$$\begin{aligned} V(t) &= P(t,T)\mathrm{E}_t^T\left((P(T,T^*) - K)^+\right) \\ &= P(t,T)\mathrm{E}_t^T\left((P(T,T,T^*) - K)^+\right). \end{aligned}$$

From the discussion in Section 4.4.1 we know that $P(t,T,T^*)$ is a Q^T-martingale characterized by the SDE

$$dP(t,T,T^*)/P(t,T,T^*) = -\left[\sigma_P(t,T^*) - \sigma_P(t,T)\right]^\top dW^T(t),$$

where W^T is a d-dimensional Q^T-Brownian motion. As this is just a GBM process with time-dependent coefficients, the Black-Scholes-Merton results in Section 1.9 apply and lead to (4.41). □

[5]Indeed, we have already used this model in an equity context — see Section 1.9.3.2.

Proposition 4.5.2 (Caplet). *Consider a European call option paying at $T+\tau$ the amount (a* caplet*)*

$$V(T+\tau) = \tau\left(L(T,T,T+\tau) - K\right)^+, \quad \tau > 0.$$

In the Gaussian HJM model (4.40), we have

$$V(t) = \tau P(t,T+\tau)\left(L(t,T,T+\tau)\Phi(d_+) - K\Phi(d_-)\right),$$

$$d_\pm = \frac{\ln\left(L(t,T,T+\tau)/K\right) \pm v/2}{\sqrt{v}}, \; v = \int_t^T \left|\sigma_P(u,T+\tau) - \sigma_P(u,T)\right|^2 du.$$

Proof. From Lemma 4.2.3 we know that $L(t,T,T+\tau)$ is a martingale in the $(T+\tau)$-forward measure. An application of Ito's lemma to the definition (4.2) reveals that

$$dL(t,T,T+\tau)/L(t,T,T+\tau) = \left[\sigma_P(t,T+\tau) - \sigma_P(t,T)\right]^\top dW^T(t),$$

and the result follows immediately along the same lines as in the proof of Proposition 4.5.1. □

Proposition 4.5.3 (Futures Rate). *In the Gaussian HJM model (4.40), futures rates are given by*

$$F(t,T,T+\tau) = \tau^{-1}\left(\left(1/P(t,T,T+\tau)\right)e^{\Omega(t,T)} - 1\right), \tag{4.42}$$

where

$$\Omega(t,T) = \int_t^T \left[\sigma_P(u,T+\tau) - \sigma_P(u,T)\right]^\top \sigma_P(u,T+\tau)\, du.$$

Proof. From Lemma 4.2.2,

$$\begin{aligned} F(t,T,T+\tau) &= \mathrm{E}_t^{\mathrm{Q}}\left(L(T,T,T+\tau)\right) \\ &= \tau^{-1}\mathrm{E}_t^{\mathrm{Q}}\left(1/P(T,T+\tau) - 1\right) \\ &= \tau^{-1}\mathrm{E}_t^{\mathrm{Q}}\left(G(T) - 1\right), \end{aligned} \tag{4.43}$$

where we have introduced an auxiliary variable

$$G(t) \triangleq P(t,T)/P(t,T+\tau) = 1/P(t,T,T+\tau).$$

Ito's lemma shows that (see also (4.32)) in measure Q

$$\begin{aligned} dG(t)/G(t) = &\left[\sigma_P(t,T+\tau) - \sigma_P(t,T)\right]^\top \sigma_P(t,T+\tau)\, dt \\ &+ \left[\sigma_P(t,T+\tau) - \sigma_P(t,T)\right]^\top dW(t), \end{aligned}$$

such that

$$\mathrm{E}_t^{\mathrm{Q}}\left(G(T)\right) = G(t)e^{\Omega(t,T)} = \left(1/P(t,T,T+\tau)\right)e^{\Omega(t,T)},$$

where Ω is as given above. The result of Proposition 4.5.3 then follows directly from (4.43). □

In any rational model $\Omega(t,T) \geq 0$, such that $F(t,T,T+\tau) \geq L(t,T,T+\tau)$, consistent with the qualitative discussion in Section 4.1.2. As shown in Chapter II of Andersen [1996], the spread (also known as futures *convexity*) between futures and forward rates can be decomposed into two components: i) a term originating from the mark-to-market mechanism of a futures contract; and ii) a term originating from the fact that a futures contract — unlike a regular forward rate agreement — pays out the rate at the date it settles (at time T) rather than one period ahead (at time $T+\tau$). Andersen [1996], Chapter II, additionally contains a number of numerical examples examining typical futures-forward spreads, and also investigates the pricing of options on futures rates.

Section 16.8 looks in detail into pricing interest rate futures under more advanced models.

4.5.2 Gaussian HJM Models with Markovian Short Rate

Although quite tractable, the Gaussian HJM model generally does not allow for a finite-dimensional Markovian representation, and typically does not imply Markov-diffusive behavior of the short rate. As shown in Carverhill [1994], the short rate can be made Markovian, however, by imposing certain conditions on the deterministic forward rate volatility function $\sigma_f(t,T)$. To explore this, first recall from Section 4.4.3 the relation

$$r(t) = f(0,t) + \int_0^t \sigma_f(u,t)^\top \int_u^t \sigma_f(u,s)\,ds\,du + \int_0^t \sigma_f(u,t)^\top dW(u),$$

where now σ_f is deterministic. Consider imposing the special choice

$$\sigma_f(t,T) = g(t)h(T), \tag{4.44}$$

where h is a positive real function and $g : \mathbb{R} \to \mathbb{R}^{d\times 1}$ can take any sign. For this case we have

$$\sigma_P(t,T) = \int_t^T \sigma_f(t,u)\,du = g(t)\int_t^T h(u)\,du,$$

and

$$\begin{aligned} r(t) &= f(0,t) + h(t)\int_0^t g(u)^\top g(u)\left(\int_u^t h(s)\,ds\right)du + h(t)\int_0^t g(u)^\top dW(u) \\ &\triangleq f(0,t) + h(t)\int_0^t m_f(t,u)\,du + h(t)\int_0^t g(u)^\top dW(u). \end{aligned} \tag{4.45}$$

Importantly, the term

$$D(t) = \int_0^t \sigma_f(u,t)^\top \, dW(u) = h(t) \int_0^t g(u)^\top \, dW(u)$$

is now Markov, since

$$D(T) = h(T) \int_0^T g(u)^\top \, dW(u) = \frac{h(T)}{h(t)} D(t) + h(T) \int_t^T g(u)^\top \, dW(u),$$

which should be compared to the general (non-Markov) expression (4.39).

To show that the short rate is Markovian, we differentiate (4.45) with respect to t, yielding

$$\begin{aligned}
dr(t) &= \frac{\partial f(0,t)}{\partial t} dt + h'(t) \left(\int_0^t m_f(t,u) \, du + \frac{D(t)}{h(t)} \right) dt \\
&\quad + h(t) \frac{\partial}{\partial t} \int_0^t m_f(t,u) \, du \, dt + h(t) g(t)^\top \, dW(t) \\
&= \frac{\partial f(0,t)}{\partial t} dt + h'(t) \left(\frac{r(t) - f(0,t)}{h(t)} \right) dt \\
&\quad + h(t)^2 \int_0^t g(u)^\top g(u) \, du \, dt + h(t) g(t)^\top \, dW(t) \\
&= \left(\frac{\partial f(0,t)}{\partial t} - \frac{h'(t)}{h(t)} f(0,t) + h(t)^2 \int_0^t g(u)^\top g(u) \, du + \frac{h'(t)}{h(t)} r(t) \right) dt \\
&\quad + h(t) g(t)^\top \, dW(t),
\end{aligned} \tag{4.46}$$

where the second equality follows from rearrangement of (4.45). This leads to the following result.

Proposition 4.5.4. *In the d-dimensional Gaussian HJM model, when (4.44) holds the short rate satisfies an SDE of the type*

$$dr(t) = (a(t) - \varkappa(t) r(t)) \, dt + \sigma_r(t)^\top \, dW(t),$$

where $\varkappa : \mathbb{R} \to \mathbb{R}$ *and* $\sigma_r : \mathbb{R} \to \mathbb{R}^{d \times 1}$ *are deterministic functions of time, and*

$$\begin{aligned}
a(t) &= \frac{\partial f(0,t)}{\partial t} + \varkappa(t) f(0,t) + \int_0^t e^{-2 \int_u^t \varkappa(s) \, ds} \sigma_r(u)^\top \sigma_r(u) \, du \\
&= \frac{\partial f(0,t)}{\partial t} + \varkappa(t) f(0,t) + \int_0^t \sigma_f(u,t)^\top \sigma_f(u,t) \, du.
\end{aligned}$$

Proof. First, by way of defining $\varkappa$ and σ_r, we set

$$h(T) = e^{-\int_0^T \varkappa(s) \, ds}; \quad g(t) = e^{\int_0^t \varkappa(s) \, ds} \sigma_r(t),$$

such that $h'(t)/h(t) = -\varkappa(t)$ and

$$\sigma_f(t,T) = e^{-\int_t^T \varkappa(s)\,ds}\sigma_r(t).$$

The result of Proposition 4.5.4 then follows directly by insertion into (4.46). □

4.5.3 Log-Normal HJM Models

To avoid the negative forward rates inherent in Gaussian HJM models, it is tempting to consider forward rate specifications of the type

$$\sigma_f(t,T) = f(t,T)\sigma(t,T), \tag{4.47}$$

where $\sigma(t,T)$ is deterministic and bounded. In the T-forward measure

$$df(t,T) = f(t,T)\sigma(t,T)^\top dW^T(t)$$

such that $f(t,T)$ is log-normally distributed. While avoiding negative rates, the specification (4.47) has severe technical problems: in Q, forward rates will explode to infinity with non-zero probability. Attempts to apply the valuation formula (4.15) will thereby result in all zero-coupon bond prices being zero, implying obvious arbitrage opportunities. To suggest a rationale for the exploding rates, consider the Q-dynamics

$$df(t,T) = \left(f(t,T)\sigma(t,T)^\top \int_t^T f(t,u)\sigma(t,u)du\right)dt + f(t,T)\sigma(t,T)^\top dW(t).$$

Loosely speaking, the drift-term is proportional to forward rates *squared*, which, in the light of the linear growth condition in Theorem 1.6.1, may cause us to suspect problems with the existence of a non-exploding solution. Morton [1988] confirms this rigorously.

One solution to the explosion problem involves enforcing a strict upper bound on $\sigma_f(t,T)$, as in

$$\sigma_f(t,T) = \min\left(f(t,T), M\right)\sigma(t,T),$$

where M is a large positive constant. For the one-factor case ($d = 1$) Heath et al. [1992] demonstrate that this specification will ensure non-negative forward rates[6] and will prevent rate explosions. Nevertheless, the model is clearly awkward in its dependence on the arbitrary constant M. A more satisfying solution is discussed in Chapter 14, where we show that the explosion problem can be circumvented by working with simply — rather than continuously — compounded forward rates. A related issue in short rate models is also discussed in Chapter 11.

[6] To see this, notice that $M > 0$ guarantees that $df(t,T) = 0$ if $f(t,T)$ should ever reach 0.

5

Fixed Income Instruments

At this point, we have established the mathematical and numerical prerequisites needed for the remaining part of the book, much of which is devoted to the development of models for fixed income derivatives. Before delving into the modeling exercise, this final foundational chapter provides a tour of actual fixed income markets as well as an overview of the types of products traded. The simpler (and more liquid) of these products will typically serve as calibration targets to parameterize the models we develop; others (the more complicated and illiquid ones) will constitute the contracts that our model are ultimately meant to price and hedge. Throughout the chapter — and, indeed, this book — our focus is on the securities tied to the so-called *Libor rate*; this will include essentially all high-end exotic securities as well as more basic instruments such as FRAs, caps, and swaptions. Our priorities dictate that we leave out government, corporate, and mortgage bonds, as well as the derivatives associated with these types of securities. A discussion of these classes of securities, along with many more details on the organization and workings of fixed income markets, can be found in specialist literature, such as Fabozzi and Modigliani [1996], Fabozzi [1985], Fabozzi [2001], and Fabozzi and Fabozzi [1989].

5.1 Fixed Income Markets and Participants

At the most fundamental level, interest rates determine the economic cost of borrowing and lending, and as such define present values of future cash flows. In general, cash flows occurring at different times are discounted at different rates, reflecting market fluctuations in demand for money and risk preferences of market participants. The dependence of interest rates on time is described by the so-called *term structure of interest rates*, easily visualized as a curve that assigns a particular interest rate (or, equivalently, a discount factor) to each future date.

For a given entity, the cost of borrowing money will depend on its credit quality. Governments of developed countries, perceived to have virtually no possibility of default, issue bonds at comparatively low interest rates that reflect this perception. While the market in government debt is vast, corporations typically find it more convenient to use and originate fixed income instruments linked to rates that are more reflective of their own financing costs (i.e., credit quality). By far the most common of such reference rates is the London Interbank Offered rate, commonly known as the *Libor rate*. The Libor rate is a filtered average of bank estimates of rates at which they can borrow for a given term in the *interbank money market*, i.e. the wholesale market in which banks provide unsecured short-term credit to each other. Libor rates are quoted for multiple deposit maturities ranging from one day to one year, and are set every business day by averaging polling results from a number of large banks. Libor rates are available for deposits in different currencies, so that there is a USD-Libor rate, a EUR-Libor rate, and so on.

While Libor rates are probably the most used reference rates for interest rate contracts, there are other important rates to be aware of. For example, in the United States, banks are required to hold certain balances ("Federal funds") with the Federal Reserve, the central bank of the US. If a bank does not have sufficient balances, it can borrow them from another bank that has an excess on its account. The overnight interest rate charged in this case is called the *(effective) Federal funds rate*[1], or sometimes simply the Fed funds rate. This rate is often considered the best available proxy for a risk-free USD rate, in part because the Fed funds rate is normally the contractual rate used to accrue interest on posted collateral[2], as explained in Piterbarg [2010]. It is worth noting that the Fed funds rate used to be closely linked to the overnight Libor rate, with the spread between the two in the single basis points. However, in the subprime crisis of 2007–2009 the two have diverged significantly; the implications of this for interest rate curve construction are discussed in Section 6.5.3. Instruments linked to averages of the (effective) Fed fund rate over different terms are actively traded, giving rise to a term structure of Fed funds linked rates.

A special feature of the US public debt markets gives rise to another set of rates. In particular, interest on bonds issued by states and other local governments of the US is often free of the federal tax. The Bond Market Association, a trade association of the bond industry, publishes the *BMA rate* (or *BMA index*) which is the estimate of borrowing by such municipalities.

[1] The target rate, set by the Federal reserve, is aptly called the *target* Fed funds rate.

[2] To mitigate credit risk, many derivatives transactions require posting of collateral (normally cash or Treasury bonds) in the amount of the current mark-to-market. ISDA [2005] contains a detailed description of collateral agreements; according to ISDA [2009], in 2009 about 65% of all OTC derivatives transactions involved such agreements.

There is a well-developed market in interest rate derivatives that are linked to the BMA rate.

The Euro and GBP markets do not have the same mechanism as the US does for Federal funds, but overnight rates that are proxies for risk-free borrowing in those currencies do exist. They are called *Eonia* (Euro OverNight Index Average) in the Eurozone and *Sonia* (Sterling OverNight Index Average) in Great Britain, and are computed as averages of all *actual* overnight lending/borrowing transactions by qualifying banks weighted by the size of the transactions. We emphasize that these rates reflect the actual transactions that have happened, in contrast to Libor which reflect banks' estimates of rates at which borrowing (for a given term) might take place. In the crisis of 2007–2009 there have been serious concerns about the integrity of the Libor rate and whether it really reflected the actual cost of funding for banks, and even some calls to scrap the Libor rate altogether. While the Libor rate has survived the crisis, the importance of overnight rates has increased dramatically, with the market in FedFunds/Eonia/Sonia linked derivatives, most importantly in *overnight index swaps*, or OIS, of various maturities growing dramatically. As with the Fed funds rate, Eonia and Sonia have diverged significantly from the corresponding Libor rates during market turbulence, and the decoupling continues to persist. As with the Fed funds rate, the implications of these developments on interest rate curve construction are discussed in Section 6.5.3.

Interest rates change day-to-day in response to changing macroeconomic and market conditions. With the cost of borrowing and lending money affecting all aspects of the economy, it is no surprise that a vast market in derivatives on interest rates has developed. Motivations of participants are diverse, ranging from locking in the cost of financing to pure speculation.

The fixed income market can be broadly split into two (overlapping) segments: the *exchange* market and the *over-the-counter*, or *OTC*, market. Contracts linked to the level of interest rates are traded on many securities exchanges. The exchanges attract all types of investors, including market makers, hedgers and speculators; see Hull [2006] for details on all. As of March 2008, notional amounts outstanding were $26 trillion in exchange traded interest rate futures, and $45 trillion in exchange traded interest rate options. While these are impressive numbers, far more fixed income derivatives trade in OTC markets than in exchange markets: as of December 2007, the notional amounts outstanding of OTC interest rate derivatives amounted to $393 trillion[3]. The OTC market can loosely be visualized as a network of banks that trade with each other under terms governed by agreements spelled out by the trade organization International Swaps and Derivatives Association (ISDA). Central to OTC markets are the *interest*

[3]All figures from the report "Semiannual OTC derivatives statistics at end-December 2007" by Bank for International Settlements, available from `www.bis.org`.

rate dealers, banks with trading desks specializing in fixed income trading. The dealers provide liquidity in various types of securities, and are typically the most sophisticated players in the market. The dealers trade either on their own account or on behalf of customers such as *financial institutions* and *corporates*.

Financial institutions include mortgage companies (organizations that originate, package or service residential and commercial mortgage loans), pension funds, mutual funds, insurance companies, hedge funds, and other entities whose primary activities are related to financial markets. Financial institutions seek to either make money directly by engaging in trading activities (hedge funds), or to hedge their exposures (mortgage originators or servicers), or to achieve superior returns on their investments (pension funds, insurance companies). Among financial institutions, an important role is played by *issuers*, companies that issue structured notes for private and public placement. Structured notes deliver appealing return profiles to investors, returns that are essentially financed by selling options back to issuers. Issuance of increasingly complicated structured notes drives the exotic end of the fixed income markets.

Corporates are companies with primary activities not directly linked to fixed income markets, but whose operational results may be affected by the interest rate environment. For instance, many companies raise funds by borrowing from banks or by selling bonds, and are therefore affected by the prevailing levels of interest rates. Corporates often seek to lock in favorable interest rates for borrowing money, to hedge their interest rate exposures, to transform their liabilities from one type (e.g., a fixed rate liability) to another (a floating rate liability), or to design custom borrowing schemes around their expected future borrowing needs.

5.2 Certificates of Deposit and Libor Rates

Having identified the main types of market participants, we now proceed to define the universe of securities that this book will cover. For technical precision, we shall occasionally need to refer to the risk-neutral measure Q, as well as its associated expectation operator $\mathrm{E} = \mathrm{E}^{\mathrm{Q}}$ and its numeraire $\beta(t)$.

We start with the *certificate of deposit* (or *CD*), a deposit of money for a pre-specified term at a pre-specified interest rate. Terms may range from one week to one year or more, with the most popular being a 3 month or a 6 month term, depending on the currency of the deposit. If 1 (dollar) is deposited at time T for a period of τ years, then the amount of capital to be returned at time $T + \tau$ is given by[4]

[4] As was mentioned earlier, the computation of τ from given start- and end-dates will involve certain formal day counting rules, see Appendix 5.A.

$$1 + \tau L,$$

where L is, by definition, the interest rate for the CD. The rate is quoted as a simple rate, i.e. a rate with the compounding frequency equal to the term of the deposit. Notice that the average value of L for CDs quoted in the interbank market will, by definition, be equal to the (spot) Libor rate for tenor τ. Spot Libor rates for various tenors are calculated daily and are published by major news services such as Bloomberg or Reuters. As mentioned above, Libor serves as the primary reference rate in fixed income markets.

If $P(T, T+\tau)$ is the (Libor-based) discount factor to date $T+\tau$ as observed at T, then the discounted value of receiving $1+\tau L$ at time $T+\tau$ should be equal to 1 at time T, i.e.

$$1 = P(T, T+\tau)(1+\tau L).$$

In particular, recalling the definition (4.2) of $L(t, T, T+\tau)$, the rate L paid on the CD is a simple spot rate

$$L = L(T, T, T+\tau) = \frac{1}{\tau}\left(\frac{1}{P(T, T+\tau)} - 1\right). \tag{5.1}$$

5.3 Forward Rate Agreements (FRA)

A certificate of deposit allows a market participant to lock in an interest rate for a given period of time, effective immediately. Many market participants, however, find it convenient to lock in interest rates for a given period of time that starts in the future. Contracts that provide such a rate guarantee are known as *forward contracts* or, in a fixed income context, *forward rate agreements* (FRAs). An FRA for the period $[T, T+\tau]$ is a contract to exchange fixed rate payment (agreed at the initiation of the contract) against a payment based on the time T spot Libor rate of tenor τ. While all payments on an FRA are exchanged at, or near[5], time T, the contract is structured so that the payments are made in $T+\tau$ dollars.

Formally, consider the origination at time t, $t \leq T$, of a unit notional FRA contract with a rate of k. Ignoring payment delays, from the perspective of the fixed rate payer the net payment at time T will be

$$V_{\text{FRA}}(T) = \tau\left(L(T, T, T+\tau) - k\right) / \left(1 + \tau L(T, T, T+\tau)\right),$$

with the (contractually specified) factor $1/(1+\tau L(T, T, T+\tau))$ applied to roll the payment to the future date $T+\tau$. We note that

[5] Typical market conventions call for a two business day payment delay, see Appendix 5.A for more details.

$$1/\left(1+\tau L(T,T,T+\tau)\right)=P(T,T+\tau)$$

so, by the fundamental pricing result (4.13), the value of this contract at time t is equal to

$$V_{\mathrm{FRA}}(t)=\beta(t)\mathrm{E}_t\left(\beta(T)^{-1}\tau\left(L\left(T,T,T+\tau\right)-k\right)P\left(T,T+\tau\right)\right)$$

(recall that $\beta(\cdot)$ is the money market account). Substituting (5.1) we obtain

$$V_{\mathrm{FRA}}(t)=\beta(t)\mathrm{E}_t\left(\beta(T)^{-1}\left(1-P\left(T,T+\tau\right)-\tau kP\left(T,T+\tau\right)\right)\right).$$

Since $P(\cdot,T+\tau)$ is a traded asset, its price deflated by the numeraire $\beta(\cdot)$ is a martingale. Thus

$$\begin{aligned}V_{\mathrm{FRA}}(t)&=P\left(t,T\right)-P\left(t,T+\tau\right)-\tau kP\left(t,T+\tau\right)\\&=\tau P\left(t,T+\tau\right)\left(\frac{P\left(t,T\right)-P\left(t,T+\tau\right)}{\tau P\left(t,T+\tau\right)}-k\right).\end{aligned}\tag{5.2}$$

Most often, FRAs are issued at no cost to either party at the time of origination. The value of k that makes the FRA contract have value 0 at the contract initiation time t is given by the *forward Libor rate* (see (4.2)),

$$k=L\left(t,T,T+\tau\right)=\frac{P\left(t,T\right)-P\left(t,T+\tau\right)}{\tau P\left(t,T+\tau\right)}.$$

Thus, a forward Libor rate has the financial interpretation of being a break-even rate on an FRA contract in interbank markets.

5.4 Eurodollar Futures

FRAs, being forward contracts on Libor rates, allow market participants to either lock in favorable rates for future periods, or to speculate on the future direction of rates. FRAs trade in the OTC market, and are open only to institutions that participate in this market. Alternatively, *futures contracts* on Libor rates are available on a number of international exchanges, including the Chicago Mercantile Exchange (CME), London International Financial Futures and Options Exchange (LIFFE), and Marchè à Terme International de France (MATIF). The CME interest rate futures contract on a three-month spot Libor rate on US dollar denominated deposits is called the *Eurodollar futures* or, simply, *ED futures* contract.

At maturity T, an ED futures contract is settled at

$$100\times\left(1-L\left(T,T,T+\tau\right)\right).$$

The *futures rate* $F(t,T,T+\tau)$ at time t (see (4.1.2)) is defined to be the rate such that the *quoted futures price* at time t is equal to[6]

[6]So, if the futures rate is 5%, the quoted futures price is 95.

$$100 \times (1 - F(t, T, T + \tau)).$$

As is the case for all futures contracts, ED futures are settled (marked to market) daily. Confusing matters somewhat, the actual amount of money that is settled between holders of the long and the short positions in an ED future is determined by the daily change in the *actual futures price* defined by

$$N_{\text{ED}} \times \left[1 - \frac{1}{4} F(t, T, T + \tau)\right],$$

where N_{ED} is the notional principal of the contract ($\$1,000,000$ for the CME's ED futures). In particular, for 1 basis point (0.01%) increase in the rate $F(t, T, T+\tau)$, the CME contract buyer pays $1,000,000 \times 0.25 \times 0.0001 = 25$ dollars to the seller.

As explained in Chapter 4, futures rates $F(t, T, T + \tau)$ are generally different from forward Libor rates $L(t, T, T + \tau)$. The problem of computing the difference, the *ED convexity adjustment*, is considered in Section 16.8.

Unlike FRAs, for which the deposit period is negotiated between two parties, ED futures are standardized. Available contracts expire on four specific dates, one each in March, June, September and December, over the next ten years. Such standardization increases liquidity in each particular contract.

5.5 Fixed-for-Floating Swaps

A *swap* is a generic term for an OTC derivative in which two counterparties agree to exchange one stream of cash flows against another stream. These streams are called the *legs* of the swap. A *plain vanilla fixed-for-floating interest rate swap* (a *plain vanilla swap*, or just a *swap* if there is no confusion) is a swap in which one leg is a stream of fixed rate payments and the other a stream of payments based on a floating rate, most often Libor. The legs are denominated in the same currency, have the same notional, and expire on the same date. Payment streams are made on a pre-defined schedule of contiguous time intervals, known as *periods*. Typically, the floating rate is observed (or *fixed*) at the beginning of each period, with both fixed and floating rate coupons being paid out at the end of the period. A plain-vanilla swap is economically equivalent[7] to a multi-period FRA, and serves the same purpose in the market as regular FRAs. Between interest rate dealers

[7]This is true up to subtle but potentially important discounting issues. As we have pointed out in Section 5.3, the net payment of an FRA is *contractually* discounted using Libor rate from $T + \tau$ to T, whereas in a swap, the net payment for a given period is discounted at the money market account rate from the end to the beginning of the accrual period. The two types of discounting can in fact be different in the presence of discounting-index *basis*, see Sections 6.5.2 and 6.5.3.

and financial institutions, swaps of different maturities are often traded to adjust interest risk positions of the parties involved, or to simply make bets on future direction of interest rates. Swaps are also used by corporates, often in conjunction with bond or note issuance, to transform fixed rate obligations into floating ones, or vice versa.

To formally define a fixed-floating swap, one specifies a tenor structure, i.e. an increasing sequence of maturity times, normally spaced roughly equidistantly (see Section 4.1.3)

$$0 \leq T_0 < T_1 < T_2 < \ldots < T_N, \quad \tau_n = T_{n+1} - T_n. \tag{5.3}$$

In a fixed-floating swap with fixed rate k, one party (the fixed rate payer) pays simple interest based on the rate k in return for simple interest payments computed from the Libor rate fixing on date T_n, for each period $[T_n, T_{n+1}]$, $n = 0, \ldots, N-1$. The payments are exchanged at the end of each period, i.e. at time T_{n+1}. In practice, the payments are netted, and only their difference changes hands. From the perspective of the fixed rate payer, the net cash flow of the swap at time T_{n+1} is therefore given by (on a unit notional)

$$\tau_n \left(L_n(T_n) - k\right), \quad L_n(t) = L\left(t, T_n, T_{n+1}\right),$$

for $n = 0, \ldots, N-1$. Dates when the Libor rates are observed are typically called *fixing dates*; dates when payments occur are called *payment dates.*

By the fundamental valuation result (4.13), the value of a swap is equal to the expected discounted value of its (netted) payments. Specifically, the value to the fixed rate payer of a unit notional fixed-floating swap at time t, $0 \leq t \leq T_0$, is given by[8]

$$\begin{aligned} V_{\text{swap}}(t) &= \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta\left(T_{n+1}\right)^{-1} \left(L_n(T_n) - k\right)\right) \\ &= \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta(T_n)^{-1} \left(L_n(T_n) - k\right) P\left(T_n, T_{n+1}\right)\right). \end{aligned}$$

Using the definition of Libor rates $L_n(T_n)$,

$$V_{\text{swap}}(t) = \beta(t) \sum_{n=0}^{N-1} \mathrm{E}_t(\beta(T_n)^{-1}(1 - P(T_n, T_{n+1}) - \tau_n k P(T_n, T_{n+1}))).$$

For each n, $P(\cdot, T_n)$ is a traded asset, so its price deflated by the numeraire $\beta(\cdot)$ is a martingale. Hence

$$V_{\text{swap}}(t) = \sum_{n=0}^{N-1} \left(P\left(t, T_n\right) - P\left(t, T_{n+1}\right) - \tau_n k P\left(t, T_{n+1}\right)\right).$$

[8] This is a somewhat idealized expression. See Appendix 5.A for more details on market day counting conventions and related topics.

Recalling the definition of $L_n(t)$, this can be rewritten as

$$V_{\text{swap}}(t) = \sum_{n=0}^{N-1} \tau_n P(t, T_{n+1}) (L_n(t) - k).$$

An important observation is that a vanilla fixed-floating swap can be valued on date t using only the term structure of interest rates observed on that date. In particular, swap values are not affected by the *dynamics* of interest rates, only their current levels.

The swap valuation formula above can be rewritten as follows,

$$V_{\text{swap}}(t) = \left(\sum_{n=0}^{N-1} \tau_n P(t, T_{n+1}) \right) \left(\frac{\sum_{n=0}^{N-1} \tau_n P(t, T_{n+1}) L_n(t)}{\sum_{n=0}^{N-1} \tau_n P(t, T_{n+1})} - k \right).$$

Using the definitions (4.8), (4.10) and (4.11) from Chapter 4:

$$A(t) \triangleq A_{0,N}(t) = \sum_{n=0}^{N-1} \tau_n P(t, T_{n+1}), \tag{5.4}$$

$$S(t) \triangleq S_{0,N}(t) = \frac{\sum_{n=0}^{N-1} \tau_n P(t, T_{n+1}) L_n(t)}{\sum_{n=0}^{N-1} \tau_n P(t, T_{n+1})}, \tag{5.5}$$

we obtain the convenient formula

$$V_{\text{swap}}(t) = A(t) (S(t) - k). \tag{5.6}$$

The quantity $A(\cdot)$ is the *annuity* of the swap (or its *PVBP*, for Present Value of a Basis Point), and the quantity $S(t)$ is the *forward swap rate*. Clearly, $S(t)$ is the value of the fixed rate that makes the swap have value 0 to both parties at time t; S is consequently often referred to as a *par* or *break-even* rate.

For plain-vanilla swaps, the fixed rate and the swap notional are constant through time. More general swaps are, however, not bound by such restrictions and both the fixed rate and the notional may vary from period to period. A non-standard swap with a notional schedule $\{q_n\}_{n=0}^{N-1}$ (non-constant but deterministic) and a fixed rate schedule $\{k_n\}_{n=0}^{N-1}$ has the value

$$\begin{aligned} V_{\text{genswap}}(t) &= \beta(t) \sum_{n=0}^{N-1} \tau_n q_n \mathrm{E}_t \left(\beta(T_{n+1})^{-1} (L_n(T_n) - k_n) \right) \\ &= \sum_{n=0}^{N-1} \tau_n q_n P(t, T_{n+1}) (L_n(t) - k_n). \end{aligned}$$

Certain general swaps have dedicated names, such as *amortizing swaps* (notional decreases with time) and *accreting swaps* (notional increases with time).

As we mentioned in Section 5.1, swaps linked to overnight rates (Fed-Funds/Eonia/Sonia) have recently become more popular. Among them the overnight index swap (OIS) is probably the most liquid, and is defined as a swap that pays a *compounded* overnight rate against fixed rate payments. To write down its definition, let us assume that a tenor structure (5.3) is given, and denote by $\{t_{n,i}\}_{i=1}^{K_n}$ the collection of all business days in the period $[T_n, T_{n+1})$, so that $T_n = t_{n,1} < \ldots < t_{K_n} < T_{n+1}$. Then the net payment of the OIS with fixed rate k at time T_{n+1} is given by

$$\tau_n \left(\overline{L}_n - k\right),$$

where the floating rate $\overline{L}_n$ for the n-th period of OIS is given by

$$\overline{L}_n = \frac{1}{\tau_n}\left(\prod_{i=1}^{K_n-1}\left(1 + (t_{i+1} - t_i)L(t_i, t_i, t_{i+1})\right) - 1\right). \tag{5.7}$$

Here we used the notation $L(t_i, t_i, t_{i+1})$ to denote the overnight rate. Equating the overnight rate with the short rate, we can use a more mathematically convenient (although not exactly correct) expression

$$\overline{L}_n = \frac{1}{\tau_n}\left(e^{\int_{T_n}^{T_{n+1}} r(t)\,dt} - 1\right). \tag{5.8}$$

5.6 Libor-in-Arrears Swaps

Allowing the fixed rate and the notional to vary through time is not the only way to generalize a swap. For a *Libor-in-arrears swap*, Libor rates are observed (fixed) at the end of each period rather than at the beginning. Thus, a value of a Libor-in-arrears payer swap is equal to

$$V_{\mathrm{LIA}}(t) = \beta(t)\sum_{n=0}^{N-1}\tau_n \mathrm{E}_t\left(\beta\left(T_{n+1}\right)^{-1}\left(L_{n+1}\left(T_{n+1}\right) - k\right)\right).$$

Interestingly, this seemingly innocuous modification makes the value of a swap model-dependent, in contrast to the standard fixed-floating swap. We will discuss pricing of in-arrears swaps in Chapter 16.

Libor-in-arrears swaps are popular in upward-sloping interest rate curve environments, i.e. when long-tenor rates are higher than shorter-tenor ones. In such a scenario, the break-even fixed rate on the Libor-in-arrears swap tends to look more "attractive" than that of a standard fixed-floating swap, thus increasing the desirability of the swap to those seeking to receive fixed rate payments.

5.7 Averaging Swaps

Libor rates are not restricted to being observed on either the start date or the end of the pay period. A popular example is the *averaging swap*, i.e. a swap where the floating rate is determined as an average of Libor rate observations taken at regular intervals over each coupon period. For example, let $\{(t^f_{n,i}, t^s_{n,i}, t^e_{n,i})\}_{i=1}^{K_n}$ be a collection of date triplets (fixing, start and end date) that define the rates to be used in calculating the payment in period n. Defining a set of weights $w_{n,i}$, $i = 1, \ldots, K_n$, the floating rate $\overline{L}_n$ for the period $[T_n, T_{n+1}]$ may be defined as

$$\overline{L}_n = \sum_{i=1}^{K_n} w_{n,i} L\left(t^f_{n,i}, t^s_{n,i}, t^e_{n,i}\right).$$

For the fixed rate swap payer, the averaging swap value is therefore

$$V_{\text{average}}(t) = \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta\left(T_{n+1}\right)^{-1}\left(\overline{L}_n - k\right)\right). \tag{5.9}$$

As a rule, the weights $w_{n,i}$ sum up to 1, $\sum_{i=1}^{K_n} w_{n,i} = 1$; the weights usually reflect the number of days (using the appropriate day counting conventions) that a given rate $L(t^f_{n,i}, t^s_{n,i}, t^e_{n,i})$ is supposed to be in effect. Computation of the valuation expression (5.9) can be done using techniques similar to those required for in-arrears swaps; see Chapter 16 for details.

Swaps linked to the average of the Federal funds rate are common examples of an averaging swap. Particularly noteworthy is the *Fed funds/Libor basis swap* which pays the average of the Fed funds rate (over a given period) against a payment based on a Libor rate for that period. This instrument is an example of a *floating-floating single-currency basis swap*, i.e., a swap that exchanges payments based on two different floating rates in the same currency. Closely related to Fed funds basis swaps are the *Fed funds futures* contracts traded on the Chicago Board of Trade (CBOT) exchange. These contract uses the 30 day running average of the Federal funds rate for settlement.

Remark 5.7.1. Going forward, in our product descriptions we shall normally assume that all cash flows pay at the end of the periods in which they fix. While this is common practice, as we have just seen the "pay-in-arrears" rule can be broken at will depending on the client's needs — the only (self-evident) restriction is that payments should be fixed by the time they are made.

5.8 Caps and Floors

A firm with liabilities funded at a floating (i.e., Libor) rate is naturally concerned with the possibility that interest rates, and thus its interest rate

payments, may increase in the future. One way to immunize against this risk is to pay fixed on a fixed-floating interest rate swap, in effect turning floating rate payments into fixed ones. While this will guarantee a fixed rate for funding payments for the duration of the swap, it will also mean forgoing the possibility of benefiting from a potential future drop in rates. An *interest rate cap* is a security that allows one to benefit from low floating rates yet be protected from high rates. Similarly, for an investor with assets earning a floating rate, a low-rate scenario is unfavorable. An *interest rate floor* is an instrument designed to protect against low interest rates yet allow the holder to benefit from high rates.

Formally, a cap is a strip of *caplets*, call options on successive Libor rates, and a floor is a strip of *floorlets*, put options on successive Libor rates. We encountered caplets already in Section 4.5.1 and recall that this instrument pays

$$\tau_n \left(L_n(T_n) - k\right)^+$$

per unit notional at time T_{n+1}. Similarly, a floorlet pays

$$\tau_n \left(k - L_n(T_n)\right)^+$$

per unit notional at time T_{n+1}. Then, N-period caps and floors have values at time t of

$$V_{\text{cap}}(t) = \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta \left(T_{n+1}\right)^{-1} \left(L_n(T_n) - k\right)^+\right),$$

$$V_{\text{floor}}(t) = \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta \left(T_{n+1}\right)^{-1} \left(k - L_n(T_n)\right)^+\right).$$

By switching to the T_{n+1}-forward measure (see Section 4.2.2) for the n-th caplet/floorlet, the valuation formulas can be written in a more convenient form

$$V_{\text{cap}}(t) = \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) \mathrm{E}_t^{T_{n+1}} \left(\left(L_n(T_n) - k\right)^+\right),$$

$$V_{\text{floor}}(t) = \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) \mathrm{E}_t^{T_{n+1}} \left(\left(k - L_n(T_n)\right)^+\right).$$

By Lemma 4.2.3, the Libor rate $L_n(\cdot)$ is a martingale under the T_{n+1}-forward measure. Hence, caplets/floorlets can be priced using "vanilla" models[9], such as the log-normal Black model (see Remark 1.9.4).

[9] By a *vanilla* model we mean a model that specifies the dynamics (or just the terminal distribution) of only a single rate, or at most a few rates, in contrast to term structure models that specify consistent dynamics for the entire term structure of interest rates. Often vanilla models are borrowed from equity or FX modeling; having the underlying rate a martingale makes such borrowing painless. We discuss vanilla models in Chapters 7, 8, 9 and 17.

The OTC market in caps/floors is very liquid. While individual caplets/floorlets are not traded, caps/floors are available in a number of maturities. This allows the volatility information for individual forward Libor rates to be extracted from market quotes for caps/floors of different maturities, at least in principle[10]. Once extracted, these volatilities may be combined with the volatilities observed from European swaption quotes (see below), to form a set of market inputs to which interest rate models for exotics are calibrated.

5.9 Digital Caps and Floors

Digital caps and floors work like regular caps and floors, except that the n-th digital caplet pays

$$\tau_n \times 1_{\{L_n(T_n)>k\}}.$$

Similarly, the n-th digital floorlet pays

$$\tau_n \times 1_{\{L_n(T_n)<k\}}.$$

Digital caps and floors provide a leveraged way to bet on the future direction of interest rates, more so than through standard caps and floors.

5.10 European Swaptions

Caps and floors have an asymmetric exposure to interest rates, a characteristic used by both hedgers and speculators. A similar exposure profile is provided by options on swaps, the so-called *European swaptions*. A European swaption gives the holder a right, but not an obligation, to enter a swap at a future date at a given fixed rate. A *payer* swaption is an option to pay the fixed leg on a fixed-floating swap; a *receiver* swaption is an option to receive the fixed leg.

Assuming the underlying swap starts on the expiry date T_0 of the option (a typical situation), the payoff for a payer swaption at time T_0 then equals

$$V_{\text{swaption}}(T_0) = (V_{\text{swap}}(T_0))^+ = \left(\sum_{n=0}^{N-1} \tau_n P(T_0, T_{n+1})(L_n(T_0) - k)\right)^+. \tag{5.10}$$

The value at an intermediate time t, $t < T_0$, must then equal

$$\begin{aligned} V_{\text{swaption}}(t) &= \beta(t)\mathrm{E}_t\left(\beta(T_0)^{-1}V_{\text{swaption}}(T_0)\right) \\ &= \beta(t)\mathrm{E}_t\left(\beta(T_0)^{-1}\sum_{n=0}^{N-1} \tau_n P(T_0, T_{n+1})(L_n(T_0) - k)\right)^+, \end{aligned}$$

[10] This "volatility bootstrap" is by no means trivial; we discuss it in Section 16.2.

which, using (5.6), can be rewritten in the more compact form

$$V_{\text{swaption}}(t) = \beta(t)\mathrm{E}_t\left(\beta(T_0)^{-1}A(T_0)\left(S(T_0)-k\right)^+\right). \tag{5.11}$$

Moreover, switching to the annuity measure, also known as the *swap measure*, Q^A from Section 4.2.5, the swaption value can be expressed as

$$V_{\text{swaption}}(t) = A(t)\mathrm{E}_t^A\left(S(T_0)-k\right)^+, \tag{5.12}$$

with the forward swap rate $S(\cdot)$ being a martingale in the swap measure Q^A; see Lemma 4.2.4.

It is evident from (5.12) that a payer European swaption is a call option — and a receiver European swaption is a put option — on the forward swap rate, struck at the fixed rate of the swap. Hence, swaptions could be priced using a vanilla model (see footnote 9), such as the Black model or similar. Conversely, values of European swaptions can be translated into market-implied distributional characteristics of forward swap rates, a topic discussed at length in Section 7.1.2. In particular, it is universal practice to quote swaption prices in terms of *implied* Black volatilities, i.e. volatilities that recover market price when used in the Black formula. In some markets (e.g., the US), it is also common to quote implied *Gaussian* volatilities, defined in the same way with regard to a Gaussian (rather than log-normal) model for the distribution of interest rates, see (7.16).

The market in swaptions is very liquid, with many different option maturities and swap underlyings actively traded. To characterize the full universe of traded instruments, given a tenor structure (5.3) we consider swaptions of different expiries $\{T_n\}_{n=0}^{N-1}$ that can be exercised into swaps that start at T_n and cover m periods[11], i.e. their last payment date is T_{n+m}. For a convenient way to denote the various swaptions, recall definitions (4.8), (4.10) and (4.11) and introduce

$$A_{n,m}(t) = \sum_{i=n}^{n+m-1} \tau_i P\left(t, T_{i+1}\right), \tag{5.13}$$

$$S_{n,m}(t) = \frac{\sum_{i=n}^{n+m-1} \tau_i P\left(t, T_{i+1}\right) L_i(t)}{\sum_{i=n}^{n+m-1} \tau_i P\left(t, T_{i+1}\right)}, \tag{5.14}$$

for $n = 0, \ldots, N-1$, $m = 1, \ldots, N-n$. Then the value of the (n,m)-swaption (a short-hand for an "m-period swaption with expiry T_n") is equal to

$$A_{n,m}(t)\mathrm{E}_t^{n,m}\left(\left(S_{n,m}(T_n)-k\right)^+\right),$$

where $\mathrm{E}_t^{n,m}$ denotes time t expectation in the appropriate swap measure, $\mathrm{Q}^{n,m}$. Note that in trader parlance, a (vanilla) T_n-maturity European swaption on a swap that runs from T_n to T_m is said to be a "T_n into $T_{m+n} - T_n$"

[11] A bit confusingly, such a swaption is often said to have *tenor* $T_{n+m} - T_n$, a characterization it inherits from the underlying swap rate.

swaption. For instance, a 5 year option on a 10 year swap would be a "5-into-10" (or "5y-into-10y", or simply "5y10y") swaption.

Clearly, when $m = 1$, the (n, m)-swaption reduces to a caplet (or floorlet) on the Libor rate $L_n(\cdot)$, so caplets and floorlets can be thought of as one-period swaptions. Whenever in this book swaptions are discussed or used, caplets and floorlets are thereby implicitly included. Collectively, all (n, m)-swaptions constitute *the swaption grid*.

Market quotes on swaptions, typically in terms of implied volatilities, in the swaption grid provide the most readily-available information on the volatility structure of interest rates. As swaptions in the grid cover overlapping sections of the term structure of interest rates, extracting clean volatility information from market quotes is a non-trivial exercise that forms the foundation for calibration of models used for exotic interest rate derivative pricing. We will have much to say about such *volatility calibration* later on.

While options on plain-vanilla swaps comprise the bulk of the liquid ("vanilla") interest rate market, options on general swaps (i.e. on swaps with non-constant notionals and fixed rates) also trade and are properly treated as exotic derivatives. Often, general swaps can be decomposed into baskets of standard swaps, in which case options on general swaps become *basket options*. Valuation of basket options requires information on the co-dependence structure of securities in the basket, information that is not readily available from the vanilla options markets. We demonstrate how to handle this complication in Section 19.4.

5.10.1 Cash-Settled Swaptions

The swaption contract discussed in the previous section involves *physical settlement*, in the sense that an actual interest rate swap is entered into, should the option be exercised at its expiry. Physically-settled swaptions are also known as *swap-settled swaptions*. An economically equivalent swaption contract is one that instead settles into a cash payment equal to the PV (present value) of the swap as observed at time T_0. Indeed, for both types, the swaption payoff (for a payer) is given by

$$A(T_0)\,(S(T_0) - k)^+\,, \tag{5.15}$$

see (5.10) and (5.11). In the European markets, a third variety of swaptions is common, the so-called *cash-settled* swaptions. For this type of option, rather than entering into a swap, the option holder will receive a cash payout upon exercise. The settlement amount is calculated by a formula similar to (5.15), except the annuity $A(\cdot)$ is not calculated by (5.4), but instead by discounting fixed rate payments at the swap rate $S(T_0)$. Specifically,

$$V_{\text{css}}(T_0) = a\,(S(T_0))\,(S(T_0) - k)^+\,,$$

where

$$a(x) = \sum_{n=0}^{N-1} \frac{\tau_n}{\prod_{i=0}^{n}(1+\tau_i x)}.$$

Notice that the cash settlement mechanism ensures a well-defined present value of the option payout, as long as the swap rate $S(T_0)$ is observable. In contrast, the value of exercise of a physically settled swaption — the computation of which requires knowledge of a strip of discount factors — may be estimated differently by different dealers, due to bid-ask spread effects and differences in curve building technology (see Chapter 6). Technically, however, the cash settlement mechanism induces certain valuation complications, and cash-settled swaptions cannot, strictly speaking, be considered vanilla options that can be priced using, e.g., a Black-type formula[12]. This follows from the fact that in the measure associated with the deflator $X(t) = a(S(t))$, the swap rate $S(\cdot)$ is *not* a martingale, and certain drift adjustments are required. We discuss valuation of cash-settled swaptions in Section 16.6.12. As they are the most liquidly-traded OTC interest rate options in the European market, cash-settled swaptions still could (and should) be used to extract information on the volatility structure of interest rates; the procedure, however, is necessarily more involved.

5.11 CMS Swaps, Caps and Floors

As the market in plain vanilla swaps is both deep and very active, market quotes of corresponding swap rates can be used as "indexes", i.e. market variables that can themselves be used in defining payoffs of other securities. The demand for such products is often driven by particular segments of fixed income markets. For example, mortgage lenders are primarily concerned with hedging interest rate risk arising from holding residential loans, some of which may have maturities as long as thirty years. Because of potential prepayments, the interest rate risk of a pool of such mortgages is often assumed to be closely connected to movements in the 10 year swap rate; hence, mortgage lenders are natural consumers of interest rate securities linked to the 10 year swap rate.

A constant-maturity swap (CMS) rate is defined as a break-even swap rate (see (5.5)) on a standard swap of a fixed maturity, e.g. 10 years or 30 years. A *CMS swap* works just like a standard fixed-floating (Libor) swap, except for the fact that floating leg payments are based on CMS, rather than Libor, rates. Formally, let $S_{n,m}(\cdot)$ be the m-period swap rate with the first fixing date T_n, as defined by (5.14). Then an m-period (payer) CMS swap's value is given by

[12]Nevertheless, this practice has been widespread until recently, and may still be in use in some institutions.

$$V_{\text{cmsswap}}(t) = \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta \left(T_{n+1}\right)^{-1} \left(S_{n,m}(T_n) - k\right) \right)$$

or, using the T_{n+1}-forward measure for each period,

$$V_{\text{cmsswap}}(t) = \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) \mathrm{E}_t^{T_{n+1}} \left(\left(S_{n,m}(T_n) - k\right) \right).$$

While standard swaps can be valued solely from knowledge of the term structure of interest rates, CMS swaps require an interest rate model for valuation; we return to a complete discussion in Chapter 16.

CMS caps and floors are defined as strips of European options on CMS rates, just like regular caps and floors are strips of European options on Libor rates:

$$V_{\text{cmscap}}(t) = \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) \mathrm{E}_t^{T_{n+1}} \left(\left(S_{n,m}(T_n) - k\right)^+ \right),$$

$$V_{\text{cmsfloor}}(t) = \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) \mathrm{E}_t^{T_{n+1}} \left(\left(k - S_{n,m}(T_n)\right)^+ \right).$$

CMS caplets are related to European swaptions, as both are European-style options on swap rates. The connection between the two types of securities is, however, subtle, as we shall discuss later in this book.

5.12 Bermudan Swaptions

A Bermudan swaption is an option to enter into a fixed-floating swap on any (or any from a given subset) of its fixing dates. For a given tenor structure (5.3), the holder of a standard Bermudan swaption has the right to exercise it on any of the dates $\{T_n\}_{n=0}^{N-1}$. Once exercised on date T_n, say, the option goes away, and the holder enters the swap with the first fixing date T_n and the final payment date T_N. The period up to $T_0 > 0$ is known as the *lockout* or *no-call* period. In common jargon, a Bermudan swaption on, say, a 10 year swap with a 2 year lockout period (at inception) is known as a "10 no-call 2", or "10nc2", Bermudan swaption.

Formally, at time T_n, the value of a payer[13], if exercised, is therefore

$$\begin{aligned} U_n\left(T_n\right) &= \beta(t) \sum_{i=n}^{N-1} \tau_i \mathrm{E}_{T_n} \left(\beta \left(T_{i+1}\right)^{-1} \left(L_i(T_i) - k\right) \right) \\ &= \sum_{i=n}^{N-1} \tau_i P\left(T_n, T_{i+1}\right) \left(L_i\left(T_n\right) - k\right). \end{aligned}$$

[13]Upon exercise, the holder of a payer (receiver) Bermudan swaption will pay the fixed (floating) leg of the swap.

Here, $U_n(T_n)$ here denotes the *exercise* value of the Bermudan swaption; loosely speaking, a Bermudan swaption contract is an option to chose between $U_n(T_n)$ for different $n = 0, \ldots, N-1$. More succinctly, we recall from Section 1.10 that the Bermudan option value at time T_n will be the maximum of $U_n(T_n)$ and the *hold value* $H_n(T_n)$, the latter defined as the value of a Bermudan swaption with the exercise dates $\{T_i\}_{i=n+1}^{N-1}$ only (compare to Sections 1.10 and 3.5).

Demand for Bermudan swaptions comes from different segments of fixed income markets. Mortgage companies use them to hedge pools of mortgages, with the flexibility of Bermudan exercise convenient in matching the uncertain timing of prepayments in mortgage pools. Investors seeking higher current income sell Bermudan-style options on swaps to increase the coupons they receive, as explained later in the context of callable Libor exotics. Bermudan swaptions are also used as hedges for callable coupon bonds.

While it may be tempting to think of Bermudan swaptions as straightforward generalizations of European swaptions, they are substantially more difficult to model and price. Indeed, it is fair to say that many valuation methods and techniques covered in this book were developed in response to the need to value and risk manage Bermudan swaptions. Bermudan swaptions are, by far, the most liquid exotic fixed income securities, with all interest rate dealers holding large inventories.

5.13 Exotic Swaps and Structured Notes

With market sophistication ever on the rise, clients demand increasingly complicated payouts, often in a familiar swap or bond format (although the appetite has waned somewhat post-crisis). In an *exotic swap*, a regular floating Libor leg is swapped against structured coupons that are allowed to be arbitrary functions of observed interest rates (such as Libor or CMS rates). A standard fixed-floating vanilla swap is an obvious and trivial example where the structured coupon simply is a fixed rate. A cap (or a floor) can be seen as another, less trivial, example. In particular, note that

$$\begin{aligned}(k - L_n(T_n))^+ &= \Big((k - L_n(T_n))^+ + L_n(T_n)\Big) - L_n(T_n) \\ &= \max\left(k, L_n(T_n)\right) - L_n(T_n),\end{aligned}$$

which demonstrates that a floor can be represented as an exotic swap in which a Libor rate is exchanged for a *floored payoff* $\max(k, L_n(T_n))$.

Exotic swaps often start their life as bonds, or notes, sold by banks to investors. In a structured note, the investor pays an up-front *principal amount* (e.g., \$10,000,000) to the issuer of the note, who in turn pays the investor a structured coupon, and repays the principal at the maturity of the note. The principal amount is invested by the issuer (or the trading

desk to which the issuer passes the note for risk management), and pays the Libor rate plus or minus a spread. From the perspective of the issuer (or the trading desk), the net cash flows of the note are those of an exotic swap.

In terms of valuation, if C_n is the structured coupon for the n-th period, the value of the exotic swap is equal to (from the perspective of structured leg buyer)

$$V_{\text{exotic}}(t) = \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta \left(T_{n+1}\right)^{-1} \left(C_n - L_n(T_n)\right)\right),$$

where we for brevity have assumed that both legs of the swap pay at the end date of each coupon period (see Remark 5.7.1). As discussed earlier, in this valuation equation, the coupon C_n can be a complicated function of interest rates, structured to reflect investors' views on the market, or to take advantage of current interest rate market conditions. For example, a floored payoff can be offered to an investor who believes that interest rates are poised for a fall in the future.

There is no universally agreed "taxonomy" for exotic swaps, but for our purposes we can distinguish between exotic swaps that are i) Libor-based, ii) CMS-based, iii) multi-rate, iv) range accruals, and v) generally path dependent. We proceed to described each type of swap in more details.

5.13.1 Libor-Based Exotic Swaps

In a Libor-based exotic swap, the structured coupon is a function of a Libor rate:

$$C_n = C_n \left(L_n(T_n)\right).$$

A large variety of structured coupons $C_n(\cdot)$ can be used. For example:

- A standard swap,
$$C_n(x) = k.$$
- Capped and floored floaters. For strike s, gearing g, cap c and floor f,
$$C_n(x) = \max\left(\min\left(g \times x - s, c\right), f\right). \tag{5.16}$$
- Capped and floored inverse floaters. For spread s, gearing g, cap c and floor f,
$$C_n(x) = \max\left(\min\left(s - g \times x, c\right), f\right). \tag{5.17}$$
- Digitals. For strike s and coupon k,
$$C_n(x) = k \times 1_{\{x>s\}}$$
or
$$C_n(x) = k \times 1_{\{x<s\}}.$$

- "Flip-flops" or "tip-tops". For strike s and two coupons, k_1 and k_2,

$$C_n(x) = \begin{cases} k_1, \, x \leq s, \\ k_2, \, x > s. \end{cases}$$

Different coupon types can be combined together to create new types of structured coupons.

A Libor-based exotic swap can usually be decomposed[14] into a sum of simpler instruments such as ordinary swap floating legs, fixed legs, caps and floors, and digital caps and floors. Therefore, if the prices of these simple contracts are available in the market (as is typically the case), Libor-based exotic swaps can be perfectly replicated by a one-time transaction in market-available instruments, a strategy referred to as *static replication*. Hence, by themselves, these instruments rarely present major valuation challenges. They do, however, serve as building blocks for more complicated securities.

5.13.2 CMS-Based Exotic Swaps

The payoffs from the previous section can be applied to CMS, rather than Libor, rates. Structured coupons are then deterministic functions of CMS rates. If an m-period rate is used, then a structured coupon for period n can be defined by

$$C_n = C_n \left(S_{n,m}(T_n) \right),$$

with $C_n(x)$ as defined in the previous section.

CMS-based exotic swaps can be decomposed into linear combinations of CMS swaps and CMS caps/floors and rarely present any extra modeling difficulties beyond those already present in CMS swaps and caps.

5.13.3 Multi-Rate Exotic Swaps

Multi-rate exotic swaps differ from the structures in Sections 5.13.1 and 5.13.2 by referencing multiple market rates (Libor or CMS) for the calculation of structured coupons. The most common example is a *CMS spread coupon*. To describe this contract, let $S_{n,a}(\cdot)$ and $S_{n,b}(\cdot)$ be two collections of CMS rates, fixing on T_n, $n = 0, \ldots, N-1$, and covering a and b periods, respectively. A CMS spread coupon with gearing g, spread s, cap c and floor f is then defined by

$$C_n = \max \left(\min \left(g \times \left(S_{n,a}(T_n) - S_{n,b}(T_n) \right) + s, c \right), f \right).$$

A typical example would be a 10 year/2 year (often abbreviated as 10y2y) CMS call spread option where a is 40 (40 quarterly periods to cover 10 years) and b is 8, with the quarterly coupon given by

[14] Indeed, this is the case for all the payouts listed above, a fact that we invite the reader to verify.

$$C_n = \max\left(S_{10y}(T_n) - S_{2y}(T_n), 0\right)$$

(using somewhat loose notation). A relatively liquid broker market exists for spread options on Euro and US dollar CMS rates.

A more general example is obtained by using one of the payoff functions $C_n(x)$ defined in Section 5.13.1, applied to the spread $x = S_{n,a}(T_n) - S_{n,b}(T_n)$. In particular, digital and flip-flop CMS spread swaps are quite popular.

Multi-rate exotic swaps typically cannot be decomposed into "standard" instruments (such as vanilla swaps, caps, etc.). Therefore, they, as a rule, cannot be valued by replication arguments, and a valuation model is required. Such a model, however, does not always need to be a full-blown term structure model: we shall show later that some types of spread-linked payoffs can be efficiently valued and risk managed by vanilla models (see footnote 9).

It should be noted that more than two rates can be used in the definition of a coupon. For example, in the so-called *curve cap* one takes a standard capped and floored payoff on a Libor or CMS (or CMS spread!) rate — see (5.16) — and makes the cap c and the floor f functions of, potentially different, CMS spreads:

$$\begin{aligned} C_n(x) &= \max\left(\min\left(g \times x - s, c\right), f\right), \\ c &= \max\left(\min\left(g_1 \times \left(S_{n,a_1}(T_n) - S_{n,b_1}(T_n)\right) + s_1, c_1\right), f_1\right), \\ f &= \max\left(\min\left(g_2 \times \left(S_{n,a_2}(T_n) - S_{n,b_2}(T_n)\right) + s_2, c_2\right), f_2\right). \end{aligned} \tag{5.18}$$

5.13.4 Range Accruals

A *range accrual* structured coupon is defined as a given rate — fixed in the simplest case, but potentially a Libor, CMS or a CMS spread rate — that only "accrues" when a different *reference* rate is inside (or, sometimes, outside) a given range. So, let $R_n(t)$ be the payment rate and $X_n(t)$ be the reference rate, and let l be the low bound and u be the upper bound. A range accrual coupon then pays

$$C_n = R_n(T_n) \times \frac{\sharp\left\{t \in [T_n, T_{n+1}] : X_n(t) \in [l, u]\right\}}{\sharp\left\{t \in [T_n, T_{n+1}]\right\}}, \tag{5.19}$$

where $\sharp\{\cdot\}$ is used to denote the number of days that a given criteria is satisfied.

The most common choice of the payment rate $R_n(t)$ is either a constant or Libor, but a CMS rate or any other structured coupon rate are also occasionally used. The reference rate $X_n(t)$ can be any market-observable rate such as a Libor rate fixing at t, a CMS rate fixing at t, or even a CMS spread rate.

We note that a range accrual coupon can always be decomposed into simpler digital payoffs, because

$$\sharp\{t \in [T_n, T_{n+1}] : X_n(t) \in [l, u]\} = \sum_{t \in [T_n, T_{n+1}]} 1_{\{X_n(t) \in [l,u]\}}, \tag{5.20}$$

where the sum on the right-hand side is over all business days in the period. This decomposition is particularly useful for fixed rate ($R_n(T_n) \equiv k$) range accruals, as simple digital options can be priced directly from the market information on European options (see Section 7.1.2). For floating, or more complicated, range accruals the decomposition is useful but requires further work to turn it into valuation formulas — see Section 17.5 for further details.

The basic payout (5.19) can be extended to include more than one range condition. In a *dual range accrual*, the position of two different reference rates relative to the range are monitored, and (5.19) is generalized to

$$C_n = R_n(T_n) \frac{\sharp\{\{t \in I_n : X_{n,1}(t) \in [l_1, u_1]\} \diamond \{t \in I_n : X_{n,2}(t) \in [l_2, u_2]\}\}}{\sharp\{t \in I_n\}}$$

with $I_n = [T_n, T_{n+1}]$ and $\diamond$ denoting either intersection $\cap$ or union $\cup$. In the former case, one counts the number of days when *both* reference rates $X_{n,1}$ and $X_{n,2}$ are within their ranges; in the latter case, one counts the number of days when *either* of the two reference rates are within their ranges.

In a *curve cap range accrual*, the lower and upper bounds become functions of CMS spreads themselves, similar to (5.18).

A *product-of-ranges* range accrual multiplies up all range accrual factors to date to define the multiplier that is used for the current coupon, e.g.

$$C_n = R_n(T_n) \times Y_n,$$

where

$$Y_n = Y_{n-1} \times \frac{\sharp\{t \in [T_n, T_{n+1}] : X_n(t) \in [l, u]\}}{\sharp\{t \in [T_n, T_{n+1}]\}}, \tag{5.21}$$
$$Y_{-1} = 1.$$

5.13.5 Path-Dependent Swaps

The payoff (5.21) is an example of *path-dependence* in the payoff, where a coupon depends on rate observations from previous coupon periods. More commonly, path-dependence in exotic swaps is introduced by linking a structured coupon not only to interest rates observed during the coupon period, but to previous coupon(s) as well. This is often referred to as a *snowball* feature. The "original" snowball structure involved a coupon of an inverse floating type, with the n-th coupon C_n given by

$$C_n = (C_{n-1} + s_n - g_n \times L_n(T_n))^+. \tag{5.22}$$

Here $\{s_n\}$ and $\{g_n\}$ are contractually specified deterministic sequences of spreads and gearings. This type of a swap is sometimes also called a *ratchet* or a *ladder* swap.

The term snowball originates with the tendency of high initial coupons to spill into subsequent coupons, in a compounding or "snowballing" fashion". Indeed, a little reflection reveals that a snowball has a highly leveraged exposure to its first few coupons, a feature that makes it more attractive to some investors, but also quite difficult to risk manage.

A large number of snowball-like payoffs have been created, often with "snow"-themed — and rather nonsensical — names, such as "snowrange", "snowbear", and "snowstorm". For example, a "snowrange" combines a range-accrual feature and a snowball feature, in the following way

$$C_n = C_{n-1} \times Y_n + s_n + g_n \times X_{n,1}(T_n), \tag{5.23}$$

where $X_{n,1}(T_n)$ is some reference rate, and Y_n is a range-accrual factor depending on a second rate $X_{n,2}$,

$$Y_n = \frac{\sharp\{t \in [T_n, T_{n+1}] : X_{n,2}(t) \in [l, u]\}}{\sharp\{t \in [T_n, T_{n+1}]\}}.$$

The range accrual factor Y_n may be a product-of-ranges accrual factor, as in (5.21). Also, additional caps and/or floors are often added to the coupon (5.23).

Path-dependent swaps typically require a term structure model for valuation; for obvious reasons, Monte Carlo methods are often mandatory.

5.14 Callable Libor Exotics

5.14.1 Definitions

As described in Section 5.12, Bermudan swaptions are Bermudan-style options to enter a regular fixed-floating swap. If we alter the swap underlying the Bermudan swaption from a regular swap to an exotic swap (see previous section), then a so-called *callable Libor exotic* (CLE) is created. CLEs most often emerge as part of callable structured notes in which an issuer receives the principal from an investor and pays a structured coupon in return. In addition, the issuer has the right to cancel — or *call* — the note on a schedule of dates; typically, this call schedule will coincide with coupon fixing dates, after some initial lock-out (or *no-call*) period. Should a note be called by its issuer[15], the principal is returned to the investor and no future coupons are paid.

A callable structured note is typically passed through by the issuer to an exotics trading desk (which could, but does not have to, be internal to the issuing bank) to deal with its risk management. Also, the principal is

[15]The call decision is most often made by the issuer's swap counterparty who is actually managing the risk, see next paragraph.

invested and pays a Libor rate, plus or minus a spread depending on the cost of financing. From a trading desk perspective what is left is an exotic swap paying structured coupons and receiving Libor, plus a Bermudan-style right to cancel the swap. For clarity, Figures 5.1–5.3 list the cash flow diagrams of a callable structured note[16].

Fig. 5.1. Callable Note: Flows at Inception

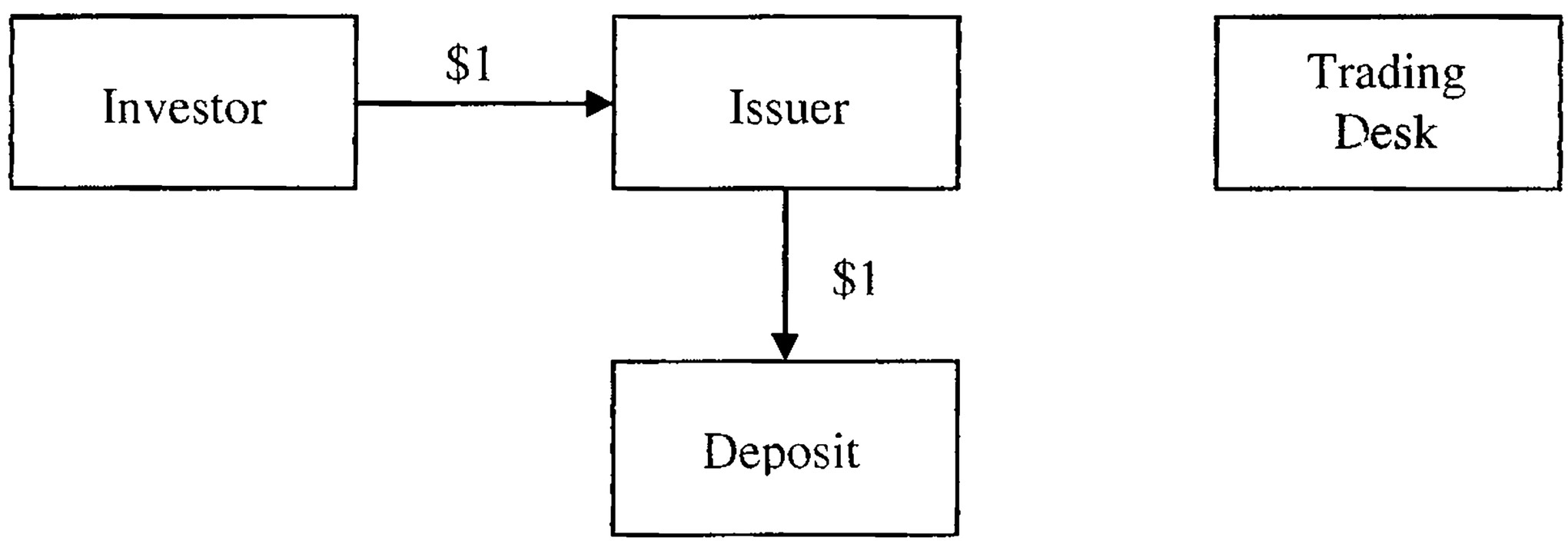
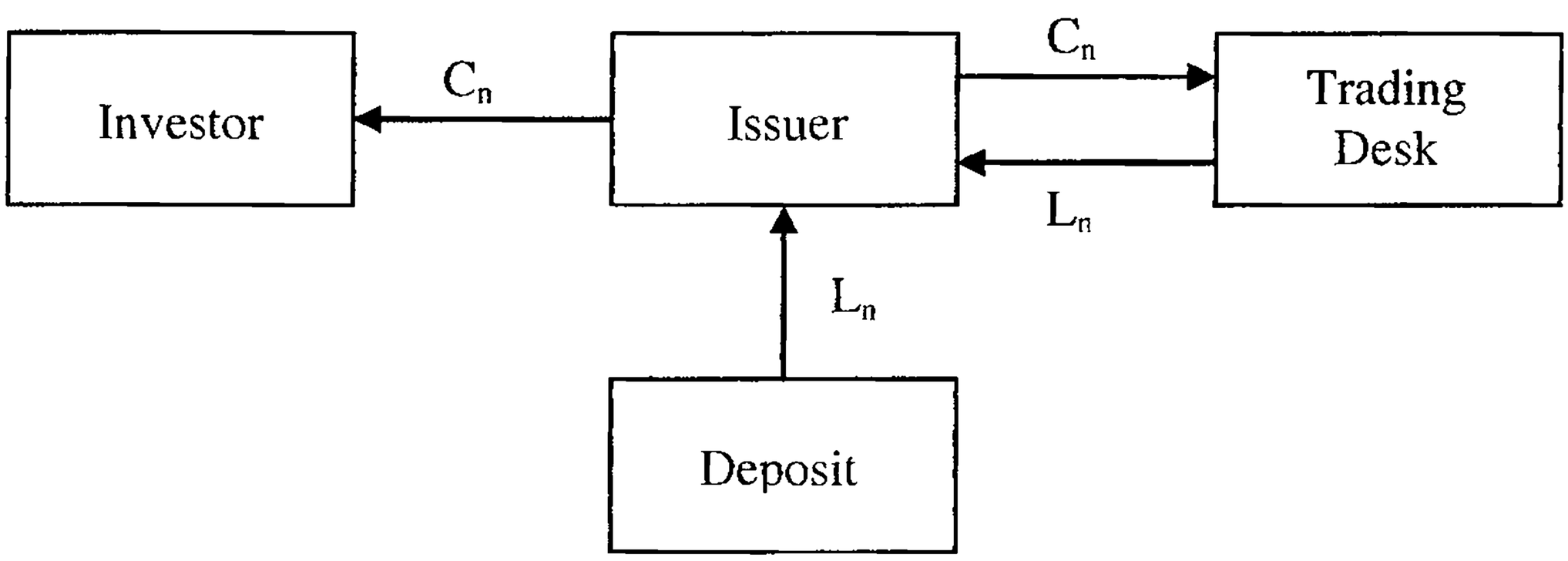


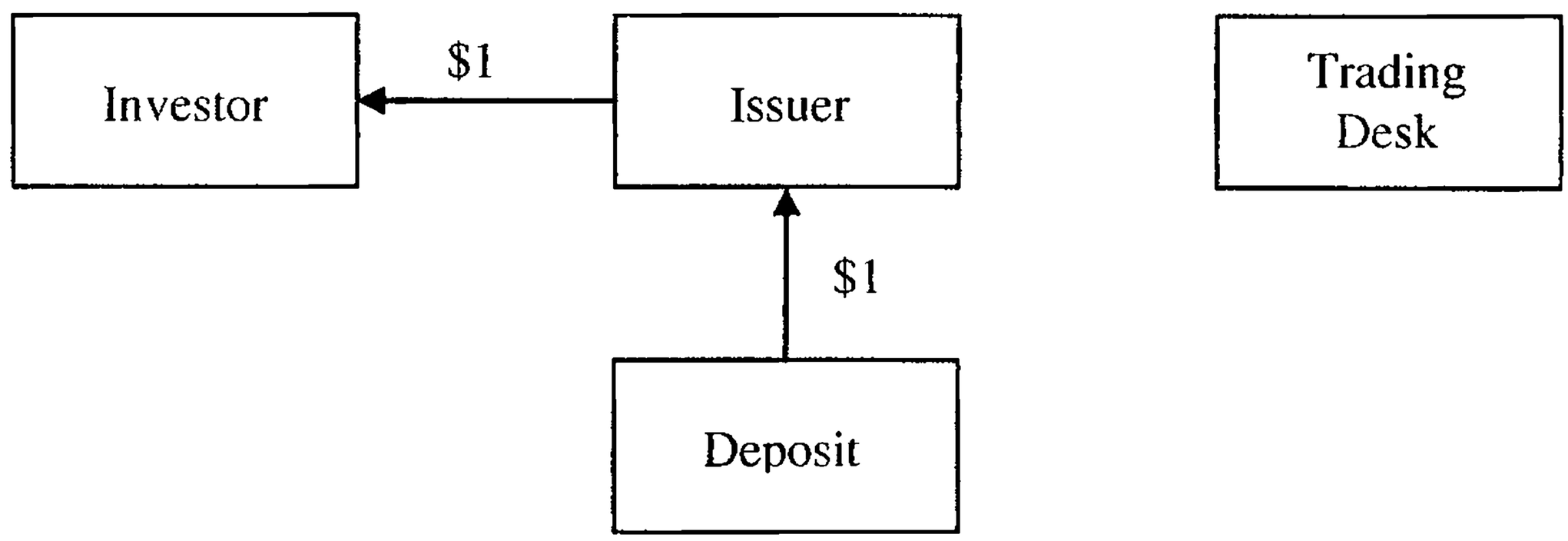
Fig. 5.2. Callable Note: Flows at Payment Times



Sometimes it is convenient to represent a cancelable exotic swap as a straight exotic swap, plus a Bermudan-style option to enter a reverse swap, i.e. a swap where legs are reversed relative to the original one. Beyond providing a break-down that is convenient for valuation purposes, this representation emphasizes the fact that the cancelability feature of a CLE benefits the party that owns it (typically a structured note issuer). Indeed, the feature is

[16] While, conceptually, the principal is deposited into a Libor-paying account, in practice it is used as part of cash management activities by the issuer. A structured note issuance program often provides cheaper funding to a bank than would be attainable by other means.

Fig. 5.3. Callable Note: Flows at Termination (Maturity or early Cancellation)



often added to a structured note as a way to offer a more attractive coupon to the investor, in return for the Bermudan-style option. Often, the coupons inside the non-call period are fixed rate coupons, and a typical way for the issuer to "pay" for the Bermudan option is to make these first coupons high, often much higher than the return available elsewhere. This "optical illusion" of high rate of return on investment is, at least in part, what drives investor interest in structured callable notes.

Consider a CLE on an exotic swap with structured coupons $\{C_n\}_{n=0}^{N-1}$. As for regular Bermudan swaptions, we denote the value of the exotic swap that one can exercise on date T_n by

$$U_n(T_n) = \beta(T_n) \sum_{i=n}^{N-1} \tau_i \mathrm{E}_{T_n} \left(\beta(T_{i+1})^{-1} \times (C_i - L_i(T_i)) \right). \tag{5.24}$$

Here, we recall that from a trading desk prospective, the cancelability feature of CLE involves an option to enter a reverse swap, with receipt of structure coupons and payment of Libor. Hold values are also defined analogously to the Bermudan swaption case: the n-th hold value $H_n(T_n)$ is defined as the time T_n value of the CLE on the same exotic swap, but with exercise dates $\{T_i\}_{i=n+1}^{N-1}$ only. That is, $H_n(T_n)$ is the time T_n value of the CLE provided it has not been exercised on or before T_n.

5.14.2 Pricing Callable Libor Exotics

A significant part of this book is dedicated to efficient methods for pricing and risk-managing callable Libor exotics. To provide a brief preview of the difficulties involved, we notice that the call feature embedded in CLEs may suggest application of PDE methods, using backward induction arguments outlined in previous chapters. However, often a CLE has explicit path dependence that makes the application of PDE methods impractical. In other cases, as we shall describe in greater depth later in the book, models that admit an efficient PDE representation are often too inflexible for

application to anything but the simplest of CLEs. Hence, CLEs more often than not must be valued using Monte Carlo methods. We have seen a preview of how optimal exercise can be handled in Monte Carlo in Section 3.5; many more details are provided in Chapter 18.

5.14.3 Types of Callable Libor Exotics

Any exotic swap can be used as an underlying for a callable Libor exotic. For our purposes, the taxonomy of callable Libor exotics can follow closely that of exotic swaps, see Section 5.13. We can thereby distinguish various types of CLEs, e.g. Libor-based, CMS-based, multi-rate, callable range accruals. callable snowballs, and so on. Many variations on the basic CLE design exist. most of which are driven by a desire to increase the value of the option to cancel that the investor sells the trading desk, in order for a higher coupon to be paid. It is difficult to classify all the features that have been invented: we content ourselves with merely listing some of the more popular ones.

5.14.4 Callable Snowballs

A callable snowball is a CLE with a snowball (or snowrange, etc.) underlying. From a modeling prospective, they are notable for being one of the first widely popular instruments that combine both strong path-dependence and optimal exercise. It is possible to incorporate snowball-type path-dependence into a PDE framework by introducing auxiliary variables, following the principles of Section 2.7.5; Section 18.4.5 discusses details specific to snowballs. Alternatively — and often preferably — optimal exercise can be incorporated into the Monte Carlo method, as discussed in Section 3.5 and later on in Chapter 18.

5.14.5 CLEs Accreting at Coupon Rate

Typically the notional of the underlying swap of a CLE is fixed throughout the life of the deal, but it does not have to be. For instance, it is not uncommon for the notional to vary deterministically, e.g. increase or decrease by non-random additive or multiplicative amounts each coupon period. Such deterministic accretion/amortization rarely adds extra complications from a modeling prospective. Occasionally, however, a contract specifies that the notional of the swap accretes at the structured coupon rate, in which case the accretion rate will be random. For such CLEs, the exercise value in (5.24) must be amended. Specifically, if q_i is the notional in place for the period $[T_i, T_{i+1}]$, then q_i is obtained from the notional over the previous period q_{i-1} by multiplication with the structured coupon over the previous period. Formally,

$$U_n(t) = \beta(t) \sum_{i=n}^{N-1} \tau_i \mathrm{E}_t \left(\beta \left(T_{i+1}\right)^{-1} \times q_i \times \left(C_i - L_i(T_i)\right) \right),$$
$$q_i = q_{i-1} \times \left(1 + \tau_{i-1} C_{i-1}\right),$$

where the initial notional q_0 is contractually specified.

Interestingly, the random accretion feature above can be incorporated into a PDE-based scheme without any extra cost, see Section 18.4.5.

5.14.6 Multi-Tranches

The more optionality an investor can sell to the issuer, the better coupon she can receive. As described earlier, the option to call the note is already present in a callable structured note. Another option that is sometimes embedded is a right for the issuer to increase the size of the note, i.e. to put more of the same note to the investor, whether she wants it or not. The name of this feature, a "multi-tranche" callable structured note, originates with the fact that these possible notional increases are formalized as tranches[17] of the same note that the issuer has the right to put to the investor. The times when the issuer has the right to increase the notional of the note typically come before the times when the note can be canceled altogether. Callability usually applies jointly to all tranches of the note.

By itself, the multi-tranche feature rarely presents modeling issues, although one must be mindful of it and plan for a pricing infrastructure that is flexible enough to handle it.

5.15 TARNs and Other Trade-Level Features

While sometimes the precise split is a little arbitrary, it is often helpful to think of a Libor exotic as being defined by

- A definition of its coupon, i.e. a formula that converts rates observed during a coupon period (and, sometimes, previous coupons) into the amount of money paid to the investor.
- A collection of trade-level features, i.e. features that cannot conveniently be expressed as coupon definitions, but instead "act" on the whole trade.

We have already seen examples of both features. For instance, a callable snowball CLE has a coupon definition given by the formula (5.22) on top of which callability has been added as a trade-level feature. In the next few sections we review some other trade-level features.

[17] "Slices" in French; here meaning "similar securities offered as part of the same transaction".

5.15.1 Knock-out Swaps

A knock-out swap is just an exotic swap that disappears on the first fixing date on which some reference rate is above (or below) a given barrier. If the knock-out rate for the period n is denoted by $X_n(t)$, the coupon by C_n, the Libor rate by L_n, and the knock-out barrier by R, the value of a down-and-out[18] knock-out swap is given by

$$V_{\mathrm{KO}}(t) = \beta(t)\mathrm{E}_t\left(\sum_{n=0}^{N-1}\tau_n\beta\left(T_{n+1}\right)^{-1}\times\left(C_n - L_n(T_n)\right)\times\prod_{i=0}^{n}1_{\{X_i(T_i)>R\}}\right).$$

5.15.2 TARNs

Callable structured notes have proved to be popular with investors, but suffer from the drawback that investors rarely know when the issuer will call the note — indeed, the decision to exercise a Bermudan-style option is driven by a model rarely accessible to the average investor. A relatively recent innovation, the *Targeted Redemption Note* (TARN), presents one possible solution to the problem. In a TARN (see Piterbarg [2004c]), the total investor return, defined as the sum of all structured coupons paid to date, is recorded over time. When the total return exceeds a pre-specified target level— hence the name of the structure — the note is terminated and the principal is returned to the investor.

As with callable notes, issuers do not keep TARN structures on their books, but swap them out with a trading desk. Since the principal payment from investors is invested at the Libor rate, to a trading desk a TARN looks like an exotic swap that knocks out on the total sum of structured coupons. Formally, let the structured coupon[19] for the period $[T_n, T_{n+1}]$ be C_n. The coupon over the period $[T_n, T_{n+1}]$ is only paid if the sum of structured coupons up to (and not including) time T_n is below a total return R. Thus, the value of the TARN at time 0 from the investor's viewpoint is given by

$$V_{\mathrm{tarn}}(t) = \beta(t)\mathrm{E}_t\left(\sum_{n=1}^{N-1}\tau_n\beta\left(T_{n+1}\right)^{-1}\times\left(C_n - L_n(T_n)\right)\times 1_{\{Q_n<R\}}\right), \tag{5.25}$$

$$Q_n = \sum_{i=1}^{n-1}\tau_i C_i, \quad Q_1 = 0.$$

[18] I.e. disappearing upon some variable breaching a barrier from above; compare to up-and-out options discussed in Chapter 2.

[19] In the original TARN product, an inverse floating coupon (5.17) was used, but any structured coupon can be employed.

We note that a TARN typically pays some fixed coupons to an investor before the knock-out feature starts, mirroring the non-call structure of CLEs, see Section 5.14. We omit these from the contract description as their present value can be computed statically off an interest rate curve separately, as the payments are known in advance.

Various features can be added to the description of the TARN we just described. For example, the last coupon, i.e. the coupon that pushes the total return over the target R, can be paid only partially to make the total return exactly R and not more. This feature is knowns as a *cap at trigger* or a *lifetime cap*. Also, if the total return over the life of the TARN never reaches the target R, a TARN equipped with a so-called *lifetime floor* will make a *make whole* payment at the TARN maturity to ensure that the total return exactly equals R, and not less. These features do not alter the general modeling framework for TARNs that we develop in later chapters, and we shall generally ignore them.

While it could be argued that a TARN is really just a swap with a different coupon definition (namely, $C_n \times 1_{\{Q_n < R\}}$)[20] we prefer to classify a TARN feature as trade-level, reflecting its historic importance and its relationship with the callability feature.

5.15.3 Global Cap

As discussed, a TARN can have a feature that restricts total return to an investor to be exactly the trigger level R. This feature can be decoupled from the TARN definition and used by itself, often called a *global cap*. Specifically, an exotic swap with a global cap R pays a structured coupon C_n to an investor until the sum of the coupons has reached the level R. Note that a swap typically does not terminate at this point, i.e. the trading desk will continue to receive Libor until the maturity of the trade or some other agreed termination event.

5.15.4 Global Floor

A *global floor* guarantees the investor a minimum cumulative return of the note. Specifically, if the sum of structured coupons paid to the investor does not reach a global floor value of F by the maturity of the deal, the issuer will pay a make whole amount to the client, equal to F minus the sum of actual coupons paid. The payment is made at the maturity of the swap or some other termination event (such as when the swap is canceled, in case of a callable global floor note).

[20] The same could be said about knock-out swaps previously defined.

5.15.5 Pricing and Trade Representation Challenges

Trade-level features are often combined with each other, and with various coupon formulas. For example, one can be asked to price a callable, TARN'ed snowball with a global cap. As we shall argue in later chapters, the only modeling solution that is sufficiently scalable to accommodate such arbitrary combinations of various trade features involves a generic, flexible model that is calibrated to a broad collection of market volatility information. In such a framework, adding extra features to a trade is ultimately not much of a modeling problem, but could be a significant trade representation challenge. While outside the scope of this book, such a challenge should be addressed by a software framework that is flexible enough to represent any, current or future, trade-level features, and incorporate them into a pricing engine without significant extra effort. A successful implementation of such a trade representation framework requires careful planning and considerable investment. One fairly of common route is to use a domain-specific programming language for trade representation, see for example Jones et al. [2000] for a commercially available version or Frankau et al. [2009] for an example of an in-house one.

5.16 Volatility Derivatives

In a nutshell, a *volatility derivative* is a contingent claim whose underlying is the volatility of a financial observable, rather than a financial observable itself. The simplest example of such a derivative is the *variance swap* (see Carr and Lee [2009b]), a structure that first emerged in equity and foreign exchange trading. In the last few years, similar ideas and structures have entered the fixed income derivative arena.

The demand for volatility derivatives in interest rates is driven by the same factors as in other asset classes; a common motivation is the desire of some market participants — hedge funds in particular — to have direct exposure to interest rate volatility, but not to the outright level of rates, say. In other cases, the product development follows the usual path of creating structured notes with appealing payoff profiles.

Different interest rates constitute different financial observables for defining a volatility, hence one needs to be rather specific when defining a volatility-linked payoff.

5.16.1 Volatility Swaps

A *volatility swap* in interest rates is a contract that measures realized volatility (or a related quantity) of a given rate, although it is structured somewhat differently from volatility swaps in equity or FX markets. Let

$X_n(t)$ be the rate used for period n; then the most common coupon of a volatility swap is given by

$$C_n = |X_{n+1}(T_{n+1}) - X_n(T_n)|,$$

or a capped version

$$C_n = \min(|X_{n+1}(T_{n+1}) - X_n(T_n)|, c).$$

The value of the (structured) leg of a volatility swap measures realized variation of the rate $X_n(\cdot)$,

$$\begin{aligned} V_{\text{volswap}}(t) =& \beta(t)\mathrm{E}_t\left(\sum_{n=0}^{N-1}\tau_n\beta(T_{n+1})^{-1}\times|X_{n+1}(T_{n+1}) - X_n(T_n)|\right) \\ &+ V_{\text{floatleg}}(t), \end{aligned} \tag{5.26}$$

where

$$V_{\text{floatleg}}(t) \triangleq \sum_{n=0}^{N-1}\tau_n P(t, T_{n+1})L_n(t) = 1 - P(0, T_N).$$

There are two common choices for the rate X_n. One choice, a *fixed-tenor* volatility swap, involves a swap rate of the same tenor on each of the fixing dates. Technically speaking, *different* swap rates are therefore used for different periods,

$$X_n(t) = S_{n,m}(t)$$

with a fixed value of m, the number of periods in the swap rate (see the definition (5.14)). For example, a rolling 10 year CMS rate could be used. The other choice, a *fixed-expiry* volatility swap, specifies the swap rate to have a fixed expiry and tenor, i.e.

$$X_n(t) = S_{K,m}(t).$$

With this definition, the volatility swap pays the absolute variation of a rate with the fixing date T_K and spanning m periods of the tenor structure $\{T_n\}_{n=0}^{N+m}$. Often $K = N$, so that the variability of the rate $S_{N,m}$ is measured over the whole of its life.

Recently, volatility swaps on CMS spread rates have appeared. As the name implies, they measure the variation of the spread of two rates, e.g. $X_n = S_{n,m_1} - S_{n,m_2}$.

5.16.2 Volatility Swaps with a Shout

Sometimes, the investor in a volatility swap is given an option to *shout*, that is to choose when the fixing of the rate occurs for the purposes of calculating the coupon payoff. In particular, the payoff of the n-th coupon is then

$$C_n = |X_{n+1}(\eta_n) - X_n(T_n)|,$$

where the random stopping time $\eta_n \in [T_n, T_{n+1}]$ is chosen by the investor coupon-by-coupon. For the uncapped version of this payoff, it is intuitively clear that it is always optimal to postpone the shout until the end of the period, i.e. $\eta_n = T_{n+1}$. So, the option given to the investor is actually worthless, while designed to appear to have some value. Interestingly, for the capped version, it is optimal[21] to shout at the lesser of T_{n+1} and the first time that $|X_{n+1}(\eta_n) - X_n(T_n)| \geq c$, where c is the cap level. As a consequence, a capped volatility swap with a shout option is equivalent to a volatility swap with a barrier on each coupon:

$$\begin{aligned} C_n = {} & c \times 1_{\left\{\max_{t\in[T_n,T_{n+1}]}|X_{n+1}(t)-X_n(T_n)|\geq c\right\}} \\ & + |X_{n+1}(T_{n+1}) - X_n(T_n)| \times 1_{\left\{\max_{t\in[T_n,T_{n+1}]}|X_{n+1}(t)-X_n(T_n)|<c\right\}}. \end{aligned}$$

This decomposition follows from results in Broadie and Detemple [1995] and is discussed in more detail in Chapter 20. The fact that one can replace an optimal exercise feature with a known static barrier is quite convenient and allows for easy Monte Carlo valuation.

5.16.3 Min-Max Volatility Swaps

The structured coupon for a min-max volatility swap is given by

$$C_n = M_n - m_n,$$

where

$$\begin{aligned} M_n &= \max_{t\in[T_n,T_{n+1}]} X_n(t), \\ m_n &= \min_{t\in[T_n,T_{n+1}]} X_n(t). \end{aligned}$$

The coupon represents the spread between the maximum and the minimum values that a given rate achieves during a coupon period.

While the two products appear quite different at a first glance, there is an interesting connection between min-max and standard volatility swaps. We shall explore this further in Chapter 20.

5.16.4 Forward Starting Options and Other Forward Volatility Contracts

The value of a standard European swaption and a standard fixed-expiry volatility swap are both linked to the volatility of the swap rate over its

[21] Ignoring some small convexity effects and the difference between discrete vs. continuous shout option rights.

entire life, i.e. from the valuation date to the fixing date of the swap rate. Some clients, however, prefer securities that are linked to the volatility of a swap rate as measured over only a sub-period of this time; in effect, the clients want exposure to what is often known as *forward* volatility. Precise definitions and the importance of forward volatility are left for future chapters; for now we content ourselves by listing a few relevant varieties of forward volatility derivatives.

Midcurve swaptions are swaptions whose expiry T^e is strictly before the fixing date T_0 of the underlying swap rate. Their value depend on the volatility of the swap rate over the period $[t, T^e] \subset [t, T_0]$.

Given a swap rate $S(\cdot)$ with the fixing date T_0, and a date $T^s < T_0$, a *forward-starting swaption straddle*[22] is given by the payoff

$$A(T_0) \times |S(T_0) - S(T^s)|$$

paid at T_0. Essentially, this contract is a combination of a receiver and a payer swaption, both of which will have their strikes fixed at time T^s to the then-prevailing level of the underlying swap rate. That is, the contract pays the value of the at-the-money straddle on the rate $S(\cdot)$ at time T^s with expiry T_0. The value of a forward-starting swaption straddle is driven by the volatility of the swap rate over the period $[T^s, T_0]$.

Recall that European swaptions are typically quoted in terms of their implied volatilities. Due to this convention, some clients find the forward starting straddle too indirect and instead want to receive implied volatility itself. Particularly popular are contracts that pay implied Normal[23] volatility as defined on p. 204. Fortunately, the at-the-money Normal volatility has a direct relationship to the swaption price, and the payoff of an *implied Normal volatility contract* is

$$\sqrt{\frac{\pi}{2\,(T_0 - T^s)}} \times |S(T_0) - S(T^s)|$$

paid at T_0. Apart from the factor $A(T_0)$ and a non-consequential scaling factor, the payoffs of a forward-starting straddle and the implied Normal volatility contract are identical. The differences in their prices are just a matter of a minor convexity correction, a topic we return to in Chapter 20.

[22] The term *straddle* is used to denote the sum of a put and a call option with identical strikes.

[23] Contracts paying implied log-normal, or Black, volatility are possible but less common, due to the common perception that interest rates are more Gaussian than log-normal, i.e. the implied Gaussian volatility is less sensitive to the changes in the level of interest rates than the implied Black volatility.

5.A Appendix: Day Counting Rules and Other Trivia

In this appendix, we very briefly cover some of the finer details of how schedules are constructed and how interest rate payments accrue under market conventions. We generally ignore these details in the main body of the book, and our treatment here only scratches the surface. For a full account, see Mayle [1993] or Stigum and Robinson [1996].

5.A.1 Libor Rate Definitions

Consider the 6 month Libor rate L fixing at time T. According to (4.2), we would compute this rate as simply

$$L(T) = L(T, T, T + 1/2) = \left(P(T, T + 1/2)^{-1} - 1\right)/\tau, \quad \tau = 1/2. \quad (5.27)$$

In reality, this computation ignores a number of quoting conventions. First, a 6 month USD Libor rate that fixes at time T, does not truly cover an *accrual period* of $[T, T + 1/2]$. Instead, the start date T^s of the accrual is set to be $T^s = T + \delta^s$, where δ^s is a delay of two business days[24]. In other words, the quoted spot Libor rate is in actuality based on a *forward starting* CD that is entered into with time lag of δ^s after the quotation date T. As for the end date T^e of the accrual period, it is normally determined by counting 6 months ahead starting from T^s, adjusting the resulting date to ensure that it is a valid business day. The precise mechanism used to make such a business day adjustments of T^s is determined from a *date rolling convention.* For USD Libor, one always uses the so-called "Modified-Following" convention where weekend or holiday dates are rolled forward to the next business day, unless doing so would cause T^e to lie in the next calendar month, in which case the payment date is rolled to the previous business day. Other rolling conventions are discussed in Mayle [1993] and Stigum and Robinson [1996].

Once the correct accrual period $[T^s, T^e]$ has been determined, to compute rate accrual it remains to compute the proper *year fraction* (or *accrual factor* or *day count fraction*) τ representing how many whole years are spanned by $[T^s, T^e]$. For the purposes of our book, we normally write simply $\tau = T^e - T^s$, but a little thought shows that expressions like this are ambiguous when T^e and T^s are thought of as actual (discrete) calendar dates, rather than as arbitrary numbers on the real line. For instance, given the existence of leap years, how many days are there in a standard year? For quant purposes, it is common to assume that a calendar year has 365.25 years, such that $T^e - T^s$ is obtained by simply counting the number of days between T^s and T^e and then dividing this number by 365.25. This "convention" is sometimes known as Actual/365.25 (or sometimes just A365.25), and is rarely, if ever, used for actual market quoting purposes. Instead, for quotation of Libor

[24] Libor rates in other currencies may have different delays. For instance, GBP Libor has zero business day delay.

rates the standard is to use an Actual/360 (A360) convention, where the number of days between T^s and T^e are converted to a year fraction by dividing by 360. As a consequence, the true value of τ used for 6 month Libor quotation purposes is typically slightly larger than 1/2. For additional year-count conventions (of which there are many), see Mayle [1993] and Stigum and Robinson [1996].

Due to the quoting standards used in real Libor markets, the relationship between discount bonds and quoted Libor rates is more complicated than (5.27). Specifically, if $L_{\text{mkt}}(T)$ represents the true quoted 6 month Libor rate at time T, we instead have

$$L_{\text{mkt}}(T) = \left((P(T,T^e)/P(T,T^s))^{-1} - 1 \right) \frac{360}{D(T^s,T^e)}, \tag{5.28}$$

where by $D(T^s,T^e)$ we denoted the number of days between T^s and T^e according to the convention used. Notice in particular how the formula now involves a forward starting zero-coupon bond $P(T,T^s,T^e) = P(T,T^e)/P(T,T^s)$, as a reflection of the settlement delay δ^s. Using existing "idealized" Libor rate notation (see (4.2)), we may write this expression as

$$L_{\text{mkt}}(T) = L(T,T^s,T^e) \times \frac{360}{365.25} \approx L(T,T^s,T^e) \times 0.986.$$

The difference between $L_{\text{mkt}}(T)$ and $L(T)$ is small enough for us to ignore it in most of this book, but any real system implementation obviously should use precise day counting rules when computing Libor fixings.

5.A.2 Swap Payments

The payments on swaps (and other instruments, such as CDs and FRAs) are subject to similar conventions as the Libor rate. Consider for instance a standard fixed-for-floating interest rate swap issued at time t (today). First, a[25] schedule $\{T_i\}_{i=0}^N$ for interest rate accrual must be constructed, starting from a given base frequency of the swap (e.g.: semi-annual). As was the case above, the schedule normally starts one or two business days after time t, i.e. $T_0 = t + \delta_0$ where δ_0 is some contractually specified delay. Date T_0 is known as the *effective date* of the swap. Given T_0, the remaining T_i, $i = 1, \ldots, N$, are computed by first laying out "unadjusted" dates according to the swap base frequency, and then applying a date rolling convention (typically Modified-Following) to each of the dates. As part of the swap contract, associated with each accrual period $[T_i, T_{i+1}]$ are then:

- A fixing date T_i^f: the date on which the floating leg index (Libor, most often) is observed. Typically T_i^f is two business days before time T_i.

[25] We assume that the fixed and floating legs pay interest on the same schedule, but in reality, this may not be the case. For instance, in USD, the standard frequency for the fixed leg is six months, and three months for the floating leg.

- A payment date T_i^p: the date on which the swap payments are made. Normally $T_i^p = T_{i+1}$, but it is not uncommon to have payment delays of 1 or 2 business days after T_{i+1}.
- A fixed leg year fraction τ_i^{fix}: the year fraction used to determine the payment at time T_i^p on the fixed leg. In the US, the most common convention for the fixed leg is[26] 30/360.
- A floating leg year fraction τ_i^{flt}: the year fraction used to determine the payment at time T_i^p on the floating leg[27]. In the US, the most common convention for the floating leg is Actual/360.

At time t, the value of a payer swap paying a fixed coupon c against Libor is therefore

$$V_{\text{swap}}(t) = \sum_{i=1}^{N} P(t, T_i^p) \mathrm{E}^{T_i^p} \left(\tau_i^{flt} L_{\text{mkt}}(T_i^f) - \tau_i^{fix} c \right),$$

where we have used the T_i^p-forward measure to state the valuation, and where L_{mkt} is defined in (5.28). In the book, we normally simplify this to

$$V_{\text{swap}}(t) = \sum_{i=1}^{N} P(t, T_i) \tau_i \mathrm{E}^{T_i} \left(L(T_{i-1}) - c \right), \quad \tau_i = T_i - T_{i-1}.$$

[26] When counting days in the 30/360 convention, each month is assumed to have 30 days. The number of days used to determine interest rate accrual will therefore differ from the *actual* number of days (which distinguishes 30/360 from Actual/360).

[27] Another, often ignored, complication with the floating leg is that the periods that define the (payment) year fractions τ_i^{flt} are sometimes slightly different from those that define forward Libor rates, due to certain conventions surrounding date adjustments. Again, see details in Mayle [1993] or Stigum and Robinson [1996].

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)

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

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

$$\mathbf{V} = \mathbf{c}\mathbf{P} \tag{6.4}$$

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 $y : [0, \mathcal{T}] \to \mathbb{R}_+$ 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,

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

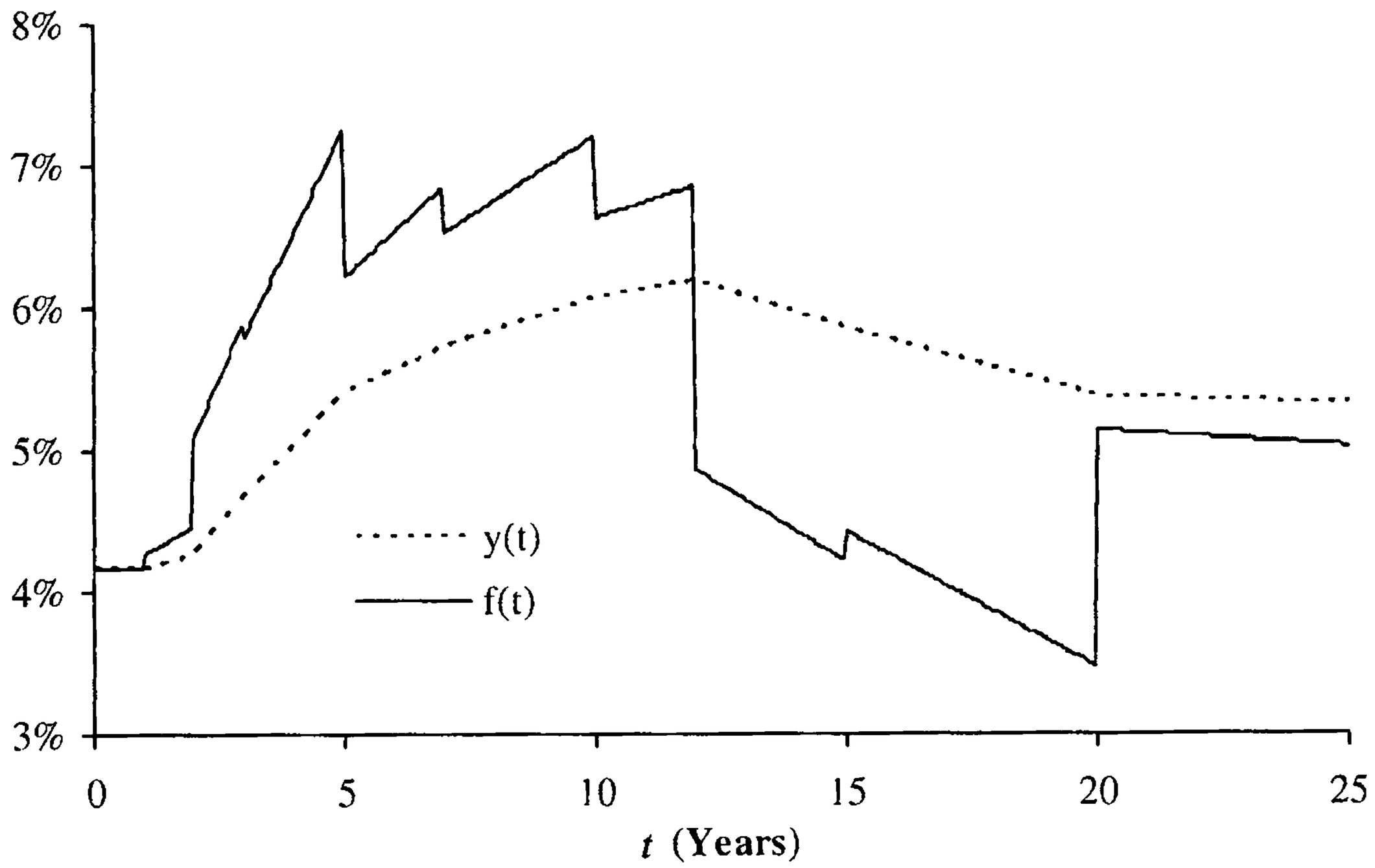
Setting $y(t) = y(1)$, $t < 1$, and then running the bootstrap procedure on the swap price expression (6.2) results in the yield shown in Figure 6.1. The same figure also shows the continuously compounded forward curve, as computed by equation (6.7). The discontinuous "saw-tooth" shape of the forward 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 $T \in [T_i, T_{i+1})$,

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

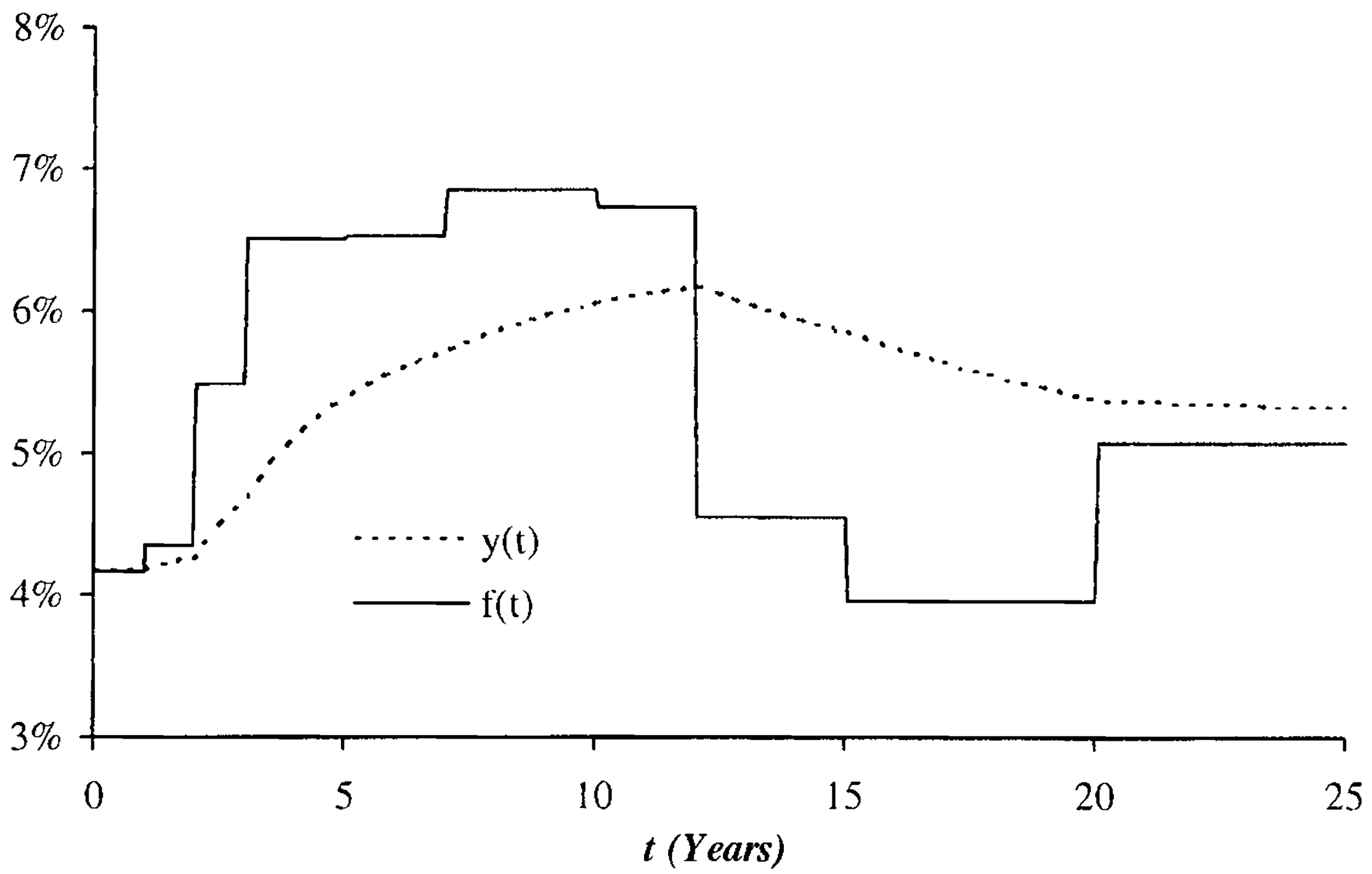
or

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

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

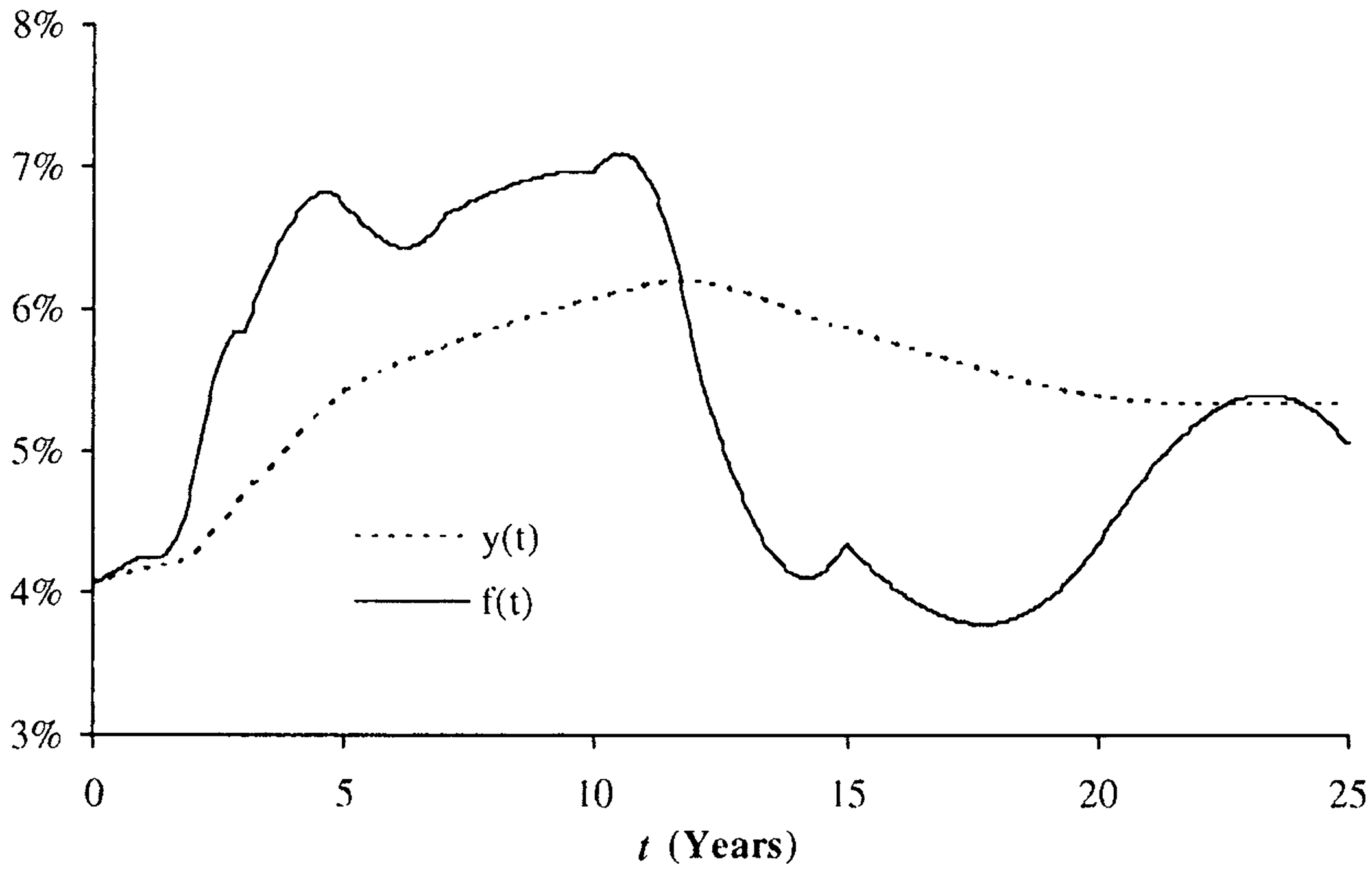
$$N^{-1} \sum_{i=1}^{N} \left(y_i^{(k+1)} - y_i^{(k)} \right)^2 < \varepsilon^2,$$

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.

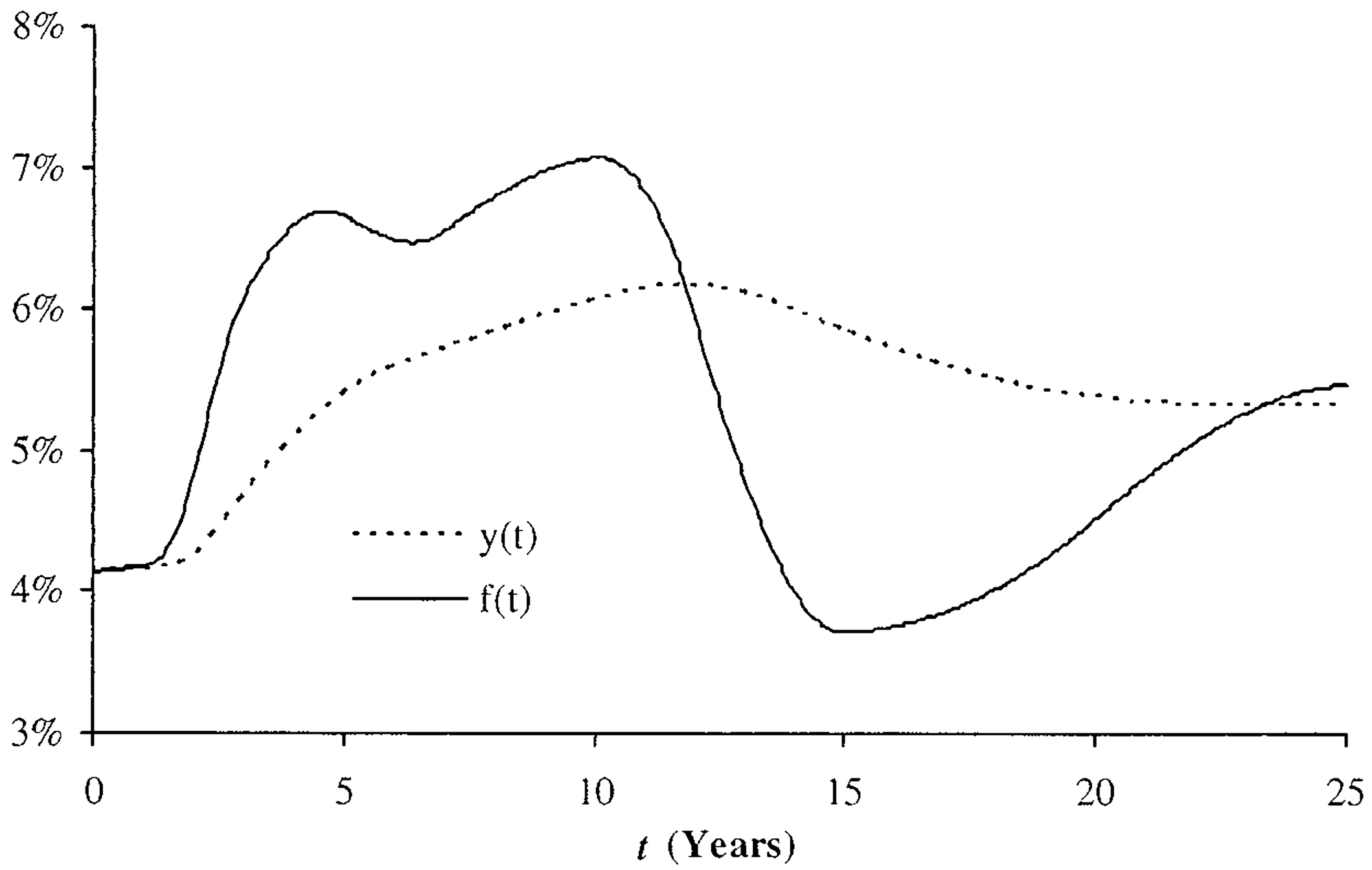
To solve for the correct values of $y_1, y_2, \ldots, y_N$, we iterate on Steps 1–4 using a non-linear root-search algorithm, terminating when (6.4) is satisfied to within acceptable tolerances. The fitting problem is typically good-natured, and virtually all standard root-search packages (see Press 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 $h_i = T_{i+1} - T_i$.

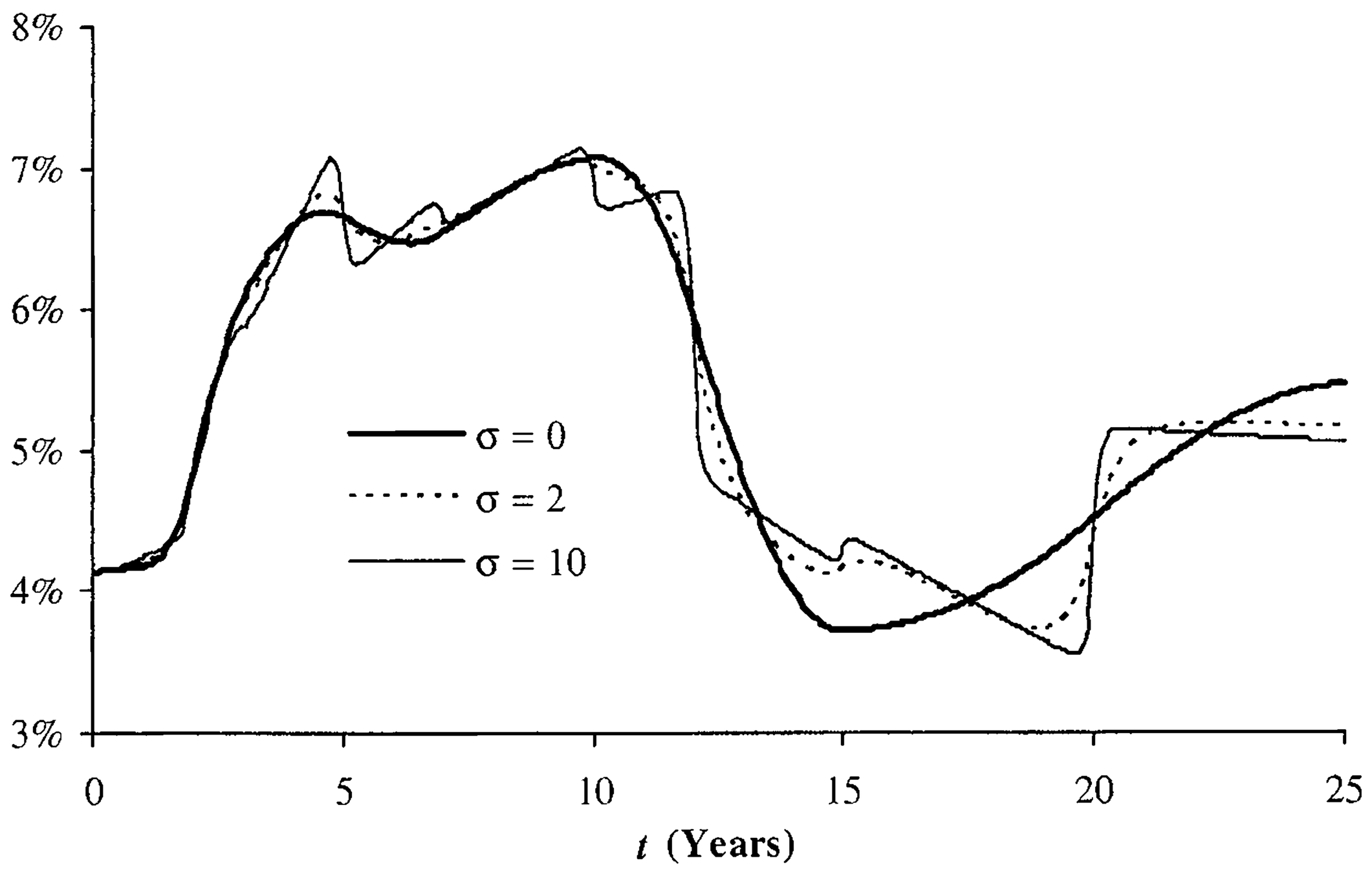
Appendix 6.A.3 discusses a number of properties of tension splines, the most important perhaps being the fact that setting $\sigma = 0$ will recover the ordinary C^2 cubic spline, whereas letting $\sigma \to \infty$ will make the tension spline uniformly approach a linear spline (i.e. the spline we used 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 **M** and the matrices **B** and **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

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

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*

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

and

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

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

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

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

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

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

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

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

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

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

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.

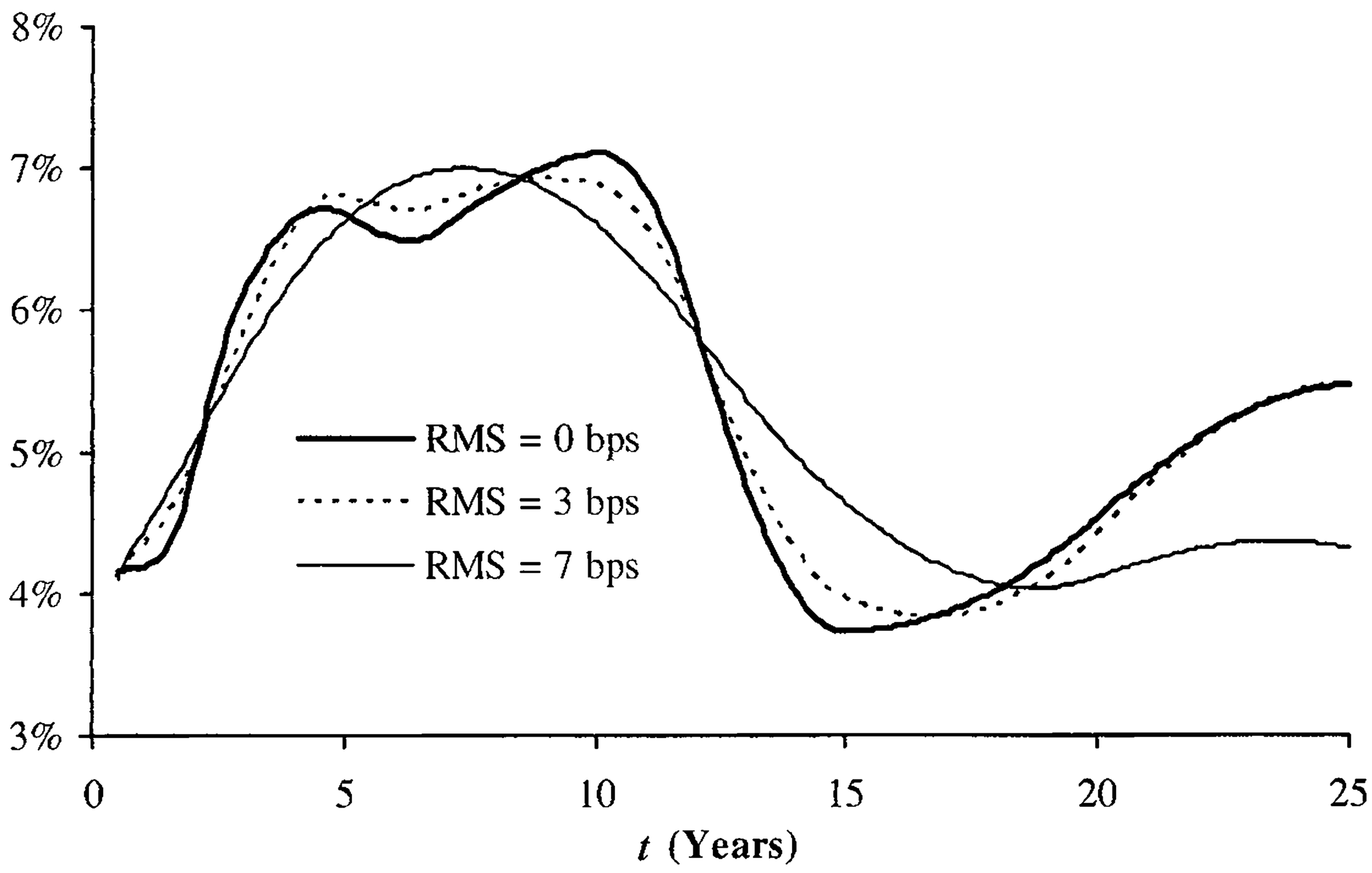
6.3.3 Example

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

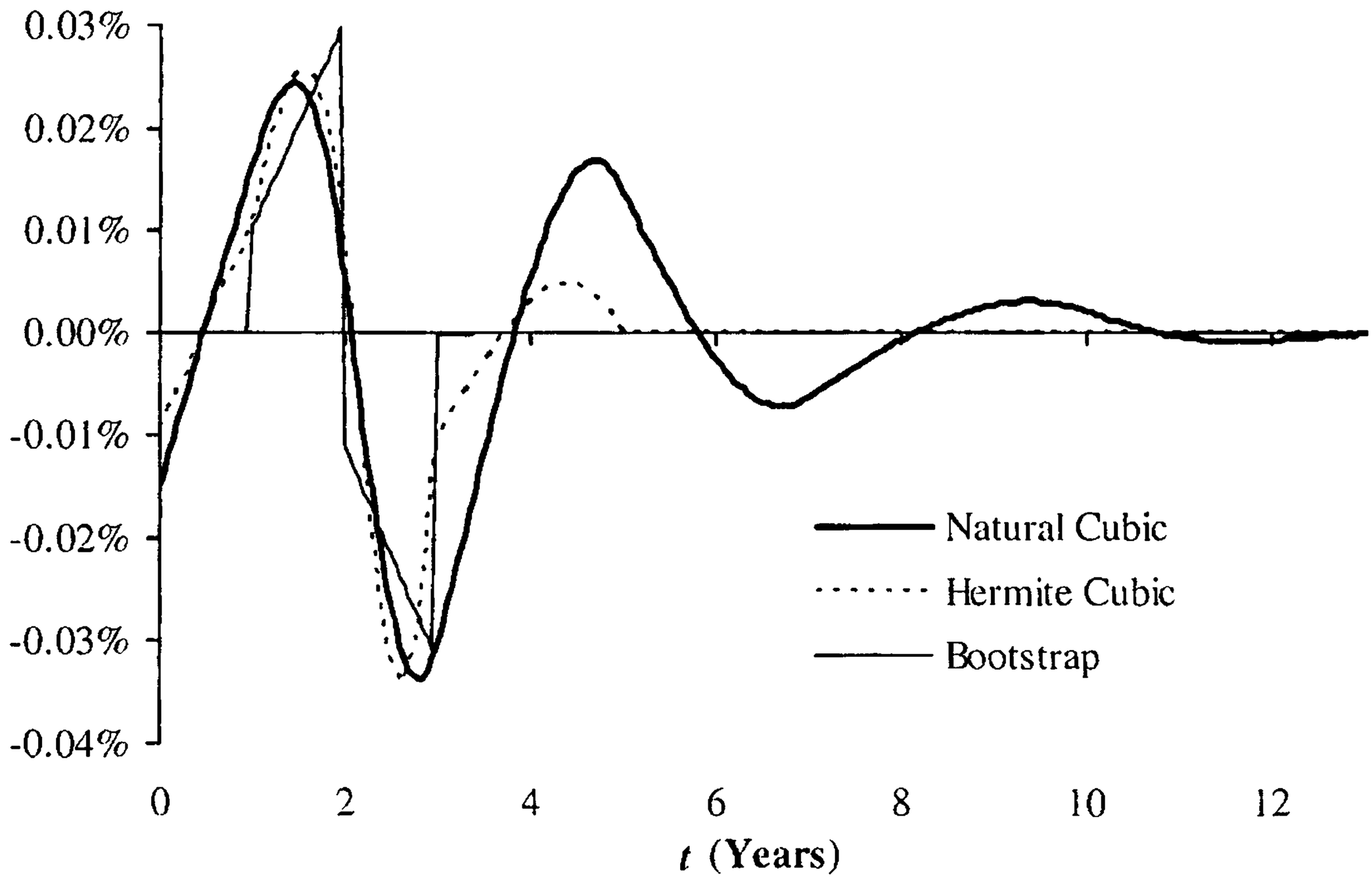
The simplest approach to computation of the delta $\partial V_0/\partial V_i$ involves a manual bump[14] to V_i, followed by a reconstruction of the yield curve, and a subsequent repricing of the portfolio V_0. This procedure is sometimes known as the *par-point approach*, and resulting derivatives *par-point 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.

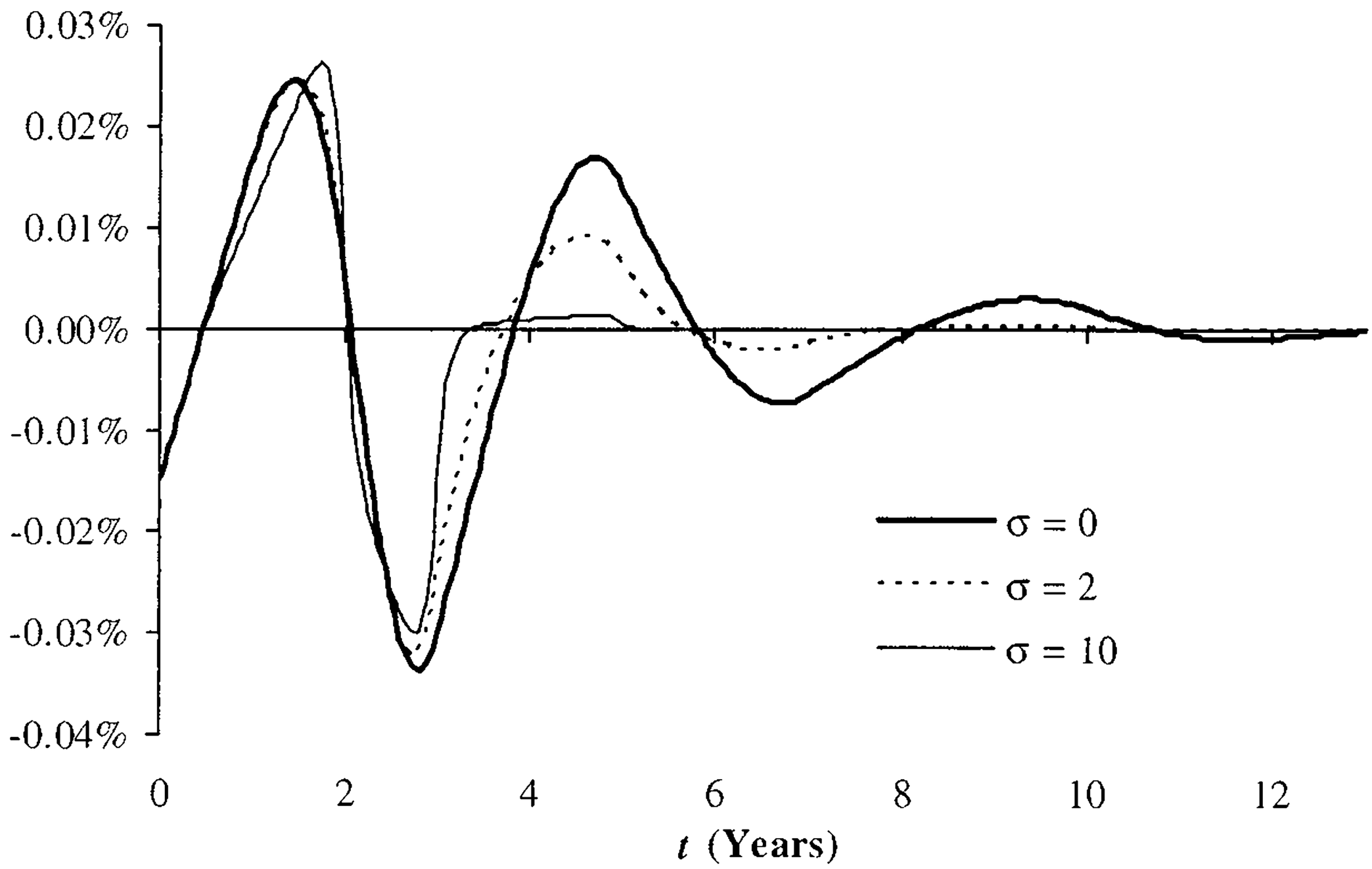
In Figure 6.8, we have followed the recommendations of Section 6.2.4 and added tension to the C^2 spline, causing a dampening of the perturbation noise. Clearly, the usage of a tension factor can have a beneficial impact on risk reports produced by the par-point approach.

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 $L = K$, $\mathbf{W} = \mathbf{1}$, $\mathbf{U} = \mathbf{0}$ 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:

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

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

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

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'' ,

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

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.

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

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*

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

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

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

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*

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

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*

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

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

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,

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

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*

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

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

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

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

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))

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

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

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 $\Omega(\tau,S)$:

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

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.

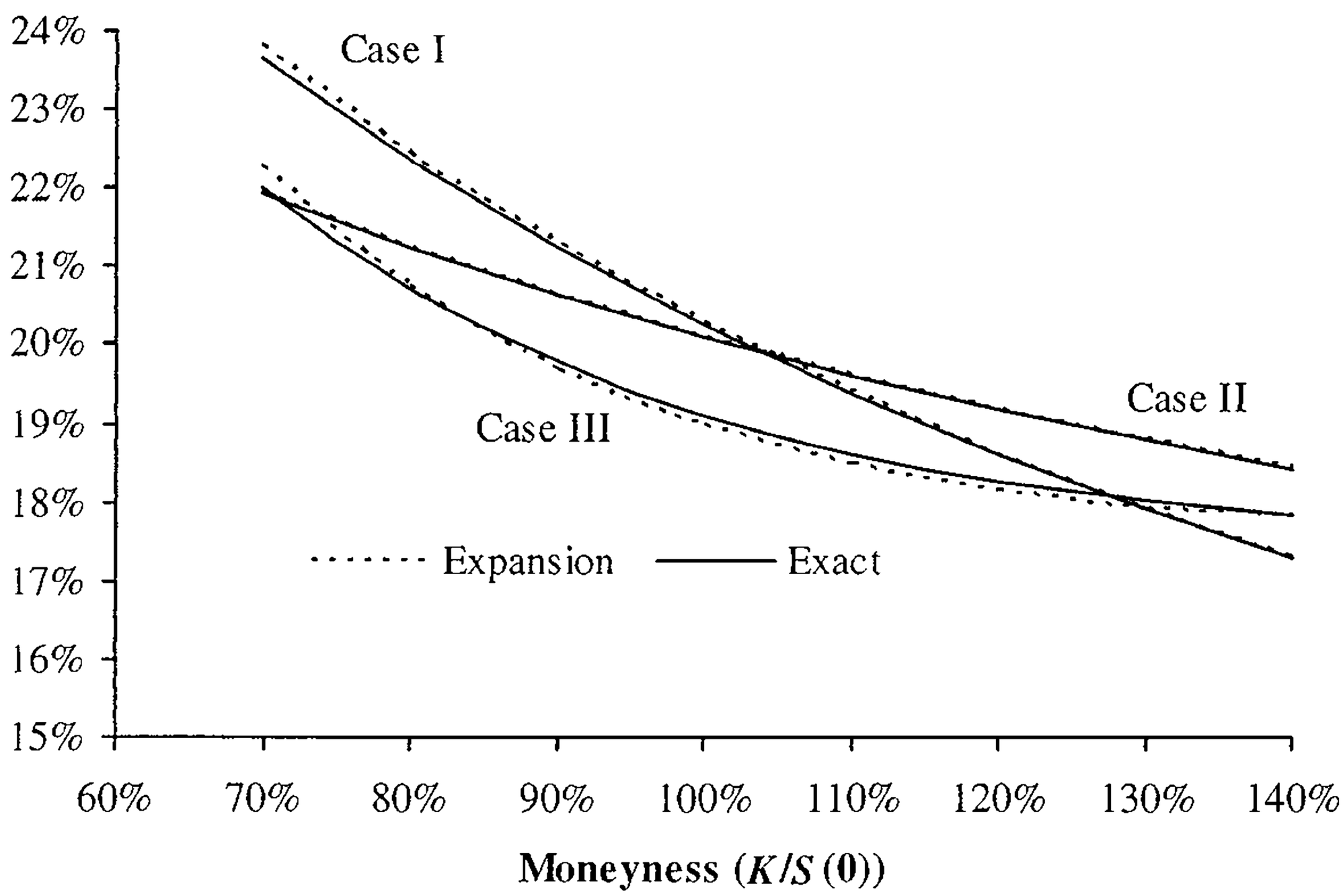
The best choice of the parameters β and ζ is not always obvious. One choice is to use Remark 7.2.14. Alternatively, we could think of a more global approach and, roughly speaking, set them in such a way that the straight line $\beta + \zeta x$ would provide as good a fit to $\varphi(x)$ as possible,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

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

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

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*

$$\Psi(\tau, S) = \Psi_0(S)\tau^{1/2} + \Psi_1(S)\tau^{3/2} + O\left(\tau^{5/2}\right),$$

$$\Psi_0(S) = (S - K)\left(\int_K^S \varphi(u)^{-1}\,du\right)^{-1},$$

$$\Psi_1(S) = -\frac{\Psi_0(S)^3}{(S-K)^2}\ln\left(\Psi_0(S)\left(\varphi(S)\varphi(K)\right)^{-1/2}\right).$$

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

$$\Psi_0(K) = \varphi(K),$$

$$\Psi_1(K) = \frac{1}{24}\varphi(K)\left(2\varphi(K)\varphi''(K) - \varphi'(K)^2\right).$$

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

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

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

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

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* $\bar{b}$ *in (7.62) for the model defined by the time-dependent local volatility function (7.60) is given by*

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

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 $\bar{g}(x)$ given by (7.60) and (7.62), we have

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

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 $\bar{p}$,

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

where $p(\cdot)$ and $\bar{p}$ 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

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

This demonstrates that instantaneous skews $b(t)$ for larger t contribute more to $\bar{b}$ 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

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

Similarly,

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

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*

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

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

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

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

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

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

The (necessary) condition for a minimum is

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

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 $\overline{c}$. 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 $\overline{b} = \overline{b}^*$. When $\overline{b} = \overline{b}^*$, however, we can observe that

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

which suggests the simplified condition[8]

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

[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)

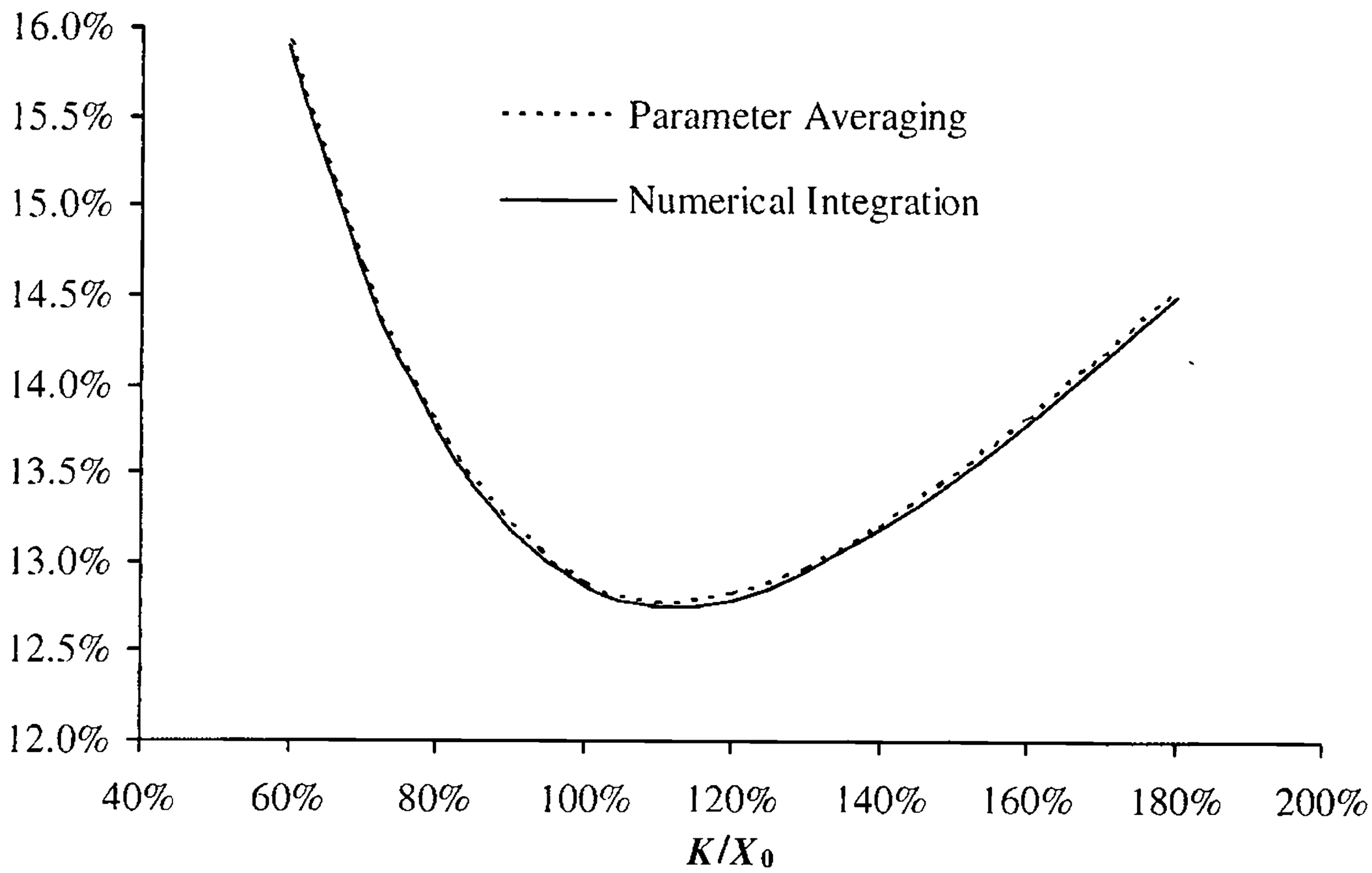
$$C(0, X_0 \colon K, T) = \mathrm{E}\left(C(T', X(T'); K, T)\right).$$

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

$$dS(t) = \lambda\,(bS(t) + (1-b)\,L)\,\sqrt{z(t)}\,dW(t), \tag{8.3}$$

$$dz(t) = \theta\,(z_0 - z(t))\;dt + \eta\sqrt{z(t)}\,dZ(t), \quad z(0) = z_0 = 1, \tag{8.4}$$

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,

$$d\sqrt{z(t)} = O\left(dt\right) + \frac{\eta}{2}\, dZ(t).$$

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

$$dz(t) = \left(\theta\left(z_0 - z(t)\right) + \rho\eta\lambda b u z(t)\right)dt + \eta\sqrt{z(t)}\,d\widetilde{Z}(t), \quad z(0) = z_0, \tag{8.12}$$

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*

$$dz(t) = \widetilde{\theta}\left(t, X(t)\right)\left(\widetilde{f}\left(t, X(t)\right) - z(t)\right) dt + \eta\sqrt{z(t)}\, d\widetilde{Z}(t),$$

where

$$\widetilde{\theta}(t,x) = \theta - \eta\rho\lambda b x \frac{\partial \ln\left(V(t,x)\right)}{\partial x}, \quad \widetilde{f}(t,x) = \frac{\theta z_0}{\widetilde{\theta}(t,x)},$$

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*

$$\beta = 2v/\eta^2 > 0, \quad \alpha = 2\left(\rho\eta\lambda b u - \theta\right)/\eta^2, \quad D = \alpha^2 - 4\beta.$$

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

$$(f_1 * f_2)(k) \triangleq \int_{-\infty}^{\infty} f_1(k-x)\, f_2(x)\, dx$$

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.,

$$(\mathcal{F}f)(\omega) \triangleq \int_{-\infty}^{\infty} e^{i\omega x} f(x)\, dx, \tag{8.14}$$

$$\left(\mathcal{F}^{-1}g\right)(x) \triangleq \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega x} g(\omega)\, d\omega. \tag{8.15}$$

As is well known, the Fourier transform of a convolution is a product of Fourier transforms, so

$$\begin{aligned}&\int_{-\infty}^{\infty}\left[\min\left(e^{-(k-x)},1\right)e^{\alpha(k-x)}\right]\left[e^{\alpha x}p(x)\right]dx\\ &\quad = (f_1 * f_2)(k) = \left(\mathcal{F}^{-1}\left(\mathcal{F}(f_1 * f_2)\right)\right)(k) = \left(\mathcal{F}^{-1}\left(g_1(\omega)\, g_2(\omega)\right)\right)(k),\end{aligned}$$

where

$$\begin{aligned}g_1(\omega) &= \int_{-\infty}^{\infty} e^{i\omega x}\min\left(e^{-x},1\right)e^{\alpha x}\, dx,\\ g_2(\omega) &= \int_{-\infty}^{\infty} e^{i\omega x}e^{\alpha x}p(x)\, dx.\end{aligned}$$

Simple calculations lead to

$$\begin{aligned}g_1(\omega) &= \int_{-\infty}^{0} e^{x(\alpha+i\omega)}dx + \int_{0}^{\infty} e^{x(-1+\alpha+i\omega)}dx\\ &= \frac{1}{\alpha+i\omega} - \frac{1}{\alpha-1+i\omega}\\ &= \frac{1}{(\alpha+i\omega)(1-\alpha-i\omega)},\\ g_2(\omega) &= \chi(\alpha+i\omega),\end{aligned}$$

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

$$\begin{aligned} \mathrm{E}\left((\xi - k)^+\right) &= \chi'(0) - k \\ &\quad - e^{\alpha k} \int_{-\infty}^{\infty} \left[\min\left(-\left(k - x\right), 0\right) e^{-\alpha(k-x)}\right] \left[e^{-\alpha x} p(x)\right] dx, \end{aligned}$$

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*

$$c_{\mathrm{SV}}(0,S;T,K) = \frac{1}{b}c_{\mathrm{B}}(0,S';T,K',\lambda b) - \frac{K'}{2\pi b}\int_{-\infty}^{\infty}\frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)}q(1/2+i\omega)}{\omega^2+1/4}\,d\omega, \quad (8.17)$$

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,

$$\begin{aligned} c_{\mathrm{SV}}(0,S;T,K) &= \mathrm{E}\left(S(T)-K\right)^+ \\ &= \frac{1}{b}\mathrm{E}\left(S'e^{\ln X(T)} - K'\right)^+ \\ &= \frac{S'}{b}\mathrm{E}\left(e^{\ln X(T)} - e^{\ln\left(K'/S'\right)}\right)^+. \end{aligned}$$

By Theorem 8.4.1 and the definition (8.10) of $\Psi_X(u;t)$,

$$c_{\mathrm{SV}}(0,S;T,K) = \frac{1}{b}\left(S' - \frac{K'}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-(\alpha+i\omega)\ln\left(K'/S'\right)}\Psi_X(\alpha+i\omega;T)}{(\alpha+i\omega)(1-\alpha-i\omega)}d\omega\right), \quad (8.19)$$

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

$$\Psi_X^0\left(u;T\right)\triangleq\mathrm{E}\left(e^{u\left(\lambda b\sqrt{z_0}W(T)-\frac{1}{2}\lambda^2b^2z_0T\right)}\right)=e^{\frac{1}{2}\lambda^2b^2z_0T\left(u^2-u\right)}. \tag{8.21}$$

On the other hand,

$$c_{\mathrm{DLN}}\left(0,S;T,K\right)=\frac{1}{b}c_{\mathrm{B}}\left(0,S';T,K',\lambda b\right),$$

so that

$$\frac{1}{b}c_{\mathrm{B}}\left(0,S';T,K',\lambda b\right)$$
$$-\frac{1}{b}\left(S'-\frac{K'}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-(\alpha+i\omega)\ln\left(K'/S'\right)}\Psi_X^0\left(\alpha+i\omega;T\right)}{\left(\alpha+i\omega\right)\left(1-\alpha-i\omega\right)}d\omega\right)=0.$$

Adding the left-hand side of this identity, which is zero, to the right-hand side of (8.19), we obtain

$$c_{\mathrm{SV}}\left(0,S;T,K\right)=\frac{1}{b}c_{\mathrm{B}}\left(0,S';T,K',\lambda b\right)$$
$$-\frac{K'}{2\pi b}\int_{-\infty}^{\infty}\frac{e^{-(\alpha+i\omega)\ln\left(K'/S'\right)}q\left(\alpha+i\omega\right)}{\left(\alpha+i\omega\right)\left(1-\alpha-i\omega\right)}d\omega,$$

where

$$q(u)=\Psi_X\left(u;T\right)-\Psi_X^0\left(u;T\right).$$

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 $\Psi_{\overline{z}}$ 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

$$q_\infty\triangleq\frac{\lambda b z_0}{\eta}\left(\sqrt{1-\rho^2}+i\rho\right)\left(1+\theta T\right). \tag{8.22}$$

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

$$A\left(v\left(\omega\right),u\left(\omega\right)\right)\sim-\frac{\lambda b z_0}{\eta}\theta\left(i\rho+\rho^c\right)T\omega.$$

Hence,

$$\begin{aligned}-\frac{1}{\omega}\ln\left(\Psi_X\left(1/2+i\omega;T\right)\right)&=-\frac{1}{\omega}\ln\left(\Psi_{\overline{z}}\left(v\left(\omega\right),u\left(\omega\right);T\right)\right)\\&=\frac{1}{\omega}\left(A\left(v\left(\omega\right),u\left(\omega\right)\right)+z_0B\left(v\left(\omega\right),u\left(\omega\right)\right)\right)\\&\to\frac{\lambda b z_0}{\eta}\left(\rho^c+i\rho\right)\left(1+\theta T\right)\end{aligned}$$

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,

$$\begin{aligned}
&\left|\int_{\omega_{\max}}^{\infty} \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)} q\left(1/2+i\omega\right)}{\omega^2+1/4}\, d\omega\right| \\
&\quad \le \int_{\omega_{\max}}^{\infty} \left|e^{(1/2+i\omega)\ln\left(S'/K'\right)}\right| \frac{\left|q\left(1/2+i\omega\right)\right|}{\omega^2}\, d\omega \\
&\quad \le \sqrt{\frac{S'}{K'}} e^{-\mathrm{Re}\,(q_\infty)\omega_{\max}} \int_{\omega_{\max}}^{\infty} \frac{d\omega}{\omega^2} \\
&\quad = \sqrt{\frac{S'}{K'}} \frac{e^{-\mathrm{Re}\,(q_\infty)\omega_{\max}}}{\omega_{\max}}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathrm{Re}\left(\int_0^{\infty} \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)} q\left(1/2+i\omega\right)}{\omega^2+1/4}\, d\omega\right) \\
&\quad \approx \frac{\omega_{\max}}{N_\omega} \sum_{n=0}^{N_\omega-1} \frac{e^{\ln\left(S'/K'\right)/2}}{\omega_n^2+1/4} \mathrm{Re}\left(e^{i\omega_n \ln\left(S'/K'\right)} q\left(1/2+i\omega_n\right)\right),
\end{aligned} \tag{8.25}$$

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

$$\int_0^\infty \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)} q(1/2+i\omega)}{\omega^2+1/4}\,d\omega = I_1 + I_2 + I_3, \tag{8.29}$$

where

$$I_1 = \int_0^{\omega'_{\max}} \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)} q(1/2+i\omega)}{\omega^2+1/4}\,d\omega, \tag{8.30}$$

$$I_2 = \int_{\omega'_{\max}}^\infty \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)}}{\omega^2+1/4}\left(q(1/2+i\omega) - e^{r_\infty(\omega)}\right)d\omega, \tag{8.31}$$

$$I_3 = \int_{\omega'_{\max}}^\infty \frac{e^{(1/2+i\omega)\ln\left(S'/K'\right)}}{\omega^2+1/4} e^{r_\infty(\omega)}\,d\omega. \tag{8.32}$$

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 $1/\sigma$.

$$
\begin{aligned}
I_2 &\approx e^{q_0}\sqrt{\frac{S'}{K'}}\sum_{n=0}^{N''_\omega-1}\frac{e^{r(\omega''_n)-r_\infty(\omega''_n)}-1}{(\omega''_n)^2+1/4}\int_{\omega''_n}^{\omega''_{n+1}} e^{\omega(-q_\infty+i\ln(S'/K'))}\,d\omega \\
&= e^{q_0}\sqrt{\frac{S'}{K'}}\sum_{n=0}^{N''_\omega-1}\frac{e^{r(\omega''_n)-r_\infty(\omega''_n)}-1}{(\omega''_n)^2+1/4} \\
&\quad\times\frac{e^{\omega''_{n+1}(-q_\infty+i\ln(S'/K'))}-e^{\omega''_n(-q_\infty+i\ln(S'/K'))}}{-q_\infty+i\ln(S'/K')}.
\end{aligned}
\tag{8.33}
$$

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

$$bK_n + (1-b)\,L = (bS + (1-b)\,L)\,e^{x_n},$$

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

$$\begin{aligned}\mathrm{E}\left(f\left(S(T)\right)\right) \approx f\left(S(0)\right) &+ \sum_{n=-N}^{-1} p_{\mathrm{SV}}\left(0,S;T,K_n\right) f''\left(K_n\right)\left(K_{n+1}-K_n\right) \\ &+ \sum_{n=0}^{N-1} c_{\mathrm{SV}}\left(0,S;T,K_n\right) f''\left(K_n\right)\left(K_{n+1}-K_n\right),\end{aligned}$$

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 $(\mathcal{F}\widehat{g})$ 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

$$\mathrm{E}\left(\overline{z}(T)^n\right) = \left.\frac{d^n}{du^n}\Psi_{\overline{z}}\left(u, 0; T\right)\right|_{u=0}, \quad n = 1, 2, \ldots .$$

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* $u \geq 0$,

$$\mathrm{E}\left(S(T)^u\right) < \infty.$$

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 $\eta = 2\nu$,

$$\begin{aligned} dS(t) &= \lambda S(t)^c \sqrt{z(t)}\, dW(t), \\ dz(t) &= \frac{1}{4}\eta^2 z(t)\, dt + \eta z(t)\, dZ(t). \end{aligned}$$

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.

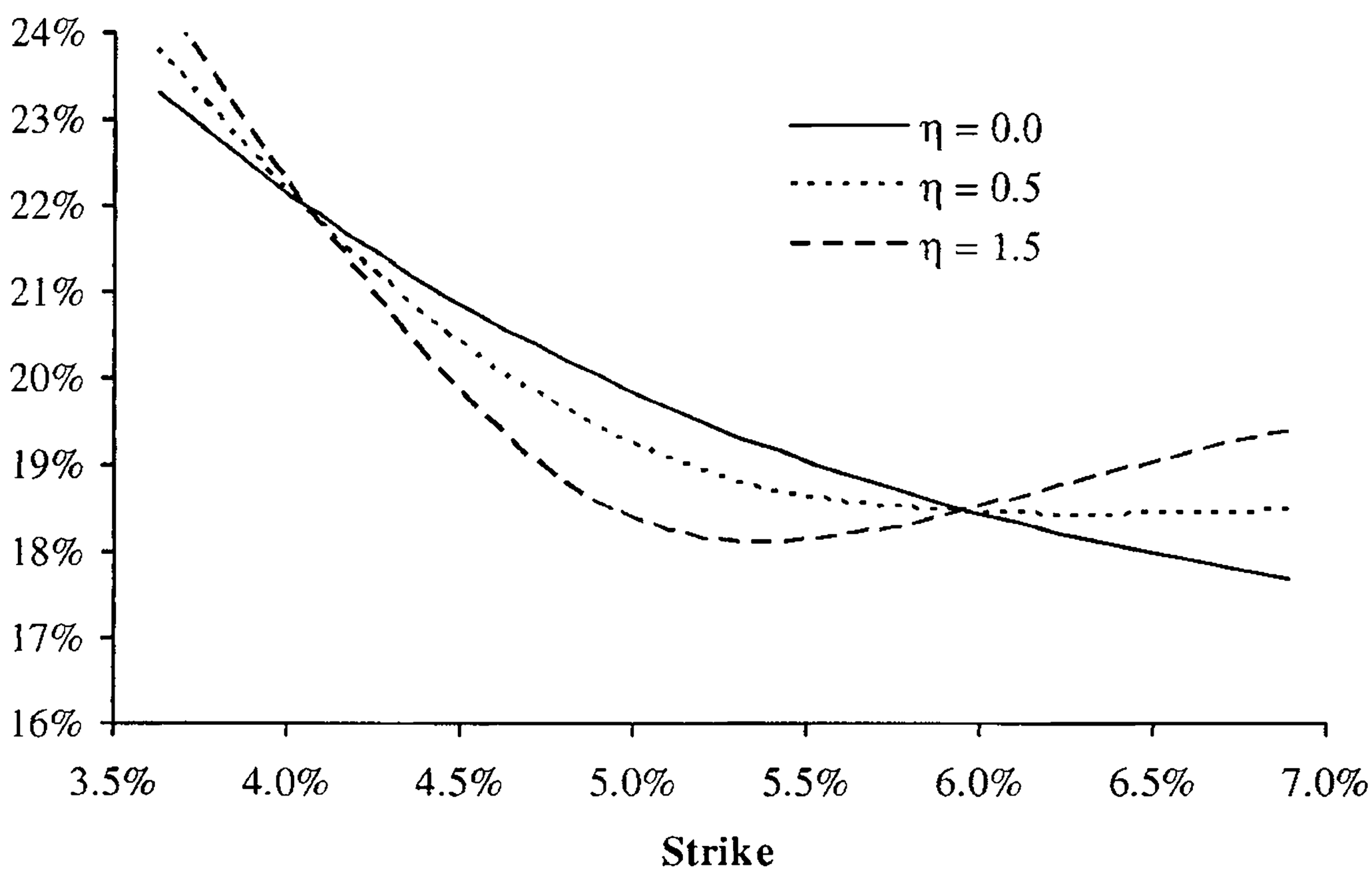
$$\sigma_{\mathrm{B}}(0, S(0); T, K) = \sigma_{\mathrm{ATM}} + R \cdot \chi + \frac{1}{2} B \cdot \chi^2 + \ldots$$

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

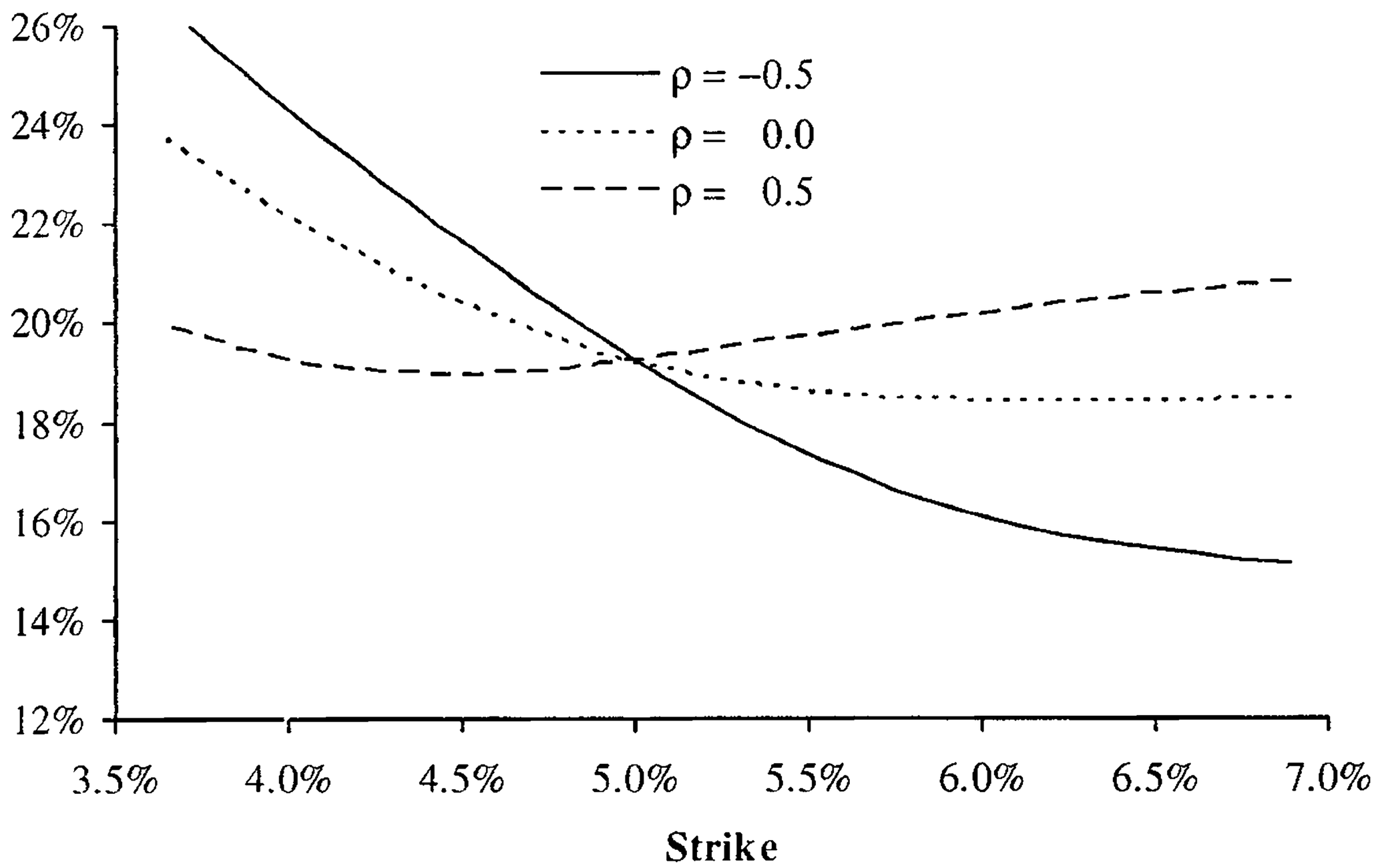


Notes: Implied volatility smile for SV model with $T = 1$, $S(0) = L = 5\%$, $z_0 = 1$, $b = 0.1$, $\lambda = 20\%$, $\theta = 0.1$, and $\rho = 0$. The volatility of variance parameter η varies as shown in the graph.

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



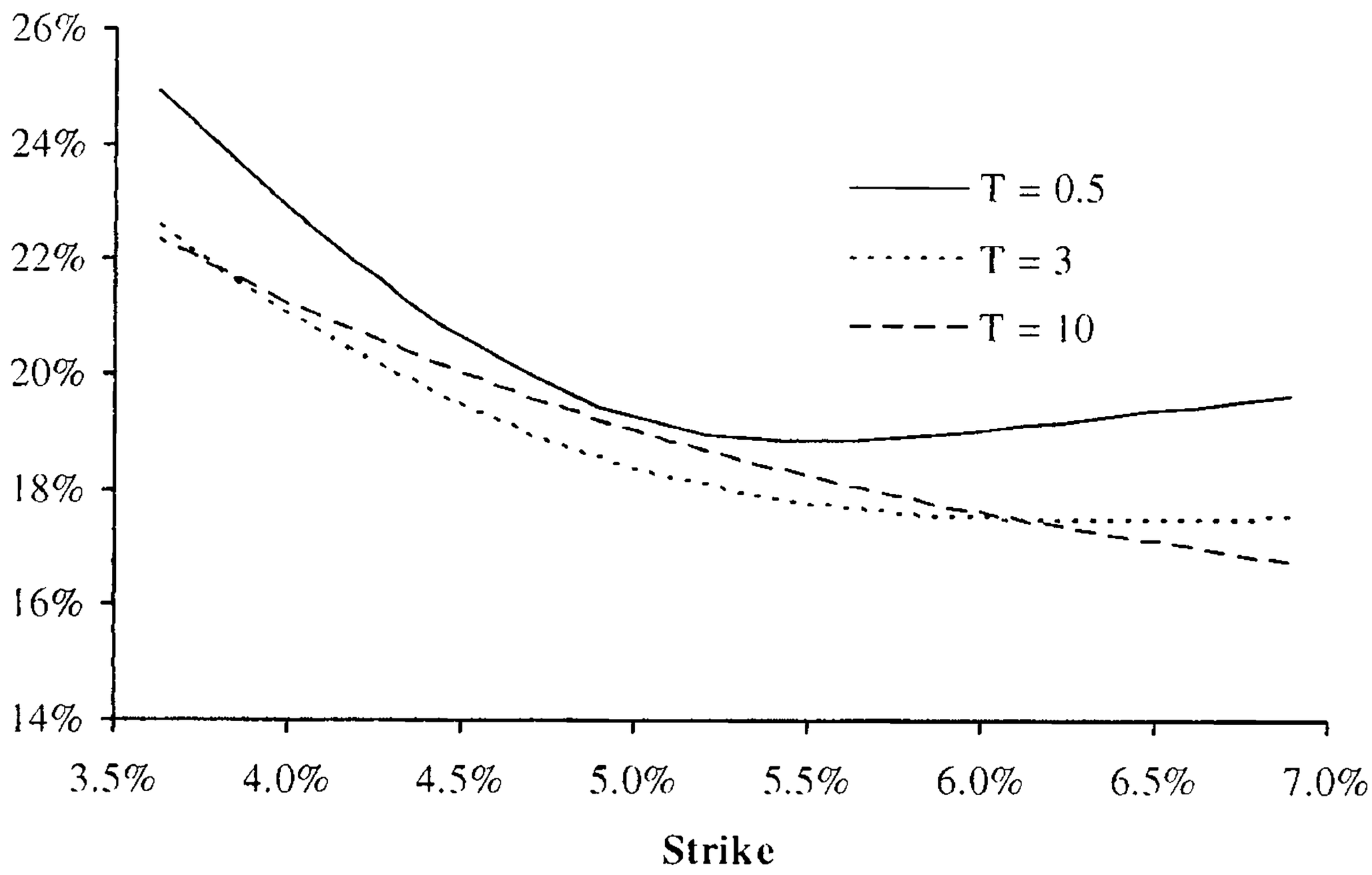
Notes: Implied volatility smile for SV model with $T = 1$, $S(0) = L = 5\%$, $z_0 = 1$, $b = 0.1$, $\lambda = 20\%$, $\theta = 0.1$, and $\eta = 1$. The correlation parameter ρ varies as shown in the graph.

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.

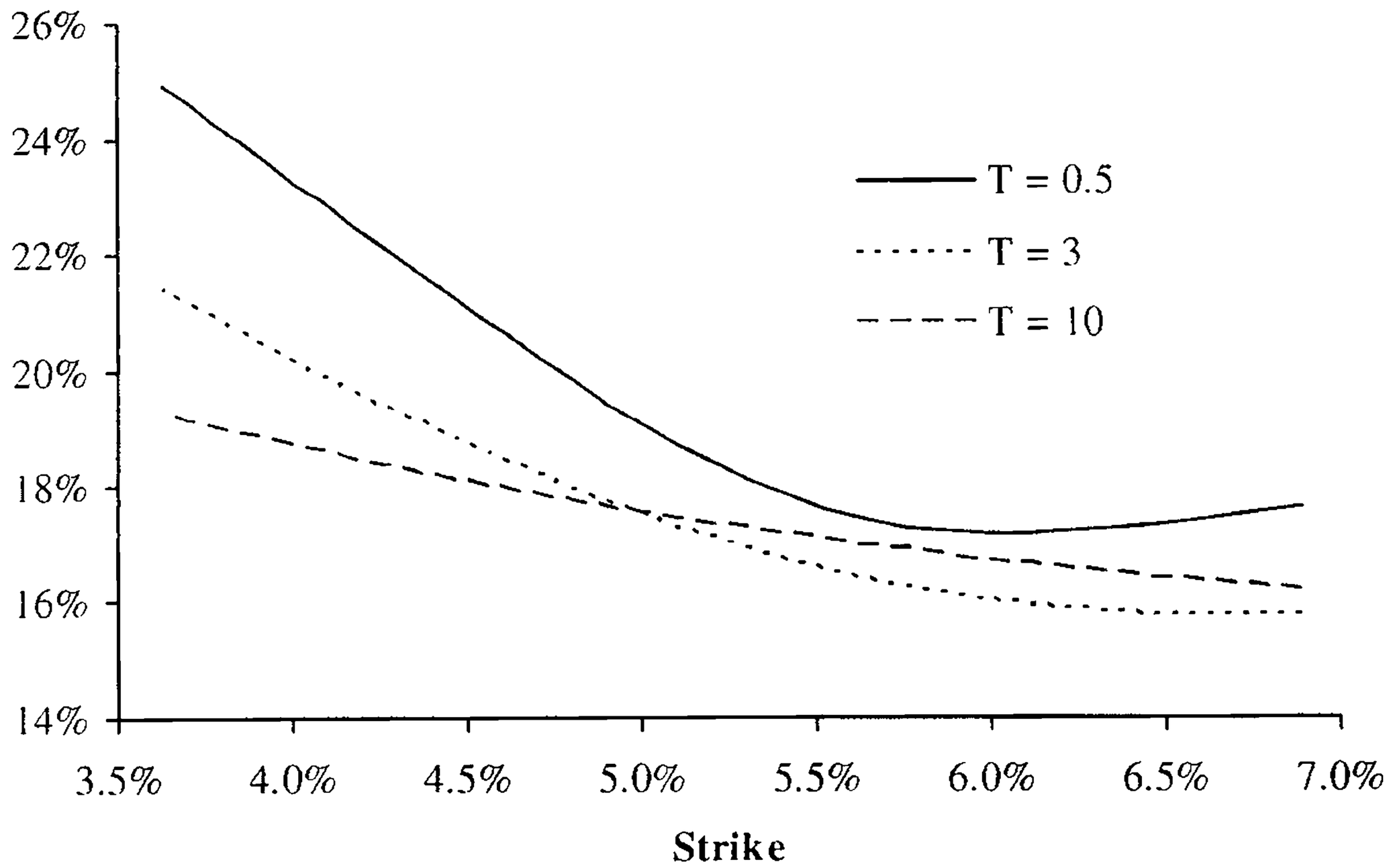
We note in passing that the ATM volatility of a constant parameter SV model is not a monotonic function of option maturity, as a quick glance at Figure 8.3 will confirm. For an analysis of the ATM volatility level and its dependence on maturity, see Lewis [2000].

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,

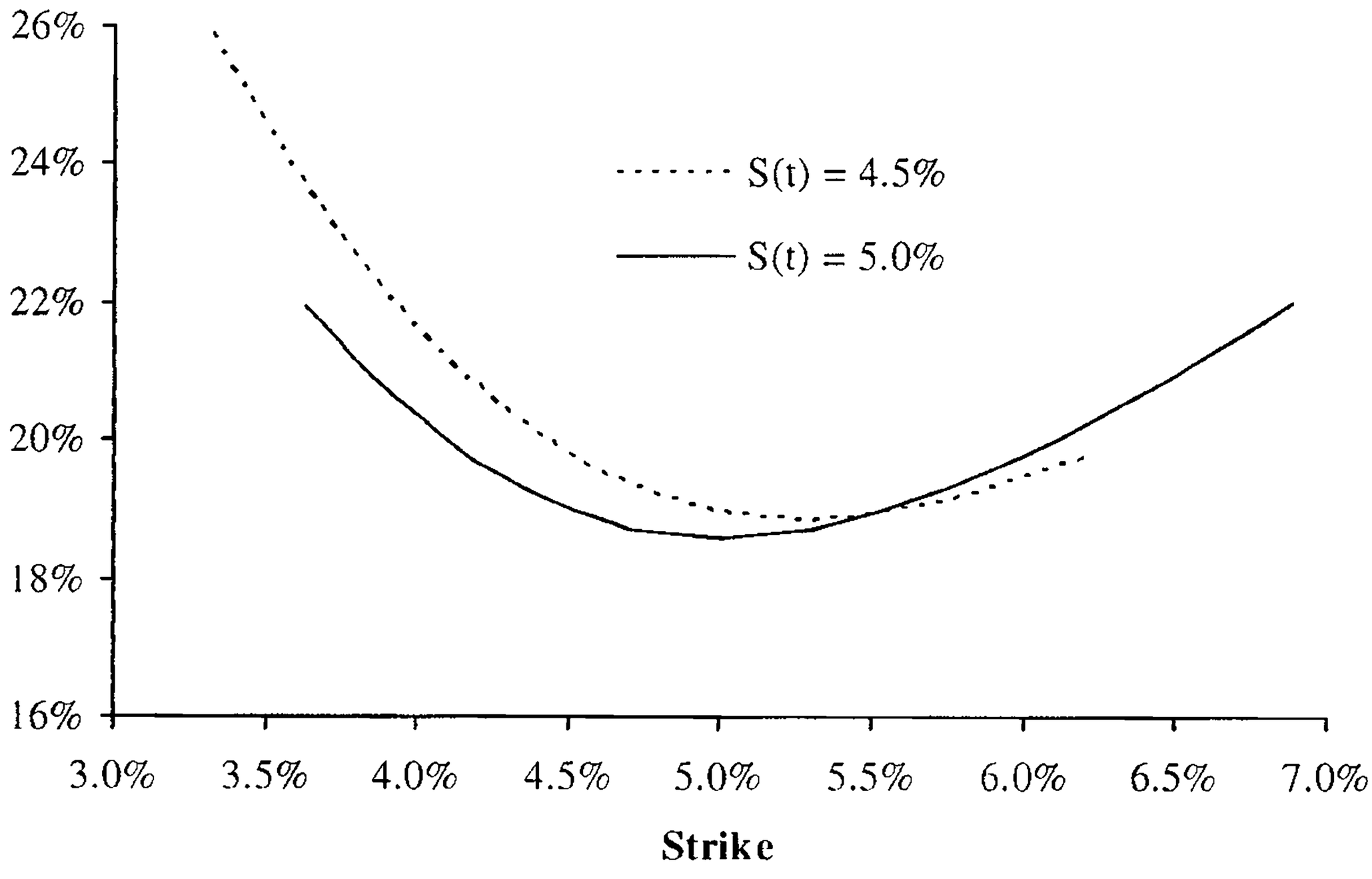
$$dS(t) = \lambda \left((1-b)L + bS(t) + \frac{1}{2}c(S(t)-L)^2 \right) dW(t). \tag{8.47}$$

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

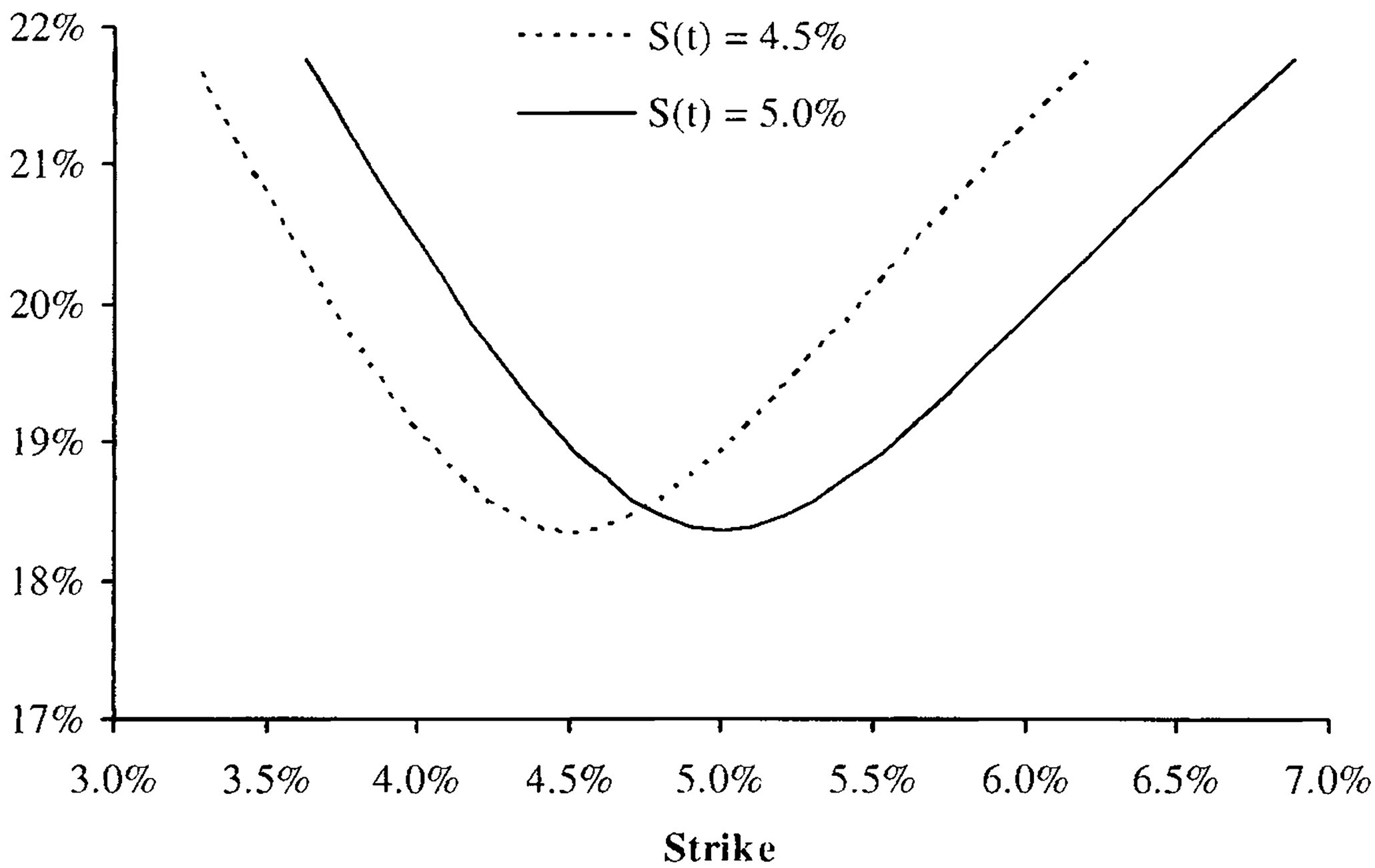


Notes: Time t implied volatility smile for quadratic LV model with $T = t + 1$, $S(0) = L = 5\%$, $b = 1$, $\lambda = 18\%$, and $c = 0.6$. Two different values for the forward rate $S(t)$ are used, as indicated in the graph.

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



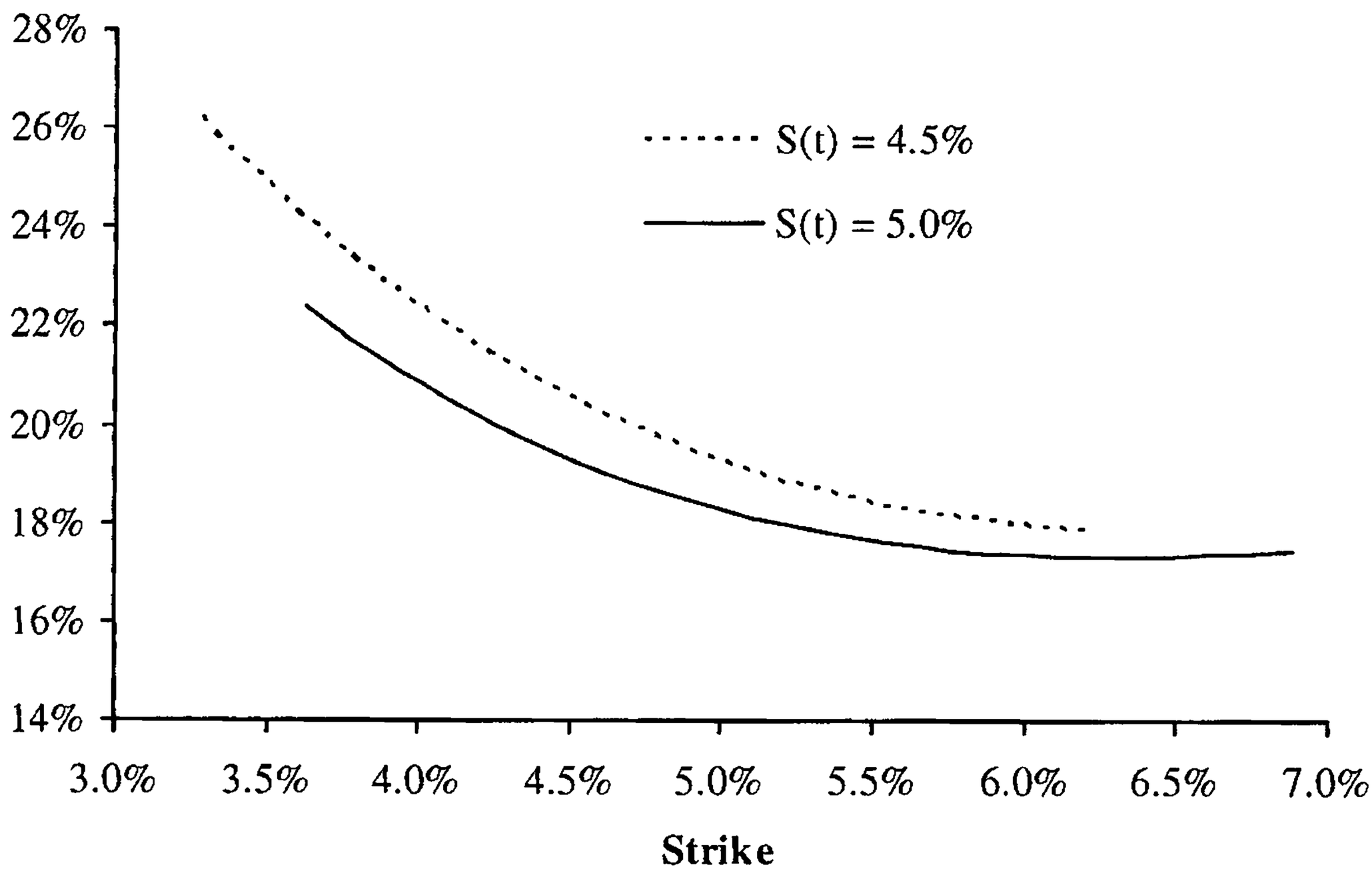
Notes: Time t implied volatility smile for SV model with $T = t + 1$, $S(0) = 5\%$, $z(t) = z_0 = 1$, $b = 1$, $\lambda = 20\%$, $\theta = 0.1$, $\rho = 0$, and $\eta = 1.5$. Two different values for the forward rate $S(t)$ are used, as indicated in the graph.

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

$$\mathrm{E}\left(z(t)|S(t)\right)$$

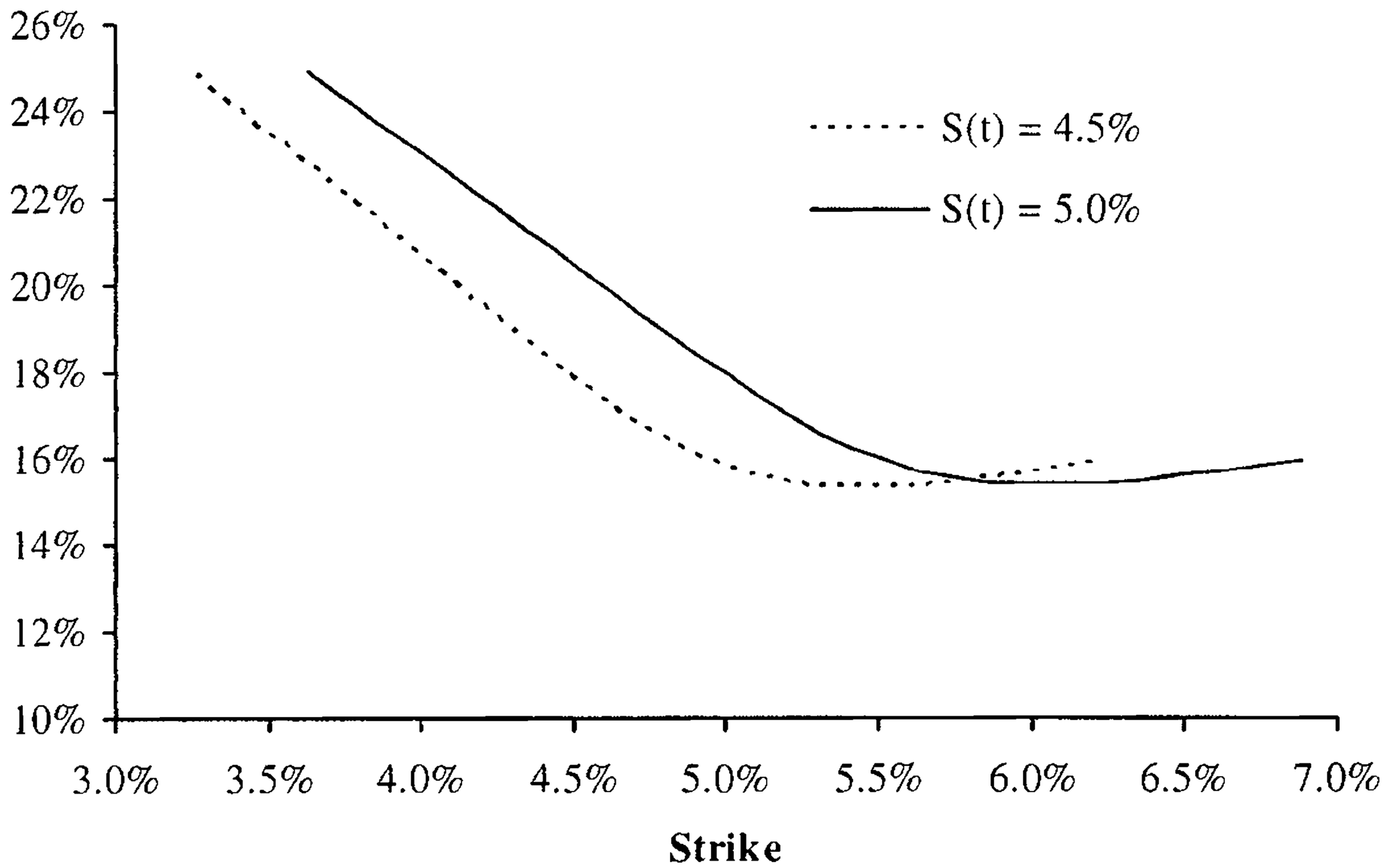
which we here compute by a simple Gaussian approximation that ignores mean reversion,

$$\mathrm{E}\left(z(t)|S(t)\right) \approx z_0 + \frac{\eta\rho}{\lambda}\frac{S(t) - S(0)}{S(0)}. \tag{8.48}$$

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.

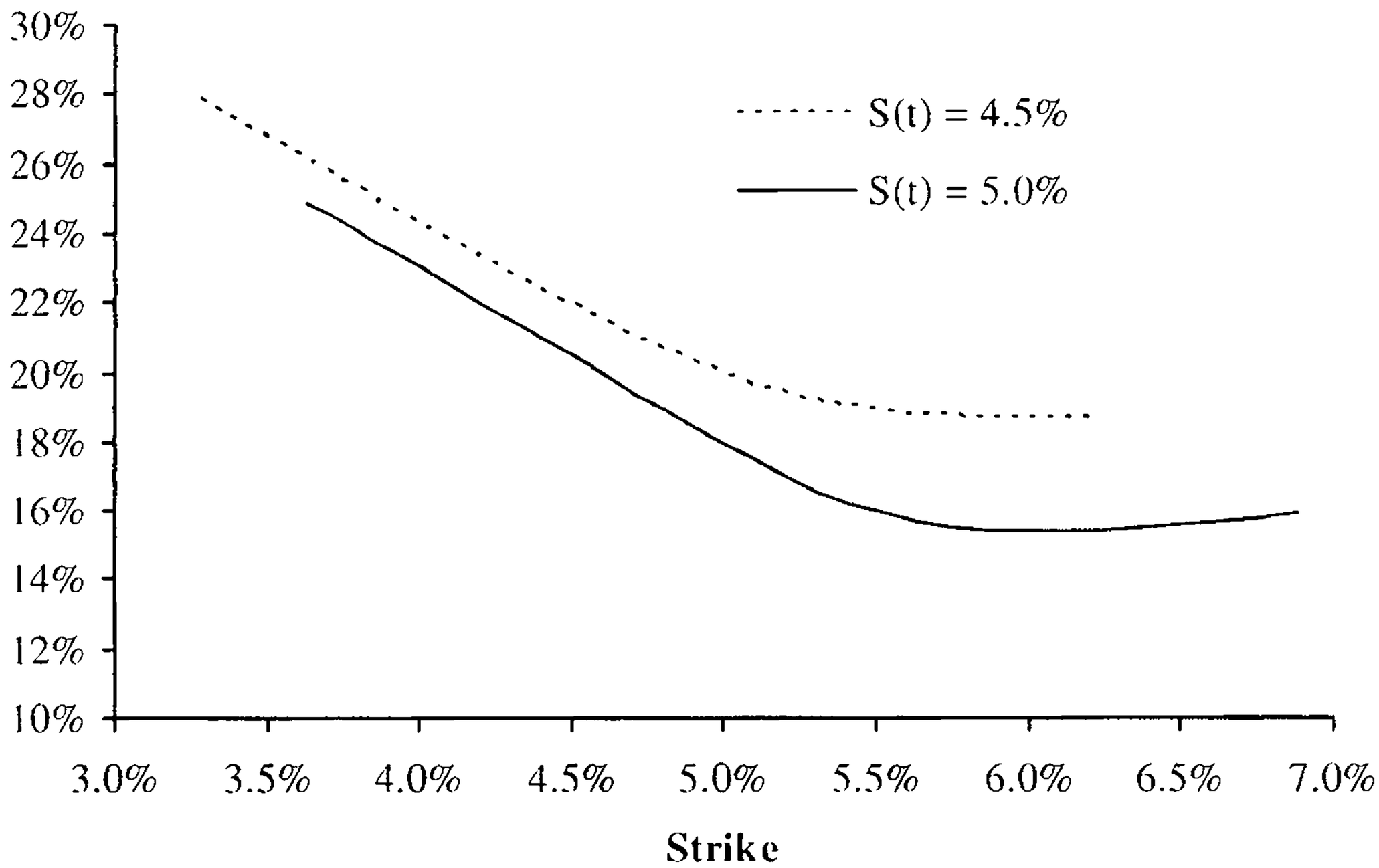
8.9 Hedging in Stochastic Volatility Models

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

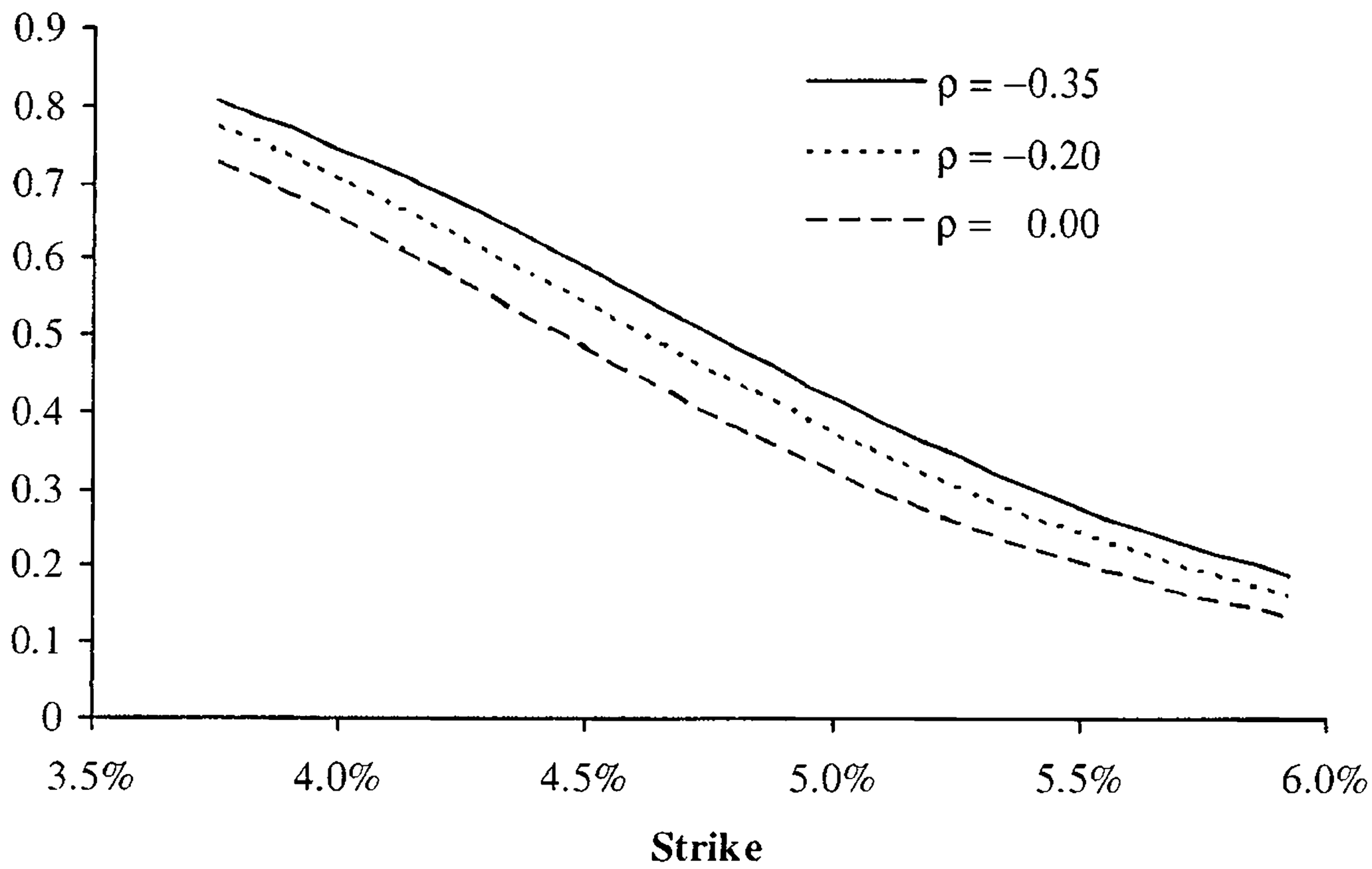
[16] Capital Asset Pricing Model, see Sharpe [1964].

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

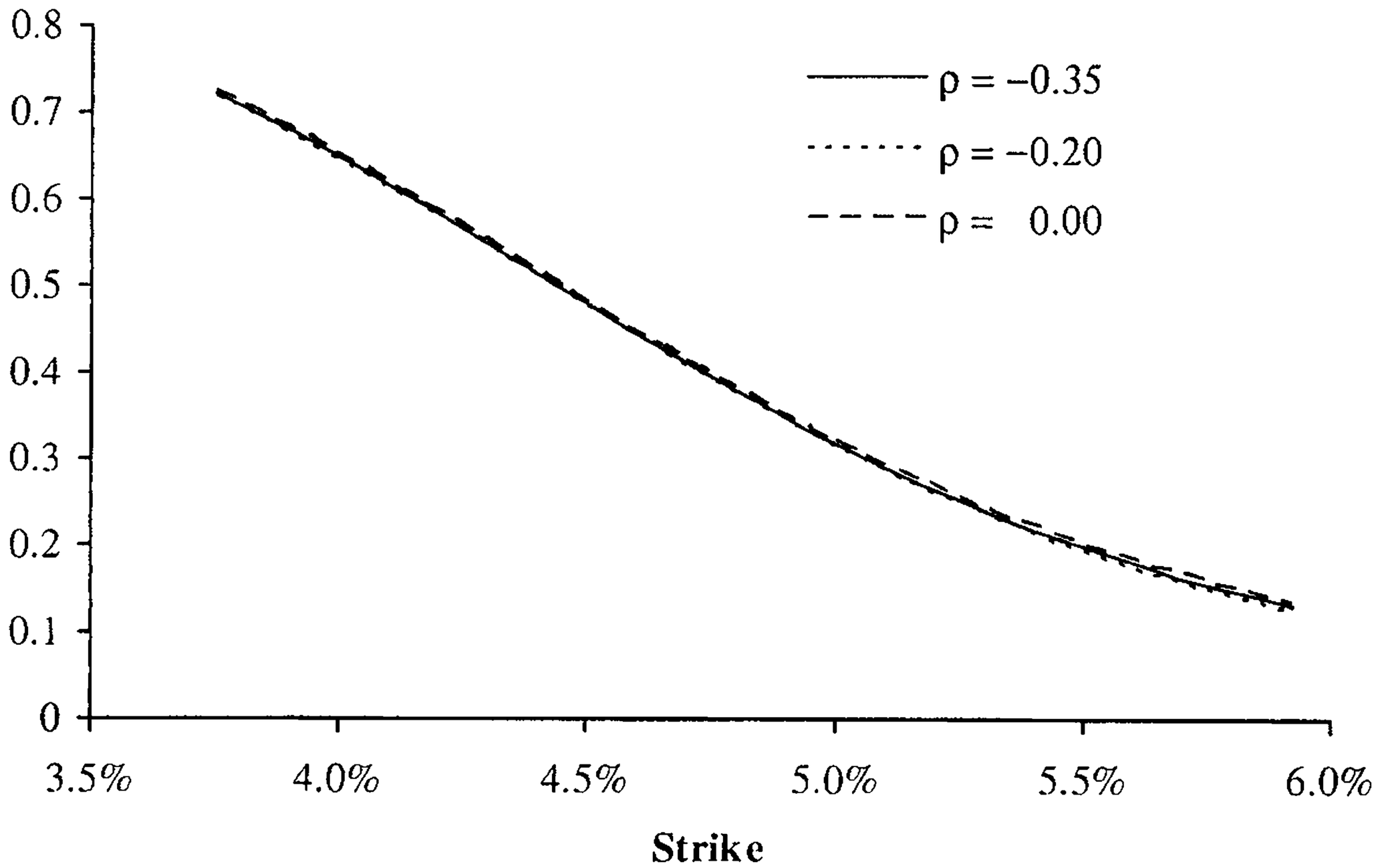
90% at $\rho = -0.3$, with other parameters being quite stable across different correlation choices. Figures 8.10 and 8.11 show the pure delta $\partial V/\partial S$ and the minimum variance delta π_{mv} for selected strikes and correlations. Clearly, the MV delta is here virtually independent of the 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),

$$dS(t) = \lambda\,(bS(t) + (1-b)\,L)\,\sqrt{z(t)}\,dW(t), \tag{8.52}$$

$$dz(t) = \theta\,(z_0 - z(t))\,dt + \eta z(t)^p\,dZ(t), \tag{8.53}$$

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

$$dS(t) = \lambda(t)\,(bS(t) + (1-b)\,L)\,\sqrt{z(t)}\,dW(t), \tag{9.1}$$

$$dz(t) = \theta\,(z_0 - z(t))\;dt + \eta\sqrt{z(t)}\,dZ(t), \tag{9.2}$$

[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

$$dz(t) = (\theta(z_0 - z(t)) + \rho\eta\lambda(t)buz(t))\, dt + \eta\sqrt{z(t)}\, d\widetilde{Z}(t), \quad z(0) = z_0, \tag{9.4}$$

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*

$$\frac{d}{dt}A(t,T) + \theta z_0 B(t,T) = 0, \tag{9.5}$$

$$\frac{d}{dt}B(t,T) - (\theta - \rho\eta bu\lambda(t))\, B(t,T) + \frac{\eta^2}{2}B(t,T)^2 + v\lambda(t)^2 = 0, \tag{9.6}$$

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,

$$\Psi_{\overline{z\lambda^2}}(v,u;T) = G(0,z_0).$$

On the other hand, by the Feynman-Kac formula, $G(t,z)$ satisfies the following PDE,

$$\frac{\partial}{\partial t}G(t,z) + (\theta z_0 - (\theta - \rho\eta b u\lambda(t))\,z)\frac{\partial}{\partial z}G(t,z) + \frac{\eta^2}{2}z\frac{\partial^2}{\partial z^2}G(t,z) + v\lambda(t)^2 zG(t,z) = 0, \quad (9.7)$$

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

$$\frac{\partial l}{\partial t} + \theta(z_0 - z)\frac{\partial l}{\partial z} + \frac{\eta^2}{2}\psi(z)^2 \left\{ \frac{\partial^2 l}{\partial z^2} + l u^2 p(t)^2 + 2up(t)\frac{\partial l}{\partial z} \right\} = 0, \tag{9.13}$$

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

$$\begin{aligned}\mathrm{E}\left(f\left(S(T)\right)\right) &= g\left(0, S(0); \overline{v}\right) \\ &\quad + \sum_{n=1}^{\infty} \frac{1}{n!T^n} \left.\frac{\partial^n g}{\partial v^n}\right|_{v=\overline{v}} \mathrm{E}\left(\left(\overline{z\lambda^2}(T) - \mu_{\overline{z\lambda^2}}(0, z_0)\right)^n\right),\end{aligned}$$

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 $O(\eta^4)$*

$$\begin{aligned}\mathrm{E}\left(f\left(S(T)\right)\right) &= g\left(0, S(0); \overline{v}\right) + T^{-2}\left(\eta^2 l_{1,2} + \eta^4 l_{2,2}\right)\frac{\partial^2 g}{\partial v^2} \\ &\quad - \eta^4 T^{-3} l_{2,3} \frac{\partial^3 g}{\partial v^3} + \frac{1}{2}\eta^4 T^{-4} l_{1,2}^2 \frac{\partial^4 g}{\partial v^4},\end{aligned}$$

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*

$$q_1 = \mu_{\overline{z\lambda^2}}(0, z_0)/T + \alpha_0\eta^2 + \alpha_1\eta^2 k^2 + O\left(\eta^4\right), \tag{9.16}$$

$$\begin{aligned}q_2 &= \mu_{\overline{z\lambda^2}}(0, z_0)/T + \left(\alpha_0\eta^2 + \beta_0\eta^4\right) \\ &\quad + \left(\alpha_1\eta^2 + \beta_1\eta^4\right)k^2 + \beta_2\eta^4 k^4 e^{-\Lambda\eta^2 k^2} + O\left(\eta^6\right),\end{aligned} \tag{9.17}$$

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 $\overline{z\lambda^2}(T)$,

$$h(\zeta_T) = a + be^{c\zeta_T}, \quad h'(\zeta_T) = bce^{c\zeta_T}, \quad h''(\zeta_T) = bc^2e^{c\zeta_T}, \tag{9.26}$$

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

$$c = \frac{h''(\zeta_T)}{h'(\zeta_T)}, \tag{9.27}$$

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*

$$\Psi_{\overline{z}}\left(\frac{h''(\zeta_T)}{h'(\zeta_T)}\overline{\lambda}^2, 0; T\right) = \Psi_{\overline{z\lambda^2}}\left(\frac{h''(\zeta_T)}{h'(\zeta_T)}, 0; T\right), \tag{9.29}$$

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 $\bar{\lambda}^2$ 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 $\bar{\lambda}$ 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 $\bar{b}$ 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

$$\bar{b} = \int_0^T b(t)w_T(t)\,dt, \tag{9.32}$$

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 $\bar{b}$ for the equation*

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

with $z(t)$ following (9.34) over a time horizon $[0, T]$ is given by

$$\bar{b} = \int_0^T b(t) w_T(t)\, dt, \tag{9.35}$$

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

$$
\begin{aligned}
&\mathrm{E}\left(\left(\int_0^T \lambda(t)^2 z(t)\,dt\right)^2\right)\\
&\quad = 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2\mathrm{E}(z(t)z(s))\\
&\quad = 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\mathrm{E}(z(s)^2)\\
&\qquad + 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2(1-e^{-\theta(t-s)})z_0\mathrm{E}(z(s)).
\end{aligned}
$$

Using (9.37) for $\mathrm{E}(z(s)^2)$ we get

$$
\begin{aligned}
\mathrm{E}\left(\left(\int_0^T \lambda(t)^2 z(t)\,dt\right)^2\right) &= 2z_0^2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\\
&\quad + 2z_0\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\int_0^s \eta(r)^2 e^{-2\theta(s-r)}dr\\
&\qquad + 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2\left(1-e^{-\theta(t-s)}\right)z_0\mathrm{E}\left(z(s)\right).
\end{aligned}
$$

Changing the order of integration for the second term, we obtain

$$
\begin{aligned}
\mathrm{E}\left(\left(\int_0^T \lambda(t)^2 z(t)\,dt\right)^2\right) &= 2z_0^2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}\\
&\quad + 2z_0\int_0^T dr\,\eta(r)^2\int_r^T ds\int_s^T dt\,\lambda(t)^2\lambda(s)^2 e^{-\theta(t-s)}e^{-2\theta(s-r)}\\
&\qquad + 2\int_0^T dt \int_0^t ds\,\lambda(t)^2\lambda(s)^2\left(1-e^{-\theta(t-s)}\right)z_0\mathrm{E}\left(z(s)\right).
\end{aligned}
$$

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

$$c_{n,m} = c_{n,m}(\mathcal{X}),$$

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

$$\{\lambda(\cdot), b(\cdot), \eta(\cdot)\} = \operatorname{argmin} \sum_{n,m} \left(c_{n,m}(\mathcal{X}) - \widehat{c}_{n,m}\right)^2, \tag{9.41}$$

[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

$$dS(t) = \widehat{\lambda}_n \left(\widehat{b}_n S(t) + \left(1 - \widehat{b}_n\right) L\right) \sqrt{z(t)}\, dW(t), \tag{9.42}$$

$$dz(t) = \theta\left(z_0 - z(t)\right) dt + \widehat{\eta}_n \sqrt{z(t)}\, dZ(t). \tag{9.43}$$

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

$$\left\{\overline{\lambda}_n\left(\mathcal{X}\right), \overline{b}_n\left(\mathcal{X}\right), \overline{\eta}_n\left(\mathcal{X}\right)\right\}$$

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,

$$\{\lambda(\cdot),\ b(\cdot), \eta(\cdot)\} = \operatorname{argmin}\left(W_\lambda \sum_n \left(\overline{\lambda}_n\left(\mathcal{X}\right) - \widehat{\lambda}_n\right)^2 \right.$$
$$\left. + W_b \sum_n \left(\overline{b}_n\left(\mathcal{X}\right) - \widehat{b}_n\right)^2 + W_\eta \sum_n \left(\overline{\eta}_n\left(\mathcal{X}\right) - \widehat{\eta}_n\right)^2\right), \tag{9.44}$$

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:

$$\sum_{i=1}^{n} \overline{\rho}_{n,i}\left(\lambda(\cdot)\right)(T_i - T_{i-1})\eta_i^2 = \left(\widehat{\eta}_n\right)^2, \tag{9.48}$$

$$\sum_{i=1}^{n} w_{n,i}\left(\lambda(\cdot),\eta(\cdot)\right)(T_i - T_{i-1})b_i = \widehat{b}_n, \tag{9.49}$$

$$F\left(\lambda(\cdot);\overline{b}_n\left(\mathcal{X}\right),\overline{\eta}_n\left(\mathcal{X}\right)\right) = \widehat{\lambda}_n, \tag{9.50}$$

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

$$\sum_{i=1}^{n} \overline{\rho}_{n,i}\left(\lambda^*(\cdot)\right)(T_i - T_{i-1})\eta_i^2 = \left(\widehat{\eta}_n\right)^2, \quad n = 1,\ldots,N.$$

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

$$dS(t) = \varphi(t, S(t)) \sqrt{z(t)}\, dW(t), \tag{9.54}$$

$$dz(t) = \theta(z_0 - z(t))\, dt + \eta(t)\psi(z(t))\, dZ(t), \tag{9.55}$$

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

$$\begin{aligned} 0 = {} & \frac{\partial}{\partial t} V(t, z, S) + \theta(z_0 - z) \frac{\partial}{\partial z} V(t, z, S) + \frac{\eta(t)^2}{2} \psi(z)^2 \frac{\partial^2}{\partial z^2} V(t, z, S) \\ & + \frac{z}{2} \varphi(t, S)^2 \frac{\partial^2}{\partial S^2} V(t, z, S) + \rho\eta(t)\psi(z)\sqrt{z}\varphi(t, S) \frac{\partial^2}{\partial z \partial S} V(t, z, S). \end{aligned} \tag{9.56}$$

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

$$\mathrm{P}(z(T) < z_{\min}) = \mathrm{P}(z(T) > z_{\max}) = q_z/2.$$

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

$$\psi(z) \to \frac{\psi(z_0)}{\sqrt{z_0}} \sqrt{z}, \tag{9.57}$$

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

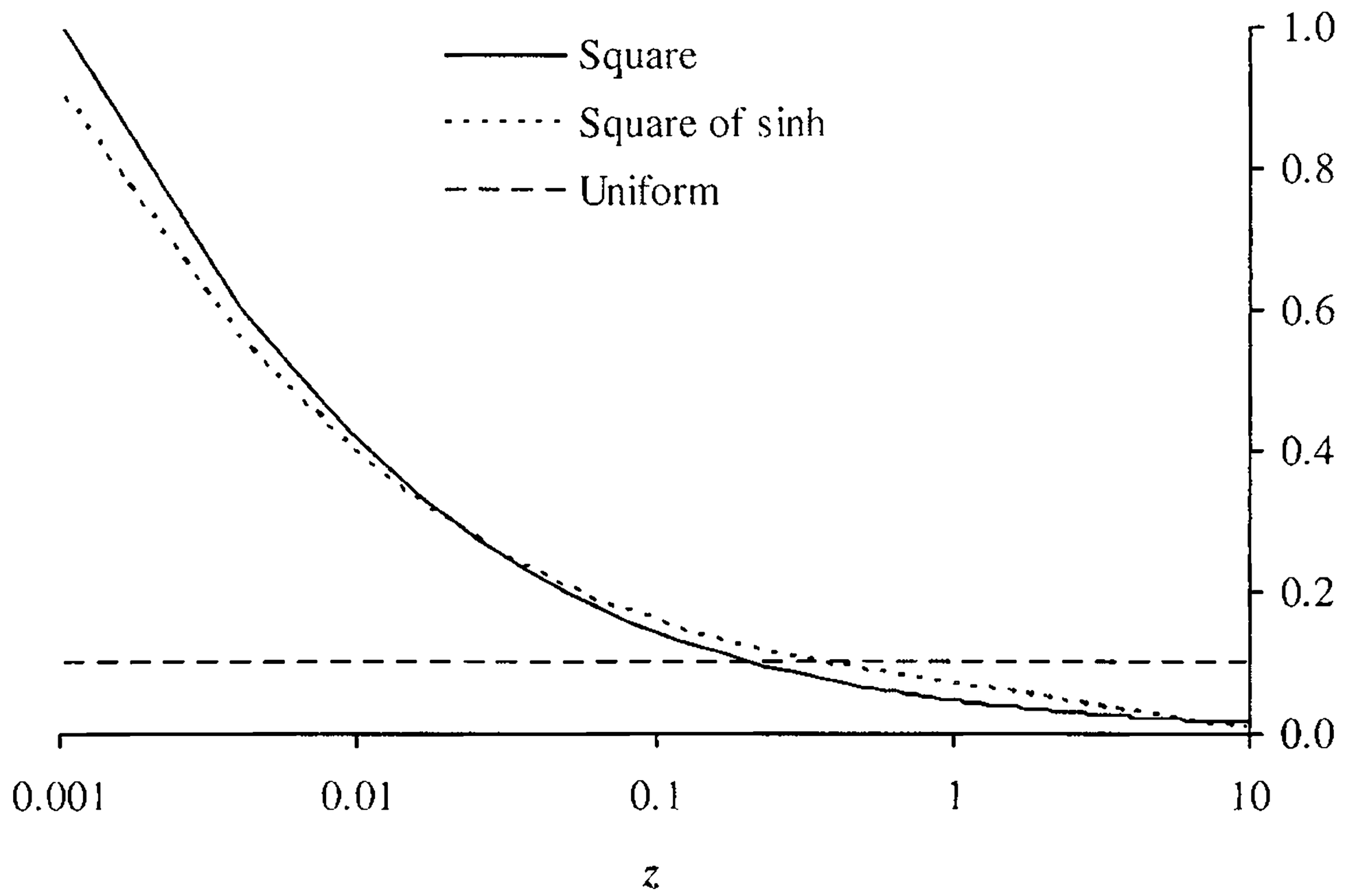
$$\zeta_n = \left(\sqrt{z_{\min}} + \frac{n}{N_z}\left(\sqrt{z_{\max}} - \sqrt{z_{\min}}\right)\right)^2, \quad n = 0, \ldots, N_z. \quad (9.60)$$

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):

$$\begin{aligned} \zeta_n &= \left(z_0 + \sinh\left(\alpha_{\min} + \frac{n}{N_z}\left(\alpha_{\max} - \alpha_{\min}\right)\right)\right)^2, \quad (9.61) \\ \alpha_{\min,\max} &= \sinh^{-1}\left(\sqrt{z_{\min,\max}} - z_0\right). \end{aligned}$$

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 $\bar{b}$ 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

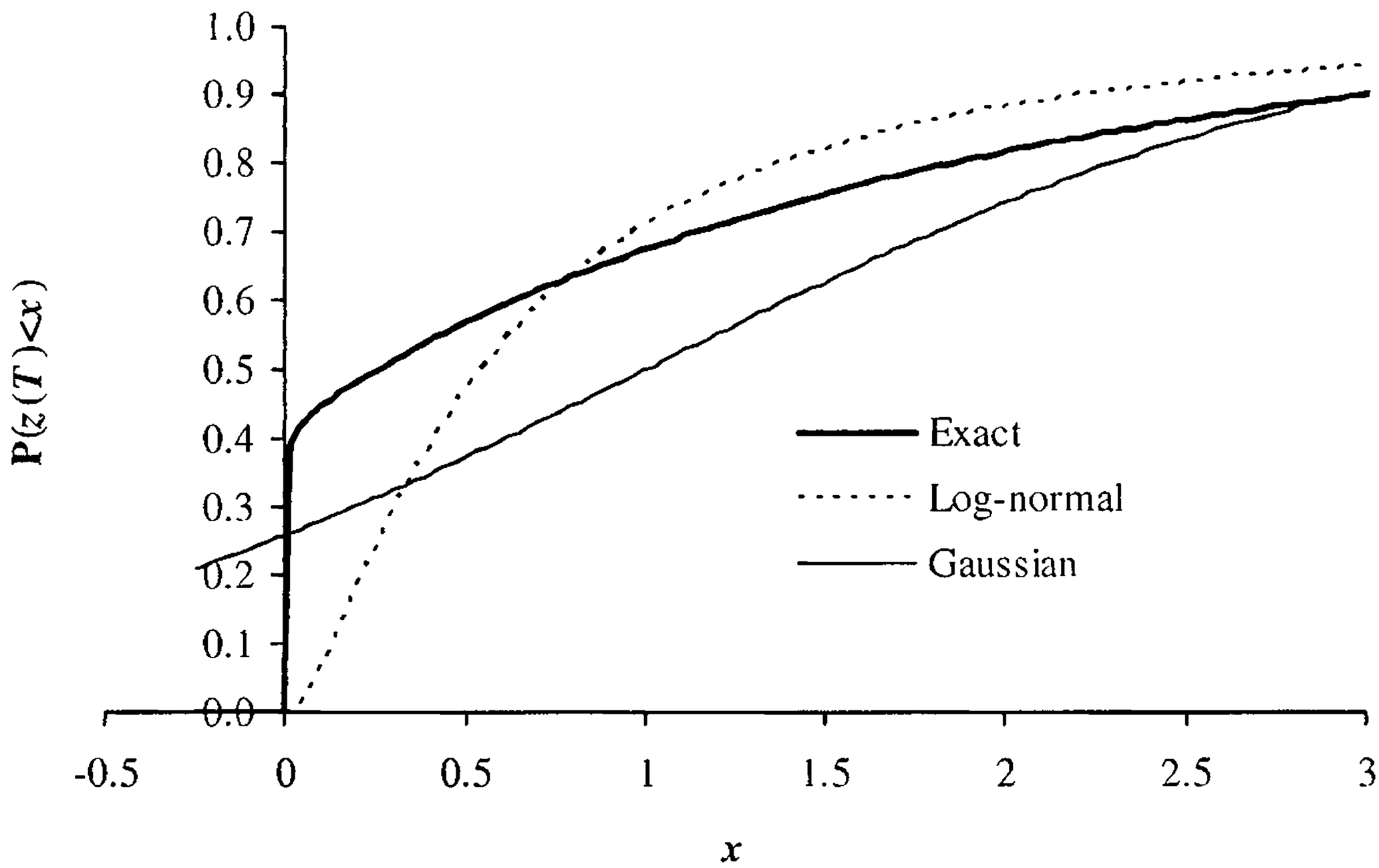
Figure 9.2 demonstrates that the density of $z(t+\Delta)|z(t)$ may sometimes be nearly singular at the origin. To accommodate this, one could contemplate inserting an actual singularity through outright truncation at the origin of a distribution that may otherwise go negative. Using a Gaussian distribution 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

$$\mathrm{P}\left(\widehat{z}(t+\Delta) \in [x, x+dx]\right) = \left(p\delta(x) + \beta(1-p)e^{-\beta x}\right)\, dx, \quad x \geq 0, \tag{9.76}$$

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

$$\Psi(x) = \mathrm{P}\left(\widehat{z}(t+\Delta) \leq x\right) = p + (1-p)\left(1 - e^{-\beta x}\right), \quad x \geq 0. \tag{9.77}$$

Here, the inverse of Ψ is readily computable:

$$\Psi^{-1}(u) = \Psi^{-1}(u; p, \beta) = \begin{cases} 0, & 0 \leq u \leq p, \\ \beta^{-1} \ln\left(\frac{1-p}{1-u}\right), & p < u < 1. \end{cases} \tag{9.78}$$

By the standard inverse distribution function method from Section 3.1.1.1, we thus get the simple sampling scheme

$$\widehat{z}(t+\Delta) = \Psi^{-1}\left(U_z; p, \beta\right) \tag{9.79}$$

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*

$$b^2 = 2\psi^{-1} - 1 + \sqrt{2\psi^{-1}}\sqrt{2\psi^{-1} - 1} \geq 0 \tag{9.80}$$

and

$$a = \frac{m}{1+b^2}. \tag{9.81}$$

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

$$\begin{aligned}\ln X(t+\Delta) &= \ln X(t) + \frac{\rho\lambda b}{\eta}\left(z(t+\Delta) - z(t) - \theta z_0 \Delta\right)\\ &\quad + \left(\frac{\theta\rho\lambda b}{\eta} - \frac{\lambda^2 b^2}{2}\right)\int_t^{t+\Delta} z(u)\,du + \lambda b\sqrt{1-\rho^2}\int_t^{t+\Delta}\sqrt{z(u)}\,dB(u),\end{aligned} \tag{9.86}$$

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 $\ln X$:

$$\ln \widehat{X}(t+\Delta) = \ln \widehat{X}(t) - \frac{1}{2}\lambda^2 b^2 \widehat{z}(t)^+ \Delta + \lambda b \sqrt{\widehat{z}(t)^+}\, W \sqrt{\Delta}\,, \tag{9.87}$$

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.

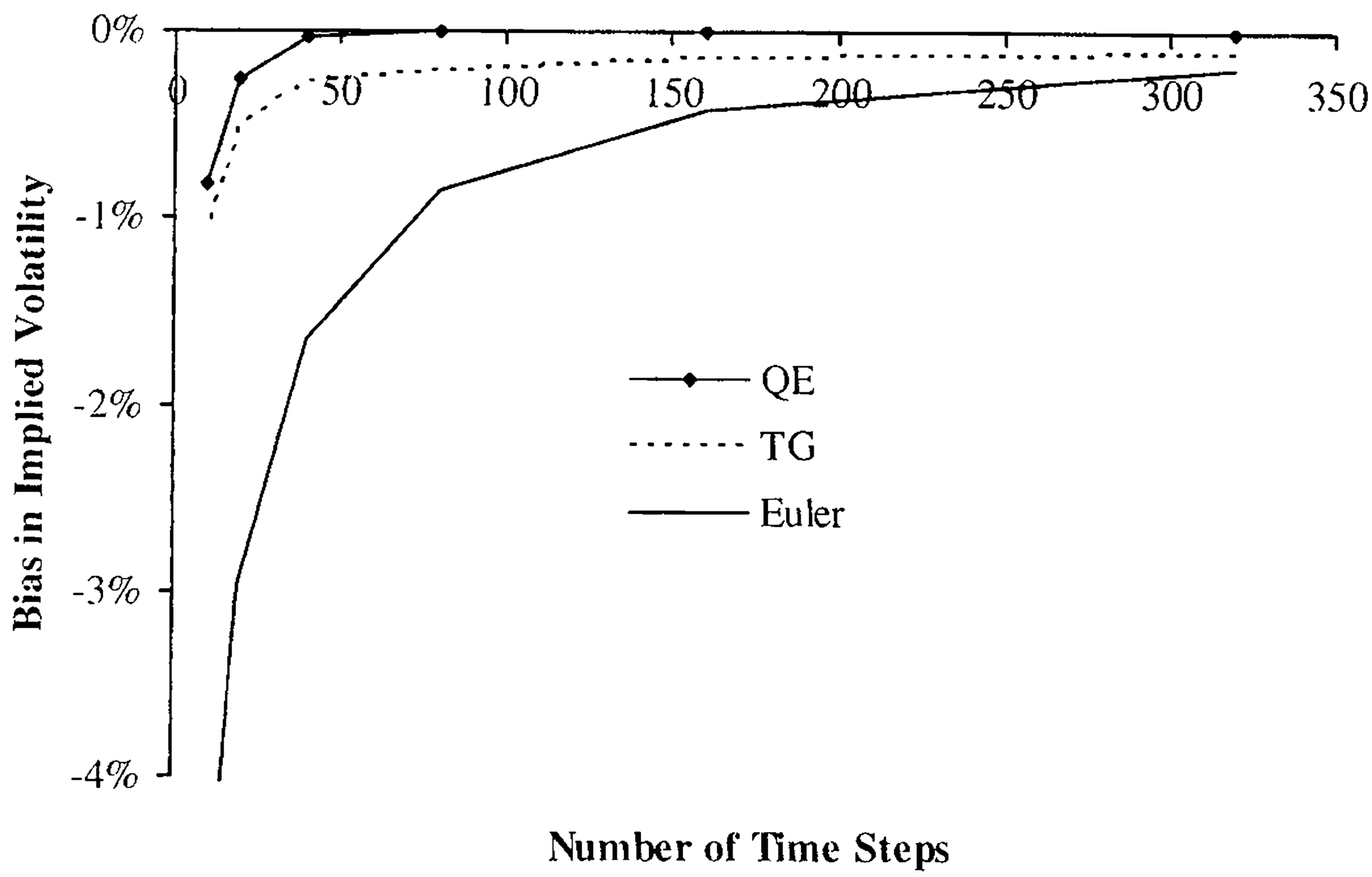
9.5.5.2 Simplified Broadie-Kaya

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,

$$\frac{1}{(i+2)!}\mathrm{E}\left(X_\epsilon(T)-x_0\right)^{i+2}=\int_{-\infty}^{\infty}\left\langle\delta_{x_0,z_0},(K-x_0)^i P_\epsilon^X(0,T)\,\pi_K\right\rangle dK,$$

(and the same for Y_ϵ) for $i=0,1$, where we have defined the payoff π_K by

$$\pi_K(x,z)=\begin{cases}(x-K)^+, & K\geq x_0,\\ (K-x)^+, & K<x_0.\end{cases}$$

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,

$$\Delta(i)=\int_{-\infty}^{\infty}(K-x_0)^i\left\langle\delta_{x_0,z_0},\left(P_\epsilon^Y(0,T)-P_\epsilon^X(0,T)\right)\pi_K\right\rangle dK.$$

By (9.96) we have,

$$\Delta(i)=\int_{-\infty}^{\infty}(K-x_0)^i \times\left(\int_0^T\left\langle\delta_{x_0,z_0},P_\epsilon^Y(0,t)\left(L_\epsilon^Y(t)-L_\epsilon^X(t)\right)P_\epsilon^X(t,T)\,\pi_K\right\rangle dt\right)dK.$$

After a series of manipulations (see Piterbarg [2005b] for details) we obtain, to order $o(\epsilon^2)$,

$$\Delta(i)=\frac{1}{2}\int_0^T\int\widehat{p}(x,z)\,(x-x_0)^i\left(\overline{f}_\epsilon(x)^2-f_\epsilon(t,x)^2\right)\lambda(t)^2\,dx\,dt, \tag{9.98}$$

$\widehat{p}(t,x)\triangleq\mathrm{E}\left(z(t)\delta(X_0(t)-x_0)\right).$

Expanding f, $\overline{f}$ to the first order around (s,x_0), we obtain

$$
\begin{aligned}
\delta(t;i) &\triangleq \int \widehat{p}(t,x)(x-x_0)^i \left(\overline{f}_\epsilon(x)^2 - f_\epsilon(t,x)^2\right) dx \\
&= 2\epsilon \left(\frac{\partial f(s\epsilon^2, x_0)}{\partial x} - \int_0^T \frac{\partial f(s\epsilon^2, x_0)}{\partial x} w(s)\, ds \right) \\
&\quad \times \int \widehat{p}(t,x)(x-x_0)^{i+1}\, dx \\
&\quad + \epsilon^2 \left(\left[\frac{\partial f(s\epsilon^2, x_0)}{\partial x}\right]^2 - \int_0^T \left[\frac{\partial f(s\epsilon^2, x_0)}{\partial x}\right]^2 w(s)\, ds \right) \\
&\quad \times \int \widehat{p}(t,x)(x-x_0)^{i+2}\, dx \\
&\quad + o\left(\epsilon^2\right).
\end{aligned}
$$

Calculating the integrals, we obtain to order $o(\epsilon^2)$,

$$
\begin{aligned}
\delta(t;i) &= 2\epsilon v(t)^2 \left(\frac{\partial f(s\epsilon^2, x_0)}{\partial x} - \int_0^T \frac{\partial f(s\epsilon^2, x_0)}{\partial x} w(s)\, ds \right), \\
\Delta(i) &= \frac{1}{2} \int_0^T \delta(t;i)\, \lambda(t)^2\, dt.
\end{aligned}
$$

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,

$$
\begin{aligned}
\mathrm{E}\left((X_0(t)-x_0)^2 z(t)\right) &= \mathrm{E}\left(z(t)\mathrm{E}\left((X_0(t)-x_0)^2 \middle| z(\cdot)\right)\right) \qquad (9.99) \\
&= \mathrm{E}\left(z(t)\int_0^t z(s)\lambda(s)^2\, ds\right) \\
&= \int_0^t \lambda(s)^2 \mathrm{E}\left(z(t)z(s)\right) ds.
\end{aligned}
$$

Clearly

$$
z(t) - z_0 = e^{-\theta(t-s)}\left(z(s) - z_0\right) + O\left(dW\right),
$$

so that

$$
\begin{aligned}
\mathrm{E}\left(z(t)z(s)\right) &= z_0^2 + \mathrm{E}\left(e^{-\theta(t-s)}\left(z(s)-z_0\right) z(s)\right) \\
&= e^{-\theta(t-s)}\mathrm{E}\left(z(s)^2\right) + \left(1 - e^{-\theta(t-s)}\right) z_0^2.
\end{aligned}
$$

We also have that

$$
\mathrm{E}\left(z(s)^2\right) = z_0^2 + z_0\eta^2 \frac{1-e^{-2\theta s}}{2\theta}. \qquad (9.100)
$$

Substituting into (9.99) yields

$$
\begin{aligned}
v(t)^2 &= \mathrm{E}\left((S(t)-x_0)^2 z(t)\right) && (9.101)\\
&= \int_0^t \lambda(s)^2 \left(e^{-\theta(t-s)}\mathrm{E}\left(z(s)^2\right) + \left(1-e^{-\theta(t-s)}\right) z_0^2\right) ds \\
&= \int_0^t \lambda(s)^2 \Bigg(e^{-\theta(t-s)} z_0^2 + e^{-\theta(t-s)} z_0 \eta^2 \frac{1-e^{-2\theta s}}{2\theta} && (9.102)\\
&\quad + z_0^2\left(1-e^{-\theta(t-s)}\right)\Bigg) ds \\
&= \int_0^t \lambda(s)^2 \left(z_0^2 + z_0\eta^2 e^{-\theta(t-s)} \frac{1-e^{-2\theta s}}{2\theta}\right) ds \\
&= z_0^2 \int_0^t \lambda(s)^2\, ds + z_0\eta^2 e^{-\theta t} \int_0^t \lambda(s)^2 \frac{e^{\theta s}-e^{-\theta s}}{2\theta}\, ds.
\end{aligned}
$$

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

$$
\begin{aligned}
\beta_0 =& \tau^{-2} l_{2,2}\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right) \\
&- \tau^{-3} l_{2,3}\left(\Omega_{31} - \Omega_{21}^2 - \frac{1}{4}\Omega^2\left(\Omega_{20} + \Omega_{10}^2\right) + \left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)^2\right) \\
&+ \frac{1}{2}\tau^{-4} l_{1,2}^2\left(\Omega_{41} - 3\Omega_{31}\Omega_{21} + 2\Omega_{21}^2 - \frac{1}{4}\Omega^2\Omega_{30} - \frac{3}{4}\Omega^2\Omega_{10}\Omega_{20}\right) \\
&+ \frac{3}{2}\tau^{-4} l_{1,2}^2\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)\left(\Omega_{31} - \Omega_{21}^2 - \frac{1}{4}\Omega^2\left(\Omega_{20} + \Omega_{10}^2\right)\right),
\end{aligned}
$$

$$
\begin{aligned}
\beta_1 = {} & \Omega^{-2}\tau^{-2} l_{2,2}\Omega_{10} \\
& - \Omega^{-2}\tau^{-3} l_{2,3}\left(\Omega_{20} - 3\Omega_{10}^2 + 2\Omega_{10}\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)^2\right) \\
& + \Omega^{-2}\frac{1}{2}\tau^{-4} l_{1,2}^2\Big(\Omega_{30} - 9\Omega_{10}\Omega_{20} + 12\Omega_{10}^3 \\
& \quad + 3\Omega_{10}\left(\Omega_{31} - \Omega_{21}^2 - \frac{1}{4}\Omega^2\left(\Omega_{20} + \Omega_{10}^2\right)\right)\Big) \\
& + \Omega^{-2}\frac{3}{2}\tau^{-4} l_{1,2}^2\left(\Omega_{21} - \frac{1}{4}\Omega^2\Omega_{10}\right)\left(\Omega_{20} - 3\Omega_{10}^2\right),
\end{aligned}
$$

$$
\beta_2 = -\Omega^{-4}\tau^{-3} l_{2,3}\Omega_{10}^2 + \Omega^{-4}\frac{3}{2}\tau^{-4} l_{1,2}^2\Omega_{10}\left(\Omega_{20} - 3\Omega_{10}^2\right).
$$