NUMERICAL METHODS IN FINANCE

Dr Antoine Jacquier

wwwf.imperial.ac.uk/˜ajacquie/

Department of Mathematics

Imperial College London

Spring Term 2015-2016

MSc in Mathematics and Finance

This version: April 3, 2016

Contents

Notations and standard definitions

The notations below will be used throughout the notes. We also wish to emphasize some common notational mistakes.

$\mathbb{N}$	integer numbers $\{0,1,2,\ldots\}$ (including 0)
$\mathbb{N}^*$	non null integer numbers $\{1,2,\ldots\}$
$\mathcal{M}_{m,n}(\mathbb{R})$	set of $m \times n$ matrices with real elements
$\mathcal{M}_n(\mathbb{R})$	set of $n \times n$ matrices with real elements
A^o	interior of a set A
$\overline{A}$	closure of a set A
$\mathcal{N}$	cumulative distribution function of the standard Gaussian distribution
$X = (X_t)_{t\geq 0} \neq X_t$	a process evolving in time, as opposed to X_t, which represents the (possibly random) value of the process X at time t
$f \neq f(x)$	f represents a function and $f(x)$ the value of the function f at the point x. Equivalently the function f can be written as $x \mapsto f(x)$
$\widehat{f}$	Fourier transform of a function f
$f(x) = \mathcal{O}(g(x))$ $(x \to \infty)$	there exist $M, x_0 > 0$ such that $\lvert f(x)\rvert \leq M\lvert g(x)\rvert$ for all $x > x_0$
$f(x) = \mathcal{O}(g(x))$ $(x \to a)$	there exist $M, \delta > 0$ such that $\lvert f(x)\rvert \leq M\lvert g(x)\rvert$ for all $\lvert x-a\rvert < \delta$
$f(x) = o(g(x))$ $(x \to a)$	$\lim_{x\to a} \frac{f(x)}{g(x)} = 0$, where $a \in \mathbb{R} \cup \{\pm\infty\}$
$\mathbb{1}_{\{x\in A\}}$	indicator function equal to 1 if $x \in A$ and zero otherwise
$x \wedge y$	$\min(x,y)$
a.s.	almost surely
$(x-y)_+$	$\max(0, x-y)$

Introduction and preliminaries

0.1 Some considerations on algorithms and convergence

Before diving into the meanders of numerical methods for finance, let us recall some basic definitions of algorithms and related numerical concepts.

Definition 0.1.1. An algorithm is a set of ordered instructions that will help construct the solution to a mathematical problem.

The above definition is obviously very broad and applies to many different situations. In the context we shall be interested in, an algorithm will deliver a sequence of values. We shall say that the algorithm is *convergent* if the above sequence converges to the desired solution. As an example one could think of the following problem: using the bisection method, solve the equation $f(x) = 0$ (for the unknown x) inside the interval $[a, b]$, where f is a strictly increasing function on $[a, b]$ such that $f(a)f(b) < 0$. The bisection method constructs a sequence of couples $(x_n, y_n)_{n\geq 0}$ in $[a, b]$ defined recursively by the following algorithm:

$$(x_{n+1}, y_{n+1}) := \begin{cases} \left(\frac{x_n + y_n}{2}, y_n\right), & \text{if } f\left(\frac{x_n + y_n}{2}\right) f(y_n) < 0, \\ \left(x_n, \frac{x_n + y_n}{2}\right), & \text{if } f\left(\frac{x_n + y_n}{2}\right) f(y_n) > 0, \\ (x_n, y_n), & \text{if } f(x_n) f(y_n) = 0, \end{cases}$$

for all $n \geq 0$, where the algorithm is started at $(x_0, y_0) := (a, b)$. The above example will clearly stop in the third case, where an exact solution is found (either x_n or y_n). If no such solution is found, the algorithm will never stop, and one hence needs a stopping criterion, namely the tolerance, i.e. a strictly positive real number $\varepsilon > 0$ such that if there exists $n \geq 1$ for which $|f(y_n) - f(x_n)| < \varepsilon$, then the algorithm is interrupted. In that case we shall say that the approximate solution is $\frac{x_n + y_n}{2}$ with an ε-tolerance.

More generally, we shall encounter several types of errors:

- the discretisation error, namely the error due to the approximation of a continuous-time random variable by a discrete-time one;

- the truncation error, for instance when computing $\int_a^b f(x)\mathrm{d}x$ instead of $\int_{-\infty}^{\infty} f(x)\mathrm{d}x$;
- the rounding error, when one truncates an exact number (with possibly an infinite number of decimals) to a finite number of decimals: for instance 3.14 in place of π.

We shall finally distinguish *well-conditioned* from *ill-conditioned* problems depending on how sensitive the solution of the problem is to a small perturbation of the initial data. Consider for instance a function $f : \mathbb{R} \to \mathbb{R}$ with a continuous derivative. From real data we are however only able to observe a noisy approximation f^{ε}, such that $|f^{\varepsilon}(x) - f(x)| < \varepsilon$ for all real number x (with $\varepsilon > 0$). Our problem is to determine the derivative f'. Using the central difference approximation (which we shall define and study in more details in Section 3.3)

$$f'(x) \approx \frac{f^{\varepsilon}(x+h) - f^{\varepsilon}(x-h)}{2h},$$

for some small enough $h > 0$. However, one can prove that the equality

$$\left| \frac{f(x+h) - f(x-h)}{2h} - \frac{f^{\varepsilon}(x+h) - f^{\varepsilon}(x-h)}{2h} \right| = \mathcal{O}\left(\frac{\varepsilon}{h}\right) \tag{0.1.1}$$

holds, so that the approximation error is of order ε/h. This implies that if h is too small compared to ε, the error from the data will become too important. This is an example of an ill-posed problem. The notion of well-posed problem was defined in 1902 by Jacques Hadamard [32] as a mathematical model for which a solution exists, is unique and depends continuously on the data.

In mathematical finance, we shall be concerned—to some degree— with pricing financial derivatives, i.e. evaluating quantities such as $f(x)$ ($x \in \mathbb{R}$) and the derivatives $f'(x)$, $f''(x), \ldots$ Since most financial derivatives do not have a closed-form solution, we will have to construct approximating sequences $(f_n(x), f_n'(x), f_n''(x), \ldots)_{n\in\mathbb{N}}$. Now, suppose we are able to construct a family of functions $(f_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in \mathbb{R}$ (i.e. the sequence $(f_n)_{n\in\mathbb{N}}$ converges *pointwise* to f). Can we then conclude that the same holds for its derivatives? The answer is negative in general. Let us recall some basic facts about convergence of functions. In the following, A will be a subset of the real line, and f, $(f_n)_{n\in\mathbb{N}}$ real functions from A to $\mathbb{R}$. All the following extend naturally to $\mathbb{R}^d$ or $\mathbb{C}^d$ ($d \geq 1$).

Definition 0.1.2. The sequence $(f_n)_{n\in\mathbb{N}}$ converges *pointwise* on A to the function f if for every $x \in A$, we have $\lim_{n\to\infty} f_n(x) = f(x)$. We can write this more formally as

$$\forall x \in A, \forall \varepsilon > 0, \exists N \in \mathbb{N} : \forall n \geq N, |f_n(x) - f(x)| < \varepsilon.$$

Remark.

- In the above definition, N usually depends both on ε and on x.
- Let (f_n) be defined by $f_n : [0,1] \ni x \mapsto \max\left(0, n - n^2 \left|x - n^{-1}\right|\right)$. The sequence converges pointwise to $f \equiv 0$ even though it becomes unbounded as n tends to infinity.

- Consider the sequence (f_n) defined on $[0,1]$ by $f_n(x) := x^n$. The sequence converges pointwise to the Dirac function at 1. In this case, even if each f_n is continuous, the limit is not.

In many cases this weak definition of pointwise convergence shall hence be insufficient and we may require a stronger form of convergence, which we introduce now.

Definition 0.1.3. The sequence $(f_n)_{n\in\mathbb{N}}$ converges *uniformly* on A to the function f if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} : \forall n \geq N, \forall x \in A : |f_n(x) - f(x)| < \varepsilon.$$

Remark. The main point here is that the same N applies to all values x in A.

We now state the most important result in this area.

Theorem 0.1.4. *Assume that each f_n is continuously differentiable on A and that*

(i) there exists $a \in A$ for which $(f_n(a))_{n\in\mathbb{N}}$ converges;

(ii) the sequence $(f'_n)_{n\in\mathbb{N}}$ converges uniformly on A.

Then the sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly to a function f on A and $f'(x) = \lim_{n\to\infty} f'_n(x)$ for all $x \in A$.

Remark. For the record, in a 1821 publication [11], Augustin-Louis Cauchy asserted that the pointwise limit of a sequence of functions is always continuous. This is of course false as we saw before, and was pointed out by Joseph Fourier, Niels Henrik Abel and Gustav Dirichlet. However only Karl Weierstrass in 1841 [55] published a rigorous definition of uniform convergence.

0.2 A concise introduction to arbitrage and option pricing

The fundamental model of mathematical finance consists of a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ on which we define a random variable S. The most obvious example is when S takes values in $\mathbb{R}$ or $\mathbb{R}_+$ and represents the price of a stock (or an interest rate, the price of some commodity...) at some given time, or when it is $\mathbb{R}^n$-valued ($n \in \mathbb{N}$) and accounts for the basket of share prices in an index such as the S&P500. One can also think of it as an infinite-dimensional random variable representing the whole path of a share price process between today and some future time, i.e. $S = (S_t)_{t\geq 0}$. A financial derivative written on the underlying random variable S can then be thought of as a functional $f(S)$. Financial derivatives are usually classified according to whether S represents the value of the share price at some future time $T > 0$ (European options) or the whole trajectory between today (time zero) and time T (American options). One of the fundamental questions in mathematical finance is to evaluate such functionals, i.e. to determine at time zero (inception of the contract) the expected value of $f(S)$. Intuitively speaking we wish to answer the following question: 'how much are we willing to pay today (time zero) to receive $f(S)$ at time T?' The answer to this question lies in what is called *absence of arbitrage*, which we shall define now. We consider a portfolio—or a trading strategy—as a random process $(V_t^\theta)_{t\geq 0}$ consisting of some positions in n stock prices $S^{(1)}, \dots, S^{(n)}$:

$$V_t^\theta = \sum_{i=1}^n \theta_t^{(i)} S_t^{(i)}, \qquad \text{for all } t \geq 0,$$

where $\theta_t^{(i)}$ represents the quantity of stock i in the portfolio at time t. We have written here V_t^θ to emphasise the fact that the strategy is fully determined by the (time-dependent) vector θ. In a discrete-time setting, let us fix some time $t > 0$. At time $t+1$, the investor may want to rebalance his portfolio, i.e. change its composition, and the value of the portfolio hence becomes

$$V_{t+1}^\theta = \sum_{i=1}^n \theta_{t+1}^{(i)} S_{t+1}^{(i)}.$$

If we assume that the investor does not invest nor withdraw any amount from his portfolio, then we necessarily have $\sum_{i=1}^n \theta_t^{(i)} S_{t+1}^{(i)} = \sum_{i=1}^n \theta_{t+1}^{(i)} S_{t+1}^{(i)}$. This can be written equivalently

$$V_{t+1}^\theta - V_t^\theta = \sum_{i=1}^n \theta_t^{(i)} \left(S_{t+1}^{(i)} - S_t^{(i)} \right),$$

and we shall call such a portfolio *self-financing*. We shall further call a trading strategy *admissible* if it is self-financing and if $V_t^\theta \geq 0$ for all $t \geq 0$.

Definition 0.2.1. An arbitrage is an admissible trading strategy (or a portfolio) $V^\theta = \left(V_t^\theta\right)_{t\geq 0}$ for which there exists some time $T > 0$ such that

$$V_0^\theta = 0 \text{ a.s.}, \qquad V_T^\theta \geq 0 \text{ a.s.} \qquad \text{and} \qquad \mathbb{Q}\left(V_T^\theta > 0\right) > 0.$$

Intuitively this means that one cannot make a sure profit out of nothing. In this definition we have used a probability $\mathbb{Q}$ given ad hoc as an element of the probability space. However for practical and theoretical reasons—which shall be made clear later in this course—we might want to use other probabilities, which are *equivalent* (in some sense to be made precise). In the discrete-time setting used above, we consider a family of random variables $X := (X_{t_1}, \dots, X_{t_n}, \dots)$ indexed by time steps. Consider further the family of nested sets $(\mathcal{F}_n)_{n\geq 1}$ satisfying $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$ for any $n \geq 1$. We shall informally— and leave the rigorous definition for later—consider $\mathcal{F}_n$ as the quantity of information available at time t_n generated by the random variables $X_{t_1}, \dots, X_{t_n}$.

Definition 0.2.2. We say that the family of random variables $X = (X_{t_1}, \dots, X_{t_n}, \dots)$ is a martingale if the equality $\mathbb{E}\left(X_{t_n} | \mathcal{F}_p\right) = X_{t_p}$ holds for any $1 \leq p \leq n$.

Example. If $(Y_n)_{n\geq 1}$ forms a family of independent identically distributed random variables such that for any $n \geq 1$, $\mathbb{E}(Y_n|\mathcal{F}_n) = 0$ and $(\mathcal{F}_n)_{n\geq 1}$ is the related flow of information, then the family $(X_n)_{n\geq 1}$ defined by $X_n := \sum_{i=1}^n Y_i$ is a martingale.

Definition 0.2.3. A probability measure $\mathbb{P}$ is a martingale measure equivalent to $\mathbb{Q}$—and we denote it by $\mathbb{P} \sim \mathbb{Q}$— if discounted price processes are martingales under $\mathbb{P}$.

Theorem 0.2.4 (Fundamental theorem of asset pricing). *A model is said to be arbitrage-free—i.e. there does not exist any admissible arbitrage strategy—if and only if there exists an equivalent martingale measure* $\mathbb{P}$*.*

This fundamental theorem has the following immediate application in terms of pricing: under absence of arbitrage, the price at time zero of a financial derivative is equal to the discounted expectation of the final payoff $f(X)$ under the martingale measure $\mathbb{P}$, i.e. $V_0 = \mathbb{E}^{\mathbb{P}}\left(f(X)\right)$.

0.2.1 European options

A European option is a financial contract, the payoff of which only depends on the final value (at maturity) of an underlying asset S. The simplest example of a European option with strike $K > 0$ and maturity $T > 0$ is that of a Call option, where the payoff is given by $f(S) = \max(S_T - K, 0)$, where S_T represents the time T value of the share price process S. This therefore corresponds to the right—but not the obligation—to buy the asset S at time T at the price K. Indeed, the buyer of the option would only exercise his right if the value S_T of the stock price at maturity T is greater to the strike K. In this case, his profit at maturity is $S_T - K$. Similarly, the payoff of a Put option is given by $\max(K - S_T, 0)$.

The question of interest here is: how much is an investor willing to pay this option at time zero (the inception of the contract)? The answer lies in the following theorem:

Theorem 0.2.5. *Consider a European call option and a European put option, both written on the share price S, with strike K and maturity T. Denote C_t and P_t their respective values at time $t \in [0, T]$. Assume that there is no arbitrage opportunity and denote $B_{t,T}$ the value at time $t \in [0, T]$ of a risk-free bond paying £1 at time T. The following properties hold:*

(i) Put-Call parity: $C_t - P_t = S_t - KB_{t,T}$;

(ii) The Call option is a decreasing function of the strike;

(iii) Call and Put options are convex functions of the strike;

(iv) Call and Put options are increasing functions of the maturity.

Proof. Let us prove (i). Consider the two portfolios

- $\left(\Pi_t^1\right)_{t\geq 0}$ consists of a short position in a Call option and a long position in a stock price: $\Pi_t^{(1)} = S_t - C_t$;
- $\left(\Pi_t^2\right)_{t\geq 0}$ consisting of a short position in a Put option and a long position in the a bond paying K at time T: $\Pi_t^{(2)} = KB_{t,T} - P_t$.

At maturity T, either the stock price S_T is above the strike K (and hence $(S_T - K)_+ = S_T - K$ and $(K - S_T)_+ = 0$) or it is below, and therefore $(S_T - K)_+ = 0$ and $(K - S_T)_+ = K - S_T$. In the first case, we have $\Pi^1_T = K = \Pi^2_T$. In the second case, we have $\Pi^1_T = S_T = \Pi^2_T$. In both cases, the two portfolios have the same value at maturity. By a no-arbitrage argument, they must therefore be equal at all time $t \in [0, T]$, and the statement is proved. The other statements follow similarly, and are left as an exercise. Let us just recall that a function $f : \mathbb{R} \to \mathbb{R}$ is convex on an interval $[a, b]$ if and only if for any $(x, y) \in [a, b]^2$ and $\lambda \in [0, 1]$, the inequality $f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$ holds. □

0.2.2 American options

When buying an American option maturing at time $T > 0$, a financial agent has the right—but not the obligation—to exercise the option at any time between the inception of the contract and the maturity. The payoff $f(S)$ of the American option is therefore a function of the whole path of the process, i.e. $S = (S_t)_{0 \leq t \leq T}$. Since they give additional rights to the bearer of the option, American options are always (but not necessarily strictly) more expensive than their European counterparts.

0.2.3 Exotic options

Options have been created in order to answer growing needs on financial markets. The specificities of each deal has led to an increasing number of option types, and we just mention here a few that have become rather standard. A European barrier option is like a European call or put option, but the option can only be exercised if the share price process has remained within (or exited) a predefined range between the inception and the maturity of a contract. To be more specific, let $K > 0$ denote the strike price and $B > 0$ the—contractually agreed—barrier. Consider the case where $S_0 > B$. The payoff of a down-and-in European (call) barrier option reads $(S_T - K)_+ \mathbb{1}_{\{\tau < T\}}$, where $\tau := \inf\{t \in [0, T] : S_t < B\}$ represents the first time the share price falls below the barrier. Similarly, a down-and-out option has the payoff is $(S_T - K)_+ \mathbb{1}_{\{\tau > T\}}$. We can define analogously up-and-out or up-and-in options. Asian options have also become popular since their creation in 1987 in Tokyo. Their payoff depends on the whole path of the process S on $[0, T]$. Standard examples are (continuously monitored) Asian calls on the arithmetic average, where the payoff reads $\left(\frac{1}{T}\int_0^T S_t \mathrm{d}t - K\right)_+$ or (discretely monitored) Asian calls on the arithmetic average, with payoff $\left(\frac{1}{n}\sum_{i=1}^n S_{t_i} - K\right)_+$ for some contractually specified dates $0 \leq t_1 < \ldots < t_n \leq T$. These are standard options on a single stock used every day. Options on several stocks—*Basket options*—are designed in the same way, for instance, a European call option on the mean of two stock prices S^1 and S^2 with strike $K > 0$ and maturity $T > 0$ pays $\left(\frac{1}{2}\left(S^2_T + S^2_T\right) - K\right)_+$ at maturity. We shall see other types of options in this course and will add details as we need them.

Chapter 1

Lattice (tree) methods

In this chapter we shall present a discrete-time method for option pricing. This setting will allow us to introduce the concepts of no-arbitrage and risk-neutral expectation presented in Section 0.2 in a more rigorous way.

1.1 Binomial trees

1.1.1 One-period binomial tree

We fist consider the simple one-period case. Let S denote a stock price, whose value at time $t_0 \geq 0$ is $S_0 > 0$. At time $t_1 > t_0$, the process can take two possible values:

$$S_1 = \begin{cases} uS_0, & \text{with probability } p, \\ dS_0, & \text{with probability } 1-p, \end{cases}$$

where $0 < d < u$ and $p \in (0,1)$. Another way to understand this is to write $S_1 = XS_0$, where X is a Bernoulli random variable that takes the value u with probability p and the value d with probability $1-p$. We assume that a financial agent can invest (at time $t = 0$) in both the asset S and in a risk-free bond, i.e. borrow or sell money with a (non random) interest rate equal to $r \geq 0$ over the period $[t_0, t_1]$. We are now interested in determining the price C_0 at time t_0 of a European Call option with strike $K > 0$ and maturity t_1. At time t_1, the two possible payoffs—corresponding to the two different states of the world—are $C_1^{(u)} := (uS_0 - K)_+$ and $C_1^{(d)} := (dS_0 - K)_+$. By a simple no-arbitrage argument, one may be tempted to value it at the price $C_0 = \frac{\mathbb{E}_p(C_1)}{1+r} = \frac{1}{1+r}\left(pC_1^{(u)} + (1-p)\,C_1^{(d)}\right)$. However, this is in general false, since the probability p has been chosen from the investor's point of view, and does not necessary reflect the market's point of view. It turns out that this very probability p, called the historical (or physical) probability, does not appear at all in the pricing formula, as the following theorem shows.

Theorem 1.1.1. *In the absence of arbitrage opportunities, the price at time t_0 of a European call option written on S, with strike K and maturity t_1 is worth*

$$C_0 = \frac{\pi C_1^{(u)} + (1-\pi)\, C_1^{(d)}}{1+r}, \tag{1.1.1}$$

where $\pi := \dfrac{1+r-d}{u-d}$.

Proof. The following proof is based on the concept of the *pricing by replication*, i.e. we want to construct a portfolio consisting of shares and risk-free bonds that exactly replicates— has the same payoff as—the option we wish to evaluate. Consider a portfolio Π consisting of an amount Δ_0 of shares and with the notional ϕ invested in the risk-free bond. The value at time t_0 of the portfolio is therefore $\Pi_0 = \Delta_0 S_0 + \phi$. At time t_1, it is worth

$$\Pi_1 = \Delta_0 S_1 + (1+r)\,\phi = \begin{cases} \Delta_0 u S_0 + (1+r)\,\phi, & \text{with probability } p, \\ \Delta_0 d S_0 + (1+r)\,\phi, & \text{with probability } 1-p. \end{cases}$$

Since our portfolio Π has to replicate the option, it therefore needs to have the same payoff. This implies the following system of equations:

$$\begin{cases} \Delta_0 u S_0 + (1+r)\,\phi & = C_1^{(u)} \\ \Delta_0 d S_0 + (1+r)\,\phi & = C_1^{(d)}, \end{cases}$$

which we can solve explicitly as

$$\Delta_0 = \frac{C_1^{(u)} - C_1^{(d)}}{(u-d)\, S_0}, \qquad \text{and} \qquad \phi = \frac{1}{1+r}\frac{uC_1^{(d)} - dC_1^{(u)}}{(u-d)}.$$

By absence of arbitrage, since our portfolio Π and the Call option have the same value at maturity (same payoff), then they necessarily have the same value at inception of the contract, i.e. at time t_0, so that $C_0 = \Pi_0$. Define now $\pi := \dfrac{1+r-d}{u-d}$ and the theorem follows. □

Remark 1.1.2.

(i) The historical probability p does not appear in the final formula.

(ii) If the quantity π lies between zero and one, then we could interpret it as a (new) probability under which the option value at time t_0 is the expected value of its payoff at maturity t_1. In fact we can show (see Exercise 1) that absence of arbitrage implies the inequalities $d < 1 + r < u$ and hence $\pi \in (0,1)$. From Theorem 1.1.1 the call price at time zero therefore reads $C_0 = (1+r)^{-1}\,\mathbb{E}_\pi(C_1)$. This probability π is called the *risk-neutral probability*, under which option prices are martingales.

(iii) Beyond absence of arbitrage, the replication strategy holds because the market is complete, which means that we are able to fully replicate the option using only traded assets (here simply shares and bonds). When this is not the case, markets are said to be *incomplete* and the risk-neutral probability might not be uniquely defined any more. We shall see such an example of incomplete market in Section 1.2 below.

(iv) Note that the proof of the theorem does not rely on the particular form of the payoff of the call option. The same result therefore carries out to Put options and other European options.

Exercise 1. Show that the absence of arbitrage opportunities implies $d < 1 + r < u$, and hence $\pi \in (0, 1)$. Construct an arbitrage when this inequality does not hold.

Solution. *Assume for instance that $1 + r \leq d$. At time zero, borrow an amount S_0 and buy the stock price with this money. At maturity T, pay back the amount you owe (i.e. $S_0 (1 + r)$) and sell the asset. The net profit / loss is therefore equal to $S_T - S_0 (1 + r) > S_T - dS_0$. Since this quantity is strictly positive, this means an arbitrage opportunity exists.*

Exercise 2. Show that under the risk-neutral probability π, the expected return of the option is equal to the risk-free interest rate r.

1.1.2 Multi-period binomial tree

We now extend the one-period binomial tree approach developed above to the multi-period case. Let N be a strictly positive integer representing the number of periods, and let $0 = t_0 < t_1 < \ldots < t_N = T$ denote the discretisation time steps. We shall consider that the time increment $t_n - t_{n-1}$ is the same for any $n = 1, \ldots, N$, and we denote it τ. Between two consecutive nodes of the tree, the stock price can either jump up by a percentage u with probability p or jump down by a percentage d with probability $1 - p$, as in the following figure:

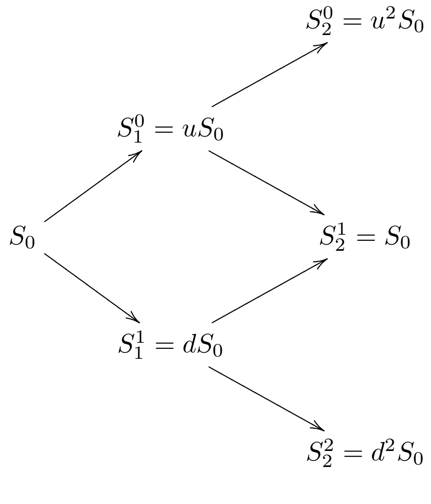

Figure 1.1: Three-period symmetric and recombining binomial tree.

It is straightforward to see that at any time t_n ($n = 0, \ldots, N$), the process S takes values in the set $\left\{u^{n-k}d^k S_0\right\}_{k=0,\ldots,n}$, and we use the notation $S_n^k := u^{n-k}d^k S_0$, where the subscript n represents the time t_n and the superscript k accounts for the state. In particular, out of all the possible paths, only $\binom{n}{k} := \frac{n!}{k!(n-k)!}$ lead to the value $u^{n-k}d^k S_0$, so that

$$\mathbb{P}\left(S_n = S_n^k\right) = \binom{n}{k} p^{n-k}\left(1-p\right)^k, \qquad \text{for all } k = 0, \ldots, n.$$

As in the one-period case, we are interested in finding the price at time t_0 of a European option V. As mentioned in Remark 1.1.2 in the one-period case, the proof of the option price does not rely on the particular form of the payoff, so that V can be a Put, a Call or any other European option. Similar to the notation S_n^k, we shall use V_n^k to denote the value of the option in state k at time t_n.

Remark 1.1.3. In the one-period binomial tree above, we have implicitly assume that the time increment was one year. Discrete compounding corresponds to the case where interest rate r corresponds to a given time period, say 1 month (1/12 year). In this case, investing an amount B at time zero yields $B\left(1 + \frac{r}{12}\right)^{12}$ one year later. More generally, if $n \geq 1$ is the number of compounding periods, then investing B at time zero yields $B\left(1 + \frac{r}{n}\right)^n$ one year later. Continuously compounding corresponds to the case where n tends to infinity. Since $\lim_{n\to\infty}\left(1 + \frac{r}{n}\right)^n = \mathrm{e}^r$, we can directly generalise this and hence the interest over a period of time t is worth e^{rt}.

Theorem 1.1.4. *In the absence of arbitrage opportunities, the price V_0 at time t_0 of the European option written on S reads*

$$V_0 = R^{-N} \sum_{k=0}^{N} \binom{N}{k} \pi^{N-k}\left(1-\pi\right)^k V_N^k, \tag{1.1.2}$$

where again $\pi := \frac{R-d}{u-d}$ and $R := \exp\left(rT/N\right)$.

Remark 1.1.5. From a computational point of view, this formula may not be optimal. Indeed one has to evaluate terms such as $n!$. When n is not even too large, this will create an overflow. For instance, for $n = 20$, we already have $n! = 2432902008176640000$. Computing the tree backward from the terminal value one step at a time will remain robust as n goes large, since no such number will need to be computed.

Proof. The proof follows a backward induction scheme. For any $n = 0, \ldots, N$ and $k = 0, \ldots, n$, we use the same notation V_n^k as for the stock price to denote the value at time t_n of the option in state k. Consider the next-to-final time t_{N-1}, when the stock price takes values in the set $\left\{S_{N-1}^k = u^{N-1-k}d^k S_0\right\}_{k=0,\ldots,N-1}$. At maturity t_N, the stock price can either go up or go down, and we can therefore apply the one-period binomial Theorem 1.1.1 to obtain

$$V_{N-1}^k = \frac{\pi V_N^k + \left(1-\pi\right) V_N^{k+1}}{R}, \qquad \text{for all } k = 0, \ldots, N-1.$$

At time t_{N-2}, we can apply the same argument and we obtain

$$V_{N-2}^k = \frac{\pi V_{N-1}^k + (1-\pi)\, V_{N-1}^{k+1}}{R} = \frac{\pi^2 V_N^k + 2\pi\,(1-\pi)\, V_N^{k+1} + (1-\pi)^2\, V_N^{k+2}}{R^2}, \qquad \text{for } k = 0,\ldots,N-2.$$

The theorem then follows by backward induction. □

Exercise 3. Consider a European call option V with strike $K > 0$, written on an underlying asset S. In the multi-period binomial model with N periods, determine the integer N_0 such that the value of the European call option at time t_0 is worth

$$R^{-N} \sum_{k=0}^{N_0} \binom{N}{k} \pi^{N-k}\,(1-\pi)^k \left(S_N^k - K\right),$$

where the risk-neutral probability π is given in Theorem 1.1.4. Prove that this is also equal to

$$S_0 \mathrm{B}\left(N_0, N, \frac{(1-\pi)\, d}{R}\right) - \frac{K}{R^N} \mathrm{B}\Big(N_0, N, (1-\pi)\Big), \tag{1.1.3}$$

where $\mathrm{B}(\cdot, n, p)$ represents the cumulative distribution function of a Binomial random variable with probability of success at each trial p and number of trials n.

Solution. *In the sum in Theorem 1.1.4, the payoff* $V_N^k := \left(S_N^k - K\right)_+$ *will be non null as soon as* $S_N^k = u^{N-k} d^k S_0 \geq K$. *Define* $N_0 := \min\left\{\sup\left\{0 \leq k : u^{N-k} d^k S_0 \geq K\right\}, N\right\}$. *Now,*

$$u^{N-k} d^k S_0 \geq K \qquad \textit{if and only if} \qquad k \leq \frac{N \log(u) - \log(K/S_0)}{\log(u) - \log(d)},$$

and hence N_0 *is the smaller between the largest integer smaller than the right-hand side of the second inequality and* N. *Recall that for a Binomial random variable* Y *with parameters* $n \in \mathbb{N}^*$ *and* $p \in [0,1]$, *we have* $\mathbb{P}(Y = k) = \binom{n}{k} p^k\,(1-p)^{n-k}$, *for any* $k \in \{0,\ldots,n\}$. *Furthermore,* $\mathbb{E}(Y) = np$ *and* $\mathbb{V}(Y) = np\,(1-p)$.

1.1.3 From discrete to continuous time

We have constructed above a discrete-time process describing the evolution of the stock price. It is natural then to wonder how this construction behaves as the time increment tends to zero: for a partition $0 = t_0 < \ldots < t_N = T$ of the interval $[0,T]$ where T is a fixed maturity, we wish to study the structure of the tree as N tends to infinity. The purpose of this section is to compute such a limit and to understand its implication on European option prices.

Kushner theorem for Markov chains

We begin by the definition of a Markov chain.

Definition 1.1.6. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given probability space. A sequence of random variables $(X_n)_{n\geq 1}$ on Ω is called a (discrete time) Markov chain if

$$\mathbb{P}\left(X_{n+1} = x | X_1 = x_1, \ldots, X_n = x_n\right) = \mathbb{P}\left(X_{n+1} = x | X_n = x_n\right), \qquad \text{for all } n \geq 1.$$

Furthermore it is called a time-homogeneous (or stationary) Markov chain if

$$\mathbb{P}\left(X_{n+1} = x | X_n = y\right) = \mathbb{P}\left(X_n = x | X_{n-1} = y\right), \qquad \text{for all } n \geq 1, x, y \in \Omega.$$

Remark 1.1.7. We used the notation S_n^k in the previous section to denote the value of the stock price at the node $(n, k) \in [0, N] \times [0, n]$. For a maturity $T > 0$ and a total number of nodes $N > 0$, we define the time increment $\tau := T/N$. We shall from now on slightly change the notation and define S_n^τ as the value of the stock price at node n, i.e. at time $n\tau$, for $n = 0, \ldots, N$. For each $n \in [0, N]$, S_n^τ is therefore a random variable taking value in $\left\{u^k d^{n-k} S_0\right\}_{k=0,\ldots,n}$. With this notation, the family of random variables $(S_n^\tau)_{n\in[0,N]}$ is a Markov chain.

In order to understand the following, we need to define a Brownian motion (in $\mathbb{R}$):

Definition 1.1.8. A one-dimensional standard Brownian motion $(W_t)_{t\geq 0}$ is a continuous-time family of random variables (namely a stochastic process) satisfying

(i) $W_0 = 0$ almost surely;

(ii) the paths of the process are almost surely continuous;

(iii) $(W_t)_{t\geq 0}$ has independent increments: for any $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2$, $W_{t_1} - W_{s_1}$ and $W_{t_2} - W_{s_2}$ are independent;

(iv) for any $0 \leq s \leq t$, $W_t - W_s$ is normally distributed as $\mathcal{N}(0, t - s)$.

Consider now a continuous-time process $(X_t)_{t\geq 0}$ satisfying the stochastic differential equation

$$\mathrm{d}X_t = b\left(X_t\right)\mathrm{d}t + \sigma\left(X_t\right)\mathrm{d}W_t, \qquad X_0 = x_0 \in \mathbb{R}, \tag{1.1.4}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion and b and σ smooth functions defined on $\mathbb{R}$.

Remark 1.1.9. We shall not delve into the exact meaning of such a representation for now. Note however that this is the canonical representation for a continuous-time diffusion process. By *diffusion*, we mean a time-evolving random process with continuous paths. On a small time interval $[t, t + \tau]$ (for some $t, \tau \geq 0$), we can represent (1.1.4) as

$$X_{t+\tau} = X_t + b(X_t)\tau + \sigma(X_t)\sqrt{\tau}Z,$$

where Z is a Gaussian random variable with zero mean and unit variance. From this representation, we clearly see that the process $(X_t)_{t\geq 0}$ evolves according to a drift $b(\cdot)$ with some random (Gaussian) perturbations amplified by a *diffusion coefficient* function $\sigma(\cdot)$. Note that from this discrete-time representation, the following two identities are immediate:

$$\mathbb{E}_t\left(X_{t+\tau}\right) = X_t + b(X_t)\tau \qquad \text{and} \qquad \mathbb{V}_t\left(X_{t+\tau}\right) = \sigma(X_t)^2\tau,$$

where $\mathbb{E}_t$ represents the expectation conditional to the information flow up to time t.

For any $\tau > 0$, let $(X_n^\tau)_{n\geq 1}$ be a Markov chain, and denote the increments $\Delta X_n^\tau := X_{n+1}^\tau - X_n^\tau$. We shall say that the family of Markov chains $(X_n^\tau)_{n\geq 1}^{\tau>0}$ is *locally consistent* with the continuous-time diffusion $(X_t)_{t\geq 0}$ if it satisfies the following three conditions:

$$\begin{aligned}
\mathbb{E}\left(\Delta X_n^\tau | X_n^\tau\right) &= b\left(X_n^\tau\right)\tau + o(\tau),\\
\mathbb{V}\left(\Delta X_n^\tau | X_n^\tau\right) &= \sigma\left(X_n^\tau\right)^2\tau + o(\tau),\\
\lim_{\tau\to 0}\sup_{n\geq 1}|\Delta X_n^\tau| &= 0.
\end{aligned}$$

The following theorem is a particular case of a more general result proved by Kushner [45], and we shall omit its proof for brevity.

Theorem 1.1.10. *Assume that the two functions b and σ are bounded, and that the family of Markov chains $(X_n^\tau)_{n\geq 1}^{\tau>0}$ is locally consistent with the diffusion $(X_t)_{t\geq 0}$ defined in* (1.1.4). *Then the process $\widetilde{X}_t^\tau := X_{[t/\tau]}^\tau$ converges in law to the diffusion $(X_t)_{t\geq 0}$.*

Reconciling the discrete and continuous time

In the previous sections, we have shown how to construct the so-called risk-neutral probability in a multi-period tree. However, we have not said anything about the upward and downward amplitudes u and d of the jumps. Let us first introduce one of the canonical models in finance.

Definition 1.1.11. Let Z be a Gaussian random variable with mean zero and unit variance and let σ be a strictly positive real number. If the stock price process S satisfies

$$S_{t+\tau} = S_t \exp\left(\left(r - \frac{1}{2}\sigma^2\right)\tau + \sigma\sqrt{\tau}Z\right),$$

for any $t, \tau \geq 0$, with $S_0 > 0$ and where $r \geq 0$ represents the risk-free interest rate, we say that the stock price follows the *Black-Scholes* model.

The following lemma gives us the mean and the variance of such dynamics. (Recall that the notation $\mathbb{E}_t$ stands for the expectation conditional on the filtration at time $t \geq 0$, see Appendix A.1).

Lemma 1.1.12. *In the Black-Scholes model, the stock price is lognormally distributed and*

$$\mathbb{E}_t\left(S_{t+\tau}\right) = S_t \mathrm{e}^{r\tau} \qquad \textit{and} \qquad \mathbb{E}_t\left(S_{t+\tau}^2\right) = S_t^2 \mathrm{e}^{\left(2r+\sigma^2\right)\tau}. \tag{1.1.5}$$

Proof. Exercise. □

We now wish to construct a binomial tree such that the limit of the process as the time increment tends to zero corresponds to this diffusion model. An amount B invested in a risk-free asset at time t yields at time $t+\tau$ the amount $(1+r\tau)B$. At time $t+2\tau$, this is worth $(1+r\tau)^2 B$, and so on. Repeat this n times ($n \geq 1$) on time periods $\widetilde{\tau}/n$, and let n tends to infinity:

$$\lim_{n\to\infty}\left(1 + \frac{r\widetilde{\tau}}{n}\right)^n B = \mathrm{e}^{r\widetilde{\tau}} B.$$

Consider now the binomial tree between two time nodes t and $t+\tau$. The expectation and variance of the asset price process read

$$\mathbb{E}_t\left(S_{t+\tau}\right) = S_t\Big(pu + (1-p)\,d\Big) \qquad \text{and} \qquad \mathbb{E}_t\left(S_{t+\tau}^2\right) = S_t^2\Big(pu^2 + (1-p)\,d^2\Big). \tag{1.1.6}$$

With Theorem 1.1.10 in mind, we equate the expectations and the variances in the continuous and in the discrete time settings in (1.1.5) and (1.1.6), divide respectively by S_t and S_t^2, so that

$$pu + (1-p)\,d = \mathrm{e}^{r\tau} \tag{1.1.7}$$

$$pu^2 + (1-p)\,d^2 = \mathrm{e}^{\left(2r+\sigma^2\right)\tau}. \tag{1.1.8}$$

There are two equations and three unknowns, p, u and d. From the first equation, we deduce $p = \dfrac{\mathrm{e}^{r\tau}-d}{u-d}$, which is the risk-neutral probability determined in Theorem 1.1.4. We therefore need an additional condition to properly determine the upward and downward amplitudes u and d. There are different possible choices, each of which shall determine a specific model. We list them here as a general overview, and each case will be treated separately in the forthcoming sections.

(i) Condition $ud = 1$: this corresponds to the figure 1.1 of a recombining tree. This model was proposed by Cox, Ross and Rubinstein in [15].

(ii) $p = \dfrac{1}{2}$: Jarrow-Rudd model.

(iii) Tian model: $pu^3 + (1-p)\,d^3 = \exp\Big(3\left(r+\sigma^2\right)\tau\Big)$.

Examples of models

The Cox-Ross-Rubinstein model

As mentioned above, the Cox-Ross-Rubinstein model, first developed in [15], corresponds to a recombining tree, i.e. imposing the condition $ud = 1$. Under this condition, the binomial model characterised by (1.1.7) and (1.1.8) is defined uniquely, as shown in the following proposition.

Proposition 1.1.13. *The unique solutions to the two equations* (1.1.7) *and* (1.1.8) *are*

$$u = \frac{\exp\left(-r\tau\right)}{2}\left(1+\xi^2+\sqrt{\left(1+\xi^2\right)^2-4\mathrm{e}^{2r\tau}}\right),$$

$$d = \frac{\exp\left(-r\tau\right)}{2}\left(1+\xi^2-\sqrt{\left(1+\xi^2\right)^2-4\mathrm{e}^{2r\tau}}\right),$$

where we define $\xi^2 := \exp\Big(\left(2r+\sigma^2\right)\tau\Big)$.

Proof. From the definition of ξ and (1.1.8), we can write

$$\xi^2 = \mathrm{e}^{\left(2r+\sigma^2\right)\tau} = p\left(u^2-d^2\right)+d^2 = \frac{\mathrm{e}^{r\tau}-d}{u-d}\left(u+d\right)\left(u-d\right)+d^2 = \mathrm{e}^{r\tau}u - 1 + \frac{\mathrm{e}^{r\tau}}{u},$$

which implies that $\mathrm{e}^{r\tau}u^2 - \left(1+\xi^2\right)u + \mathrm{e}^{r\tau} = 0$. This is a second-order polynomial in u. Its determinant $\left(1+\xi^2\right)^2 - 4\mathrm{e}^{2r\tau}$ is strictly positive since $r \geq 0$ and $\sigma > 0$. A straightforward manipulation of (1.1.8) shows that the same holds if we consider it as a polynomial in d instead of u. Since we want $d < u$, the theorem follows. □

Note in particular that the double inequality $d < 1 < u$ holds. To be historically precise, Cox, Ross and Rubinstein proposed to take $u = \exp(\sigma\sqrt{\tau})$ and $d = u^{-1}$, but did not prove Proposition 1.1.13, i.e. only proposed a truncated (to order $\sqrt{\tau}$) version of $\log(u)$ and $\log(d)$.

Corollary 1.1.14. *Up to order $\mathcal{O}(\tau^{3/2})$, a Taylor expansion in τ gives*

$$\log(u) = \sigma\sqrt{\tau}+\mathcal{O}\left(\tau^{3/2}\right), \qquad \log(d) = -\sigma\sqrt{\tau}+\mathcal{O}\left(\tau^{3/2}\right), \qquad p = \frac{\mathrm{e}^{r\tau}-d}{u-d} = \frac{1}{2}+\frac{\left(r-\frac{1}{2}\sigma^2\right)\sqrt{\tau}}{2\sigma}+\mathcal{O}(\tau).$$

With such a choice, Equations (1.1.7) and (1.1.8) are satisfied up to order $\mathcal{O}\left(\tau^2\right)$.

Note that we could also write it in the following way:

$$u = 1+\sigma\sqrt{\tau}+\frac{1}{2}\sigma^2\tau+\mathcal{O}\left(\tau^{3/2}\right) \qquad \text{and} \qquad d = 1-\sigma\sqrt{\tau}+\frac{1}{2}\sigma^2\tau+\mathcal{O}\left(\tau^{3/2}\right)$$

Proof. As the time increment τ tends to zero, we have the Taylor expansions

$$\frac{1+\xi^2}{2\mathrm{e}^{r\tau}} = 1+\frac{1}{2}\sigma^2\tau+\mathcal{O}\left(\tau^2\right) \qquad \text{and} \qquad \sqrt{\left(\frac{1+\xi^2}{2\mathrm{e}^{r\tau}}\right)^2-1} = \sigma\sqrt{\tau}+\mathcal{O}\left(\tau^{3/2}\right),$$

from which the expansions given in the corollary follow. □

The Jarrow-Rudd model: The extra condition in the Jarrow-Rudd model is $p = \frac{1}{2}$. This corresponds to approximating the Brownian motion by a random walk, i.e. an incremental process which can only move by one unit up or down between to time steps. Matching the first two moments give (we leave the proof as an exercise)

Proposition 1.1.15. *Under the condition $p = 1/2$, the unique solutions to (1.1.7) and (1.1.8) are*

$$u = \mathrm{e}^{r\tau}\left(1+\sqrt{\mathrm{e}^{\sigma^2\tau}-1}\right) \qquad \textit{and} \qquad d = \mathrm{e}^{r\tau}\left(1-\sqrt{\mathrm{e}^{\sigma^2\tau}-1}\right),$$

and Equations (1.1.7) and (1.1.8) are satisfied up to order $\mathcal{O}\left(\tau^2\right)$.

The actual parameters proposed by Jarrow and Rudd are

$$u = \exp\left(\left(r-\frac{1}{2}\sigma^2\tau\right)+\sigma\sqrt{\tau}\right) \qquad \text{and} \qquad d = \exp\left(\left(r-\frac{1}{2}\sigma^2\tau\right)-\sigma\sqrt{\tau}\right),$$

that is, the two moment equalities (1.1.7) and (1.1.8) are satisfied up to order $\mathcal{O}(\tau^2)$.

The Tian model: The Tian model adds the third moment matching as additional constraint, $pu^3+(1-p)\,d^3 = \exp\left(3\left(r+\sigma^2\right)\tau\right)$, so that we obtain

Proposition 1.1.16. *Under the Tian condition, the unique solutions to (1.1.7) and (1.1.8) are*

$$u = \frac{\phi_\tau\mathrm{e}^{r\tau}}{2}\left(1+\phi_\tau+\sqrt{\phi_\tau^2+2\phi_\tau-3}\right) \qquad \textit{and} \qquad d = \frac{\phi_\tau\mathrm{e}^{r\tau}}{2}\left(1+\phi_\tau-\sqrt{\phi_\tau^2+2\phi_\tau-3}\right),$$

where $\phi_\tau := \mathrm{e}^{\sigma^2\tau}$. Ignoring terms of orders $\mathcal{O}\left(\tau^{3/2}\right)$ in (1.1.7) and in (1.1.8), $p = \frac{1}{4}\left(2-3\sigma\sqrt{\tau}\right)$.

Remark 1.1.17. The Tian tree is not symmetric since $ud = \phi_\tau^2\mathrm{e}^{2r\tau}$ is different from one.

Convergence of CRR to the Black-Scholes model

In this section, we would like to see what happens to the multi-period binomial model developed in Section 1.1.2 when the time increment τ tends to zero. Put differently, we want to understand in what sense the discrete-time framework of the binomial model converges to some continuous-time model. As we mentioned above, Theorem 1.1.4 is not specific to a European call option, and we therefore consider here a European option V on the underlying asset S with maturity $T > 0$ and having some payoff $f(S_T)$ at maturity. We first prove the convergence in law of the discrete-time approximation to the continuous-time Black-Scholes model.

Theorem 1.1.18. *The discrete-time approximation of the stock price process in the (CRR) binomial tree converges in law to the Black-Scholes model when the time increment $\tau = T/N$ tends to zero (equivalently, as N tends to infinity).*

Proof. Without loss of generality, we may assume that the initial value S_0 of the stock price is equal to one. Let now λ be a real number, and write

$$\begin{aligned}\mathbb{E}\left(\mathrm{e}^{\mathtt{i}\lambda\log(S_N)}\right) &= \mathbb{E}\left(\exp\left(\mathtt{i}\lambda\log\prod_{n=0}^{N-1}\frac{S_{n+1}}{S_n}\right)\right)\\ &= \left(\mathbb{E}\left(\mathrm{e}^{\mathtt{i}\lambda\log(Z)}\right)\right)^N = \left(p\mathrm{e}^{\mathtt{i}\lambda\log(u)} + (1-p)\,\mathrm{e}^{\mathtt{i}\lambda\log(d)}\right)^N\\ &= \left(p\mathrm{e}^{\mathtt{i}\lambda\sigma\sqrt{\tau}+\mathcal{O}(\tau^{3/2})} + (1-p)\,\mathrm{e}^{-\mathtt{i}\lambda\sigma\sqrt{\tau}+\mathcal{O}(\tau^{3/2})}\right)^N,\end{aligned}$$

where Z is a Bernoulli random variable taking values in $\{d, u\}$ representing the returns of the stock price process between two time steps. In the second line, we appealed to the independence property of the increments of the stock price, and Corollary 1.1.14 is used in the third line. Let now $\tau = T/N$ and let N tend to infinity, Corollary 1.1.14 then implies

$$\lim_{N\to\infty}\mathbb{E}\left(\mathrm{e}^{\mathtt{i}\lambda\log(S_N)}\right) = \exp\left(\left(r-\frac{\sigma^2}{2}\right)\mathtt{i}\lambda T - \frac{\lambda^2\sigma^2 T}{2}\right),$$

so that we have the convergence of the characteristic functions and the theorem follows from the results in Appendix A.1.3. □

Remark 1.1.19.

(i) We only proved the convergence in law for the Cox-Ross-Rubinstein tree. A generalisation of this result can be found in Proposition 1.2.6 below.

(ii) The convergence in law (weak convergence) of the stock price implies the convergence of European option prices with continuous and bounded payoffs such as Put option prices, but not Call options, see Appendix A.1.3. Convergence of European Call options however follows by Call-Put parity. We shall give a direct proof of the convergence of the option pricing formula (1.1.2) later, in Theorem 1.1.21.

(iii) With n steps, Corollary 1.1.14 implies that the range of possible values for the stock price process is $[d^n S_0, u^n S_0] = \left[S_0 \mathrm{e}^{-\sigma\sqrt{nT}}, S_0 \mathrm{e}^{\sigma\sqrt{nT}}\right]$, and the grid becomes dense in the two-dimensional subspace $[0,T] \times (0,\infty)$ as the number of steps increases to infinity.

Theorem 1.1.20. *As the time increment τ tends to zero, the moment matching conditions* (1.1.7) *and* (1.1.8) *imply that the option price converges to the solution of the so-called Black-Scholes partial differential equation*

$$\partial_t V + rS_t \partial_S V + \frac{\sigma^2}{2} S_t^2 \partial_{SS}^2 V - rV_t = 0, \tag{1.1.9}$$

with boundary (payoff) condition $V_T = f(S_T)$.

Proof. Let $V(x,t)$ denote the value of the option price with at time t where the underlying stock price is worth x. Consider the backward scheme where the option price $V(x,t)$ at some time t is given in terms of the option prices $V(ux, t+\tau)$ and $V(dx, t+\tau)$ at time $t+\tau$, depending on whether the stock price has moved up or down:

$$V(x,t) = \mathrm{e}^{-r\tau}\Big(pV(ux, t+\tau) + (1-p)\,V(dx, t+\tau)\Big). \tag{1.1.10}$$

In the Cox-Ross-Rubinstein tree model, we take $u = \exp(\sigma\sqrt{\tau})$ and $d = \exp(-\sigma\sqrt{\tau})$. Furthermore, a Taylor series expansion at the point (x,t) gives

$$\begin{aligned} V\left(x\mathrm{e}^{\sigma\sqrt{\tau}}, t+\tau\right) &= V(x,t) + \left(\mathrm{e}^{\sigma\sqrt{\tau}} - 1\right) x\frac{\partial V}{\partial x} + \tau\frac{\partial V}{\partial t} + \frac{x^2}{2}\left(\mathrm{e}^{\sigma\sqrt{\tau}} - 1\right)^2 \frac{\partial^2 V}{\partial x^2} + o(\tau), \\ &= V(x,t) + \left(\sigma\sqrt{\tau} + \frac{1}{2}\sigma^2\tau\right) x\frac{\partial V}{\partial x} + \tau\frac{\partial V}{\partial t} + \frac{\sigma^2 x^2 \tau}{2}\frac{\partial^2 V}{\partial x^2} + o(\tau), \end{aligned}$$

where all the derivatives of the function C are evaluated at the point (x,t). In the first line, we have used the fact that $o\left(\left(\mathrm{e}^{2\sigma\sqrt{\tau}} - 1\right)^2\right) = o(\tau)$. In the second line, we have performed a Taylor series expansion of $\left(\mathrm{e}^{\sigma\sqrt{\tau}} - 1\right)$ for small τ. We have only expanded up to first order in τ. This is justified by the fact that the expressions for u and d are also of order 1 from Corollary 1.1.14. Likewise, the cross derivative $\frac{\partial^2}{\partial x \partial t}$ has been omitted. Similarly, we have

$$V\left(x\mathrm{e}^{-\sigma\sqrt{\tau}}, t+\tau\right) = V(x,t) - \left(\sigma\sqrt{\tau} - \frac{1}{2}\sigma^2\tau\right) x\frac{\partial V}{\partial x} + \tau\frac{\partial V}{\partial t} + \frac{\sigma^2 x^2 \tau}{2}\frac{\partial^2 V}{\partial x^2} + o(\tau).$$

If we now plug these two expansions back into (1.1.10) and omit the $o()$ terms, we obtain

$$\left(\mathrm{e}^{r\tau} - 1\right) V = \left((2p-1)\, x\sigma\sqrt{\tau} + \frac{\sigma^2 x\tau}{2}\right)\frac{\partial V}{\partial x} + \tau\frac{\partial V}{\partial t} + \frac{\sigma^2 x^2 \tau}{2}\frac{\partial^2 V}{\partial x^2} + o(\tau).$$

From Corollary 1.1.14, we know that $p = \frac{1}{2} + \frac{1}{2\sigma}\left(r - \frac{1}{2}\sigma^2\right)\sqrt{\tau} + \mathcal{O}(\tau)$. Therefore we obtain (after dividing by τ)

$$rV = rx\frac{\partial V}{\partial x} + \frac{\partial V}{\partial t} + \frac{\sigma^2 x^2}{2}\frac{\partial V^2}{\partial x^2} + o(1).$$

Taking the limit as τ tends to zero, and applying the boundary conditions, the theorem follows. □

We have seen (Theorem 1.1.18) how the stock price converges in law when the time increment tends to zero and how European Call options converge to the solution of some partial differential equation with suitable boundary conditions. We shall see in Chapter 3 how to solve (numerically) such an equation. The purpose of the following theorem is to provide a simple and elegant pricing formula that corresponds to the formula (1.1.2) when the number of steps tends to infinity.

Theorem 1.1.21. *When the increment τ tends to zero, the pricing formula* (1.1.2) *converges to*

$$V_0 = S_0\mathcal{N}\left(d_+\right) - K\mathrm{e}^{-rT}\mathcal{N}\left(d_-\right),$$

where

$$d_\pm := \frac{\log\left(S_0/K\right) + \left(r \pm \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}},$$

and where $\mathcal{N}$ represents the cumulative distribution function of the standard Gaussian random variable with zero mean and unit variance.

Proof. Recall the option pricing formula in the binomial tree with N nodes (Equation (1.1.3)):

$$S_0\mathrm{B}\left(N_0, N, \frac{(1-\pi)\,d}{R}\right) - \frac{K}{R^N}\mathrm{B}\Big(N_0, N, (1-\pi)\Big), \tag{1.1.11}$$

where $\mathrm{B}(\cdot, n, p)$ is the cumulative distribution function of a Binomial random variable $\mathcal{B}(n,p)$. In view of the formula in the theorem, it is clear that it is sufficient to prove that

$$\lim_{N\to\infty} \mathrm{B}\left(N_0, N, \frac{(1-\pi)\,d}{R}\right) = \mathcal{N}\left(d_+\right) \qquad \text{and} \qquad \lim_{N\to\infty} \mathrm{B}\Big(N_0, N, (1-\pi)\Big) = \mathcal{N}\left(d_-\right).$$

By Exercise 3 we know that there exists $\alpha \in [0,1)$ such that $N_0 = \frac{N\log(u) - \log(K/S_0)}{\log(u)-\log(d)} - \alpha$.

Let $(X_i)_{i\geq 1}$ be a sequence of independent random Bernoulli random variables with parameter $\widetilde{\pi}$. We know that $\overline{X}_n := \sum_{i=1}^n X_i$ follows $\mathcal{B}\left(n, \widetilde{\pi}\right)$, and hence $\mathbb{E}\left(\overline{X}_n\right) = n\widetilde{\pi} < \infty$ and $\mathbb{V}\left(\overline{X}_n\right) = n\widetilde{\pi}\left(1-\widetilde{\pi}\right) < \infty$. From Berry-Esséen inequality (see Theorem A.1.10) in the appendix, there exists a strictly positive constant C—universal and therefore independent of n—such that

$$\sup_x \left|\mathbb{P}\left(\frac{\overline{X}_n - n\widetilde{\pi}}{\sqrt{n\widetilde{\pi}\left(1-\widetilde{\pi}\right)}} \leq x\right) - \mathcal{N}(x)\right| \leq \frac{C\rho}{\sqrt{n}}, \tag{1.1.12}$$

where

$$\rho := \mathbb{E}\left(\frac{\left|X_1 - \widetilde{\pi}\right|^3}{\left(\widetilde{\pi}\left(1-\widetilde{\pi}\right)\right)^{3/2}}\right) = \frac{\widetilde{\pi}\left(1-\widetilde{\pi}\right)\left(\widetilde{\pi}^2 + \left(1-\widetilde{\pi}\right)^2\right)}{\left(\widetilde{\pi}\left(1-\widetilde{\pi}\right)\right)^{3/2}} = \frac{\widetilde{\pi}^2 + \left(1-\widetilde{\pi}\right)^2}{\sqrt{\widetilde{\pi}\left(1-\widetilde{\pi}\right)}}.$$

Let us now go back to the expression (1.1.11) and let us focus on the first probability. The random variable $\overline{X}_n$ is a Binomial random variable with parameters $n = N$ and $\widetilde{\pi} = \dfrac{(1-\pi)\,d}{R}$, and let $x = \dfrac{N_0 - n\widetilde{\pi}}{\sqrt{n\widetilde{\pi}\left(1-\widetilde{\pi}\right)}}$ in (1.1.12). In the CRR formula, let $\tau := T/N$ denote the time increment, so that Corollary 1.1.14 implies

$$\widetilde{\pi} = \frac{1}{2} - \frac{1}{2}\left(\sigma + \frac{r - \frac{1}{2}\sigma^2}{\sigma}\right)\sqrt{\tau} + \mathcal{O}(\tau),$$

which converges to $1/2$ as N tends to infinity (equivalently as τ tends to zero). Therefore the right-hand side in (1.1.12) tends to zero as the time increment becomes smaller. We conclude that

$$\lim_{N\to\infty} \mathbb{P}\left(\overline{X}_N \leq x\right) = \lim_{N\to\infty} \mathcal{N}\left(\frac{x - N\widetilde{\pi}}{\sqrt{N\widetilde{\pi}\left(1-\widetilde{\pi}\right)}}\right), \qquad \text{for any } x \in \mathbb{R},$$

and hence

$$\lim_{N\to\infty} \mathrm{B}\left(N_0, N, \frac{(1-\pi)\,d}{1+r}\right) = \lim_{N\to\infty} \mathcal{N}\left(\frac{N_0 - N\widetilde{\pi}}{\sqrt{N\widetilde{\pi}\left(1-\widetilde{\pi}\right)}}\right).$$

Taking now N_0 and the definition of $\widetilde{\pi}$ as above and making use of Corollary 1.1.14, we obtain the following Taylor series expansion as τ tends to zero:

$$\begin{aligned}
N\widetilde{\pi} &= \frac{T}{2\sigma\sqrt{\tau}}\left(r - \frac{\sigma^2}{2}\right) + \frac{T}{2\tau} + \mathcal{O}(1),\\
N_0 &= \frac{1}{2\sigma\sqrt{\tau}}\log(S_0/K) + \frac{T}{2\tau} + \mathcal{O}(1),\\
N\widetilde{\pi}\left(1-\widetilde{\pi}\right) &= \frac{T}{4\tau} + \mathcal{O}(1),
\end{aligned}$$

from which we deduce

$$\frac{N_0 - N\widetilde{\pi}}{\sqrt{N\widetilde{\pi}\left(1-\widetilde{\pi}\right)}} = \frac{1}{\sigma\sqrt{T}}\left(\log\left(\frac{S_0}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)T\right) + \mathcal{O}\left(\sqrt{\tau}\right),$$

which converges to d_+ given in the theorem as τ tends to zero, and the continuity of the map $\mathcal{N}$ finishes the proof. The proof of the other probability is analogous and left as an exercise. □

Example (Numerical example and convergence)**.** We consider here a European call option price written on the underlying stock price worth 100 at inception of the contract, with maturity $T = 1.1$ year and strike price $K = 95$. We assume that the continuous interest rate is $r = 3\%$ and that the volatility of the returns is $\sigma = 20\%$. To value such an option in a multi-period Cox-Ross-Rubinstein binomial tree, we consider N nodes, so that the time increment τ is equal to T/N. The price as the number of nodes tends to infinity is given by Theorem 1.1.21 and is equal to 12.6931. We are interested here in the convergence of the CRR model as the number of nodes grows large. Figure 1.2 provides us with a numerical example of the convergence of the tree to the Black-Scholes price given in Theorem 1.1.21, as a function of the number of nodes N. The corresponding MATLAB code is as follows:

Black-Scholes Call option formula

```
function C=bscall($S_0, K, r, \sigma, T$)
$d_1 = \left(\log(S_0/K) + \left(r + \frac{1}{2}\sigma^2\right)T\right)/\left(\sigma\sqrt{T}\right)$
$d_2 = \left(\log(S_0/K) + \left(r - \frac{1}{2}\sigma^2\right)T\right)/\left(\sigma\sqrt{T}\right)$
$C = S_0\mathcal{N}(d_1) - K\mathrm{e}^{-rT}\mathcal{N}(d_2)$;
```

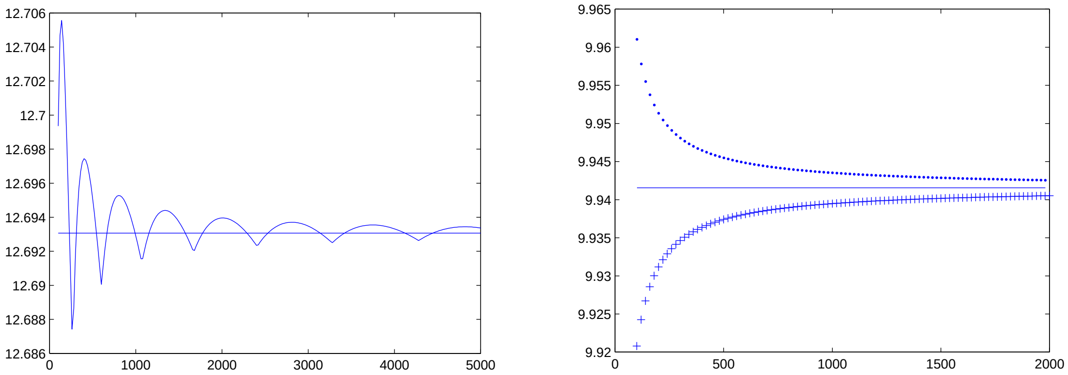

Figure 1.2: Convergence of the CRR price to Black-Scholes as a function of the number of nodes N. $K = 95$ in the left figure and the right figure is the at-the-money case $K = 100$. The dotted plot considers only odd numbers of steps whereas the crossed one considers even numbers of steps only.

Binomial tree call option pricing formula

```
function price = BinEuroCall($S0, K, r, \sigma, T, n$)
$\delta_T = T/n$;    $u = \mathrm{e}^{\sigma\sqrt{\delta_T}}$;    $d = 1/u$;    $p = (\mathrm{e}^{r\delta_T} - d)/(u - d)$;
lattice = zeros(n+1,n+1);
for i=0:n
    lattice(i+1,n+1) = max $\left(0, u^i d^{n-i} S_0 - K\right)$;
end;
for k=n-1:-1:0
    for i=0:k
        lattice(i+1,k+1) = p*lattice(i+2,k+2)+(1-p)*lattice(i+1,k+2);
    end;
end;
price = $\mathrm{e}^{-rT}$ lattice(1,1);
```

Output of the left graph on Figure 1.2

```
$S_0 = 100$;    $K = 95$;    $r = 0.03$;    $\sigma = 0.2$;    $T = 1.1$;
$\underline{n} = 100$; % minimum number of time steps
$\overline{n} = 5000$; % maximum number of time steps
$\delta_n = 20$;
BS = bscall($S_0, K, r, \sigma, T$);
BinomPrices = zeros(1,length([$\underline{n} : \delta_n : \overline{n}$]));
for $n = [\underline{n} : \delta_n : \overline{n}]$
    BinomPrices(temp) = BinEuroCall($S_0, K, r, \sigma, T, n$);
end;
plot([$\underline{n} : \delta_n : \overline{n}$], BinomPrices); hold on;
plot([$\underline{n} : \delta_n : \overline{n}$], ones(1,length([$\underline{n} : \delta_n : \overline{n}$]))*BS);
```

Remark 1.1.22. When the stock price is lognormally distributed with constant mean and constant variance, Theorem 1.1.21 gives a closed-form solution to the European call option pricing problem. In order to understand the (local) behaviour of option prices, sensitivity measures known as *Greeks* are used as fundamental tools. They are defined as follows:

$$\Delta := \frac{\partial V}{\partial S_0}, \quad \Gamma := \frac{\partial^2 V}{\partial S_0^2}, \quad \rho := \frac{\partial V}{\partial r}, \quad \Theta := \frac{\partial V}{\partial T}, \quad \upsilon := \frac{\partial V}{\partial \sigma}. \tag{1.1.13}$$

As we shall see later, they are also fundamental tools for hedging purposes, i.e. to protect oneself against moves in the underlying variables (S, r, σ).

Exercise 4. Consider a European call option with maturity $T > 0$ and strike $K > 0$ evaluated at time zero, written on a stock price process $(S_t)_{t\geq 0}$ following the Black-Scholes model (see Definition 1.1.11). Derive closed-form formulae for

- a European Put option with the same characteristics as the Call;
- the Greeks of the Call, defined in Remark 1.1.22.

The identity (to prove) $S_0\mathcal{N}'(d_+) = K\mathrm{e}^{-rT}\mathcal{N}'(d_-)$ might be helpful.

Solution. *The Put option price at time zero is*

$$P = K\mathrm{e}^{-rT}\mathcal{N}(-d_-) - S_0\mathcal{N}(-d_+),$$

and the Greeks read

$$\begin{aligned}
\Delta &= \mathcal{N}(d_+), & \upsilon &= S_0\sqrt{T}\mathcal{N}'(d_+),\\
\rho &= KT\mathrm{e}^{-rT}\mathcal{N}(d_-), & \Gamma &= \frac{\mathcal{N}'(d_+)}{\sigma S_0\sqrt{T}},\\
\Theta &= -\frac{\sigma S_0\mathcal{N}'(d_+)}{2\sqrt{T}} - rK\mathrm{e}^{-rT}\mathcal{N}(d_-).
\end{aligned}$$

Exercise 5. Consider a European call option with maturity $T > 0$ and strike $K > 0$ evaluated at time zero, written on a stock price process $(S_t)_{t\geq 0}$ following the Black-Scholes model (Definition 1.1.11). Consider the following values: $S_0 = 100$, $r = 0$ and $\sigma = 25\%$. Plot the convergence of the CRR tree when $K < 90$, $K = 100$ and $K = 110$. Plot also the convergence of the tree for even and odd increasing number of time steps. Comment the obtained results.

1.1.4 Adding dividends

The model we have assumed so far is a very simple model describing the evolution of an asset price. In practice, asset prices distribute dividends to shareholders. A consequence of this is that once the dividend is paid, the value of the stock price drops by the amount of the dividend. We shall see here how such a feature can be implemented in a binomial tree. There are two possible ways to take dividends into account. On the one hand one may consider that dividends are distributed continuously, so that the stock price does not suffer a sudden drop at some future time, but rather diffuses in time with a drift where r is replaced by $r - q$ (in the Black-Scholes model for instance). Note however that the discounting factor $\mathrm{e}^{-r\tau}$ remains the same since it only depends on risk-free bond prices, and not on the stock price. On the other hand, one may think of the dividends as distributed at some (fixed) future times. Two different models can be considered. Assume first that the dividend distributed between t_{i-1} and t_i is proportional to the value of the stock price at time t_i, i.e. is worth αS_i for some $\alpha > 0$. At time t_i the stock price is therefore worth $(1-\alpha)\, u^j d^{i-j} S_0$ instead of $u^j d^{i-j} S_0$ in the original (without dividend) binomial tree. It is easy to see that the recombining properties of the tree are preserved. If the dividend between t_{i-1} and t_i is not proportional to the stock price but rather is a fixed amount of cash, say $D > 0$, then the stock price at time t_i is worth $u^j d^{i-j} S_0 - D$ and the tree is not recombining any more.

1.2 Trinomial trees

A natural extension of the two-node (binomial) scheme above is to consider an l-node model. A popular one is $l = 3$, called the trinomial model, represented on Figure 1.3. Between two nodes i and $i + 1$, the ratio S_{i+1}/S_i takes values in $\{d, m, u\}$, where $d < m < u$, with probability by π_d, π_m and π_u. We would like to prove a theorem similar to the binomial case, namely

Conjecture 1.2.1. *Consider a European option V written on the underlying stock price S and with payoff $V_T(S_T)$ at maturity T. Let $R := \mathrm{e}^{rT}$ be the discounting factor In the absence of arbitrage opportunities, the one-period option pricing formula in the trinomial tree is*

$$V_0 = \frac{\pi_u V_T(uS_0) + \pi_m V_T(mS_0) + \pi_d V_T(dS_0)}{R}.$$

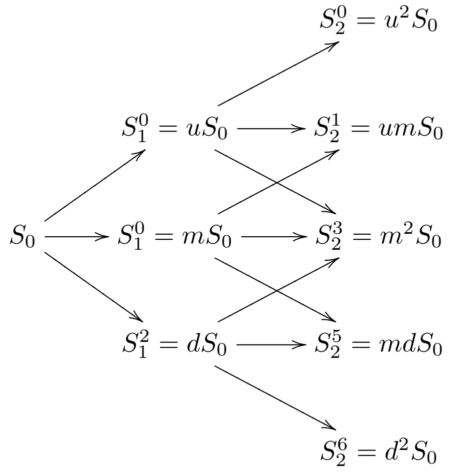

Figure 1.3: Two-period recombining trinomial tree.

However we cannot proceed as in the binomial tree case, i.e. construct a hedging portfolio with stocks and bonds that will replicate the option. In fact, if the stock price under consideration is a true martingale under some probability measure—i.e. $\mathbb{E}(S_T|\mathcal{F}_t) = S_t$ for all $0 \leq t \leq T$, where $\mathcal{F}_t$ represents the information available at time t ($(\mathcal{F}_t)_{t\geq 0}$ is called the filtration) —we know that any contingent claim can be valued using this very probability measure. As in the binomial scheme, we look at the first two moment matching conditions:

$$\begin{aligned} \pi_u u + \pi_m m + \pi_d d &= \mathrm{e}^{r\tau}, \\ \pi_u u^2 + \pi_m m^2 + \pi_d d^2 &= \mathrm{e}^{\left(2r+\sigma^2\right)\tau}. \end{aligned}$$

Together with the constraints $\pi_u + \pi_m + \pi_d = 1$ and $\pi_u, \pi_d, \pi_m > 0$, we obtain an ill-posed system of three equations with six unknown variables. Two popular models have however emerged, proposing additional constraints on the parameters in order to ensure the local consistency of the tree to the Black-Scholes model: the Boyle model and the Kamrad-Ritchken model.

Boyle model

It was developed by Phelim Boyle in 1986 [8]. The three additional conditions are $m = 1$, $ud = 1$ and $u = \exp\left(\lambda\sigma\sqrt{\tau}\right)$, where $\lambda > 0$ is a free parameter, called the *stretch parameter*. Note that the corresponding tree is recombining ($ud = m^2$).

Exercise 6. Let $W := \exp\Big(\left(2r + \sigma^2\right)\tau\Big)$. Show that the corresponding risk-neutral probabilities are given by

$$\pi_u = \frac{\left(W - R\right)u - \left(R - 1\right)}{\left(u - 1\right)\left(u^2 - 1\right)}, \quad \pi_d = \frac{\left(W - R\right)u^2 - \left(R - 1\right)u^3}{\left(u - 1\right)\left(u^2 - 1\right)}, \quad \text{and} \quad \pi_m = 1 - \pi_u - \pi_d.$$

What happens when $\lambda = 1$? Compute the limits of π_u and π_d as λ tends to infinity. What can you conclude from it?

Exercise 7. The Boyle model with $\lambda = \sqrt{3}$ is called the Hull-White model (see [38]). Show that the risk-neutral probabilities read

$$\pi_u = \frac{1}{6} + \sqrt{\frac{\tau}{3}} \frac{r - \frac{1}{2}\sigma^2}{2\sigma}, \quad \pi_d = \frac{1}{6} - \sqrt{\frac{\tau}{3}} \frac{r - \frac{1}{2}\sigma^2}{2\sigma}, \quad \text{and} \quad \pi_m = \frac{2}{3}$$

up to order $\mathcal{O}(\tau)$, for small enough τ.

Kamrad-Ritchken model

Under the risk-neutral measure, the random variable $\log(S_{t+\tau}/S_t)$ is Gaussian with mean $(r - \frac{1}{2}\sigma^2)\tau$ and variance $\sigma^2\tau$, i.e. $\log(S_{t+\tau}) = \log(S_t) + \xi$, where $\xi \sim \mathcal{N}((r - \frac{1}{2}\sigma^2)\tau, \sigma^2\tau)$. The Kamrad-Ritchken [41] symmetric trinomial tree approximates the random variable ξ by a discrete random variable $\widetilde{\xi}$ with the following distribution:

$$\widetilde{\xi} := \begin{cases} \lambda\sigma\sqrt{\tau}, & \text{with probability } \pi_u, \\ 0, & \text{with probability } \pi_m, \\ -\lambda\sigma\sqrt{\tau}, & \text{with probability } \pi_d, \end{cases}$$

where $\sigma > 0$ and $\lambda \geq 1$. Omitting terms of order higher (or equal) than $\mathcal{O}\left(\tau^2\right)$, we obtain the risk-neutral probabilities

$$\pi_u = \frac{1}{2\lambda^2} + \frac{r - \frac{1}{2}\sigma^2}{2\lambda\sigma}\sqrt{\tau}, \quad \pi_d = \frac{1}{2\lambda^2} - \frac{r - \frac{1}{2}\sigma^2}{2\lambda\sigma}\sqrt{\tau}, \quad \text{and} \quad \pi_m = 1 - \frac{1}{\lambda^2}. \tag{1.2.1}$$

Note that $\lambda < 1$ implies $\pi_m < 0$, which explains the condition $\lambda \geq 1$. In the case $\lambda = 1$, we have $\pi_m = 0$ and hence the trinomial tree reduces to a simple binomial tree.

Remark 1.2.2. As in the binomial model, we can show that the option pricing formula in Theorem 1.2.1 (extended to n nodes) converges to the Black-Scholes partial differential equation as the number of nodes tend to infinity, with an error of order $\mathcal{O}(\tau)$. Note further that the flexibility in the choice of λ implies different values for the three probabilities $\{\pi_d, \pi_m, \pi_u\}$. This corresponds to the fact that there exists, not one, but an infinity of equivalent martingale measures under which the stock price is a martingale. This corresponds to an incomplete market, as opposed to a complete one as in the binomial case, where only one such probability measure exists.

Remark 1.2.3. From a numerical point of view, it is easy to show that for a tree with n steps, the binomial scheme requires $n(n+1)/2$ additions and $n(n+1)$ multiplications, whereas the trinomial scheme requires $2n^2$ additions and $3n^2$ multiplications, so that the trinomial scheme is more demanding. When comparing a trinomial scheme with n steps to a binomial tree with $2n$ steps, it is however clear that the trinomial scheme requires less computational effort; Kamrad and Rithcken [41] have shown that it also performs better.

Remark 1.2.4. One may wonder whether there is an optimal choice of the parameter λ. It has to be chosen so that the convergence to the Black-Scholes (continuous-time) model is maximised.

Remark 1.2.5. In practice, binomial trees remain popular, sometimes over trinomial trees. This is due to the fact that the convergence is already fast enough for binomial trees, and hence the additional complexity of trinomial schemes is not always necessary for vanilla products. However this may not be true any longer for exotic options.

In Theorem 1.1.18, we proved that the CRR binomial tree converges in law to the Black-Scholes model as the time step tends to zero. The following proposition generalised this result and emphasises the order needed in the approximations of the probabilities and amplitudes.

Proposition 1.2.6. *Consider a multinomial tree with $N \geq 2$ branching possibilities, and denote $(p_i)_{1\leq i\leq N}$ the probabilities associated to the amplitudes $(\xi_i)_{1\leq i\leq N}$, such that for any $i = 1, \dots, N$, the following Taylor series expansions hold around $\tau = 0$:*

$$\begin{aligned} p_i &= p_{i,0} + p_{i,1}\sqrt{\tau} + p_{i,2}\tau + o(\tau), \\ \xi_i &= 1 + \xi_{i,1}\sqrt{\tau} + \xi_{i,2}\tau + o(\tau). \end{aligned}$$

Then the first two moment-matching equations

$$\sum_{i=1}^{N} p_i\xi_i = \mathrm{e}^{r\tau} \qquad \text{and} \qquad \sum_{i=1}^{N} p_i\xi_i^2 = \mathrm{e}^{(2r+\sigma^2)\tau}.$$

are equivalent to the convergence in law (as τ tends to zero) of the tree to the Black-Scholes model.

Proof. From the proof of Theorem 1.1.18, convergence in law will be obtained as soon as the following equality holds:

$$\sum_{i=1}^{N} p_i \mathrm{e}^{\mathtt{i}\lambda\log(\xi_i)} = 1 + \left(\mathtt{i}\lambda\left(r - \frac{\sigma^2}{2}\right) - \frac{\lambda^2\sigma^2}{2}\right)\tau + o(\tau). \tag{1.2.2}$$

From the assumptions on the form of the series expansion for p_i and ξ_i, the following constraints are immediate:

$$\sum_{i=1}^{N} p_{i,0} = 1 \qquad \text{and} \qquad \sum_{i=1}^{N} p_{i,1} = \sum_{i=1}^{N} p_{i,2} = 0.$$

Now we have, for any $i = 1, \dots, N$,

$$\begin{aligned} \mathrm{e}^{\mathtt{i}\lambda\log(\xi_i)} &= \exp\left\{\mathtt{i}\lambda\left(\xi_{1,i}\sqrt{\tau} + \xi_{2,i}\tau - \frac{1}{2}\xi_{i,1}^2\tau\right) + o(\tau)\right\} \\ &= 1 + \mathtt{i}\lambda\xi_{i,1}\sqrt{\tau} + \left(\mathtt{i}\lambda\left(\xi_{i,2} - \frac{\xi_{i,1}^2}{2}\right) - \frac{\lambda^2}{2}\xi_{i,1}^2\right)\tau + o(\tau). \end{aligned}$$

Therefore the local consistency equality (1.2.2) holds if and only if each powers of τ exactly match:

$$\begin{cases} \displaystyle\sum_{i=1}^{N} p_{i,0}\xi_{i,1} = 0, \\ \displaystyle\sum_{i=1}^{N} p_{i,0}\left(\mathtt{i}\lambda\left(\xi_{i,2} - \frac{\xi_{i,1}^2}{2}\right) - \frac{\lambda^2}{2}\xi_{i,1}^2\right) + \mathtt{i}\lambda\sum_{i=1}^{N} p_{i,1}\xi_{i,1} = \mathtt{i}\lambda\left(r - \frac{\sigma^2}{2}\right) - \frac{\lambda^2\sigma^2}{2}, \end{cases}$$

which in turn is equivalent to

$$\begin{cases}\displaystyle\sum_{i=1}^{N} p_{i,0}\xi_{i,1} = 0,\\ \displaystyle\sum_{i=1}^{N} p_{i,0}\left(\xi_{i,2} - \frac{\xi_{i,1}^2}{2}\right) + \sum_{i=1}^{N} p_{i,1}\xi_{i,1} = r - \frac{\sigma^2}{2},\\ \displaystyle\sum_{i=1}^{N} p_{i,0}\xi_{i,1}^2 = \sigma^2,\end{cases}$$

and using the third equality in the second one, we finally obtain

$$\begin{cases}\displaystyle\sum_{i=1}^{N} p_{i,0}\xi_{i,1} = 0,\\ \displaystyle\sum_{i=1}^{N} p_{i,0}\xi_{i,2} + \sum_{i=1}^{N} p_{i,1}\xi_{i,1} = r,\\ \displaystyle\sum_{i=1}^{N} p_{i,0}\xi_{i,1}^2 = \sigma^2.\end{cases} \tag{1.2.3}$$

Consider now the two moment matching equations

$$\sum_{i=1}^{N} p_i\xi_i = \mathrm{e}^{r\tau} \qquad \text{and} \qquad \sum_{i=1}^{N} p_i\xi_i^2 = \mathrm{e}^{(2r+\sigma^2)\tau}.$$

Equating powers of τ gives the following set of equations:

$$\begin{cases}\displaystyle\sum_{i=1}^{N} p_{i,0}\xi_{i,1} = 0,\\ \displaystyle\sum_{i=1}^{N} \left(p_{i,0}\xi_{i,2} + p_{i,1}\xi_{i,1}\right) = r,\\ \displaystyle 2\sum_{i=1}^{N} p_{i,0}\xi_{i,1} = 0,\\ \displaystyle\sum_{i=1}^{N} \left[2p_{i,1}\xi_{i,1} + p_{i,0}\left(\xi_{i,1}^2 + 2\xi_{i,2}\right)\right] = 2r + \sigma^2,\end{cases}$$

which is clearly equivalent to (1.2.3), and the proposition follows. □

1.3 Overture on stability analysis

We give here a slightly different approach to trees, that will serve as a first introduction to stability analysis, which we shall study more in details in Section 3.3.7. Consider the Black-Scholes model for the stock price S defined in Definition 1.1.11. For any $t \geq 0$, define the logarithm of the stock price $X_t := \log(S_t)$. Along a recombining ($ud = 1$) and symmetric ($m = 1$) trinomial tree, consider a node at time t, where the log-stock price is equal to X_t. At the next time step $t + \tau$, we have $X_{t+\tau} \in \left\{X_{t+\tau}^u, X_{t+\tau}^d, X_{t+\tau}^m\right\}$, where $X_{t+\tau}^u = \log(uS_t/S_t) = \log(u) =: \delta_x$, $X_{t+\tau}^d = -\delta_x$ and $X_{t+\tau}^m = 0$. Therefore $u = \mathrm{e}^{\delta_x}$. Over a short period of time τ, it is easy to show that

$$\mathbb{E}_t(X_{t+\tau}) = \nu\tau, \qquad \text{and} \qquad \mathbb{V}_t(X_{t+\tau}) = \sigma^2\tau + \nu^2\tau^2,$$

where $\nu := r - \frac{\sigma^2}{2}$. The moment matching conditions hence become

$$\begin{aligned}(\pi_u - \pi_d)\delta_x &= \nu\tau,\\ (\pi_u + \pi_d)\delta_x^2 &= \sigma^2\tau + \nu^2\tau^2,\\ \pi_u + \pi_d + \pi_m &= 1,\end{aligned}$$

with obvious unique solution

$$\begin{aligned}\pi_u &= \alpha\left(1 + \frac{\nu^2\tau}{\sigma^2} + \frac{\nu\delta_x}{\sigma^2}\right),\\ \pi_m &= 1 - 2\alpha\left(1 + \frac{\nu^2\tau}{\sigma^2}\right),\\ \pi_d &= \alpha\left(1 + \frac{\nu^2\tau}{\sigma^2} - \frac{\nu\delta_x}{\sigma^2}\right),\end{aligned}$$

where $\alpha := \frac{\sigma^2\tau}{2\delta_x^2}$. Although π_u is always strictly positive, the probability π_d may be negative. We therefore need to impose some conditions on the ratio τ/δ_x. Suppose indeed that ν and σ are fixed, and a specified time increment τ is given. Then if one takes the space increment δ_x larger than $\nu\tau + \sigma^2/\nu$, the probability π_d becomes negative and the tree is not properly defined any more. Note further that the tree is binomial in the case $\alpha = 1/2$, i.e. $\sigma^2\tau = \delta_x^2$.

Chapter 2

Monte Carlo methods

In Section 0.2, we showed that pricing a European option was tantamount to computing a conditional expectation. More precisely, we outlined the fact that—under some conditions—there exists a probability measure under which the value of a European contract today was worth the (discounted) value of its final payoff. In Chapter 1, we made this even more precise in a discrete-time setting for binomial and trinomial trees. This indeed turn out to be the fundamental ingredient of the backward scheme for multinomial trees. Monte Carlo methods are based on a similar idea. Consider a European option, the payoff of which at maturity $T > 0$ is given by $f(S_T)$, where f is a function from : $\mathbb{R}_+$ to $\mathbb{R}_+$ and $(S_t)_{t\geq 0}$ represents the stock price process. No-arbitrage theory tells us that the value today of such a contract is—barring the discounting factor—$\mathbb{E}\left(f(S_T)\right)$. If we are able to determine the distribution of the random variable S_T, then the computation becomes straightforward. Monte Carlo methods provide a tool to simulate such a distribution starting from the (known) initial value S_0.

2.1 Generating random variables

2.1.1 Uniform random number generator

Generating uniform random variables

The canonical random variable used for simulation purposes is the uniform random variable. Let $a < b$ be two real numbers and consider a random variable $U_{[a,b]}$ distributed uniformly on the closed interval $[a, b]$. This means that its density reads $f_{U_{[a,b]}}(x) = (b-a)^{-1}$, for all $x \in [a, b]$ and is null outside this interval. A straightforward scaling shows that the equality (in law) $U_{[a,b]} = a + (b-a)\, U_{[0,1]}$ always holds and hence it is sufficient to consider $a = 0$ and $b = 1$. Many algorithms exist to generate a random number on the interval $[0, 1]$. A popular (and robust) one is called the *Linear Congruential Generator*. Let $m > 0$, $a > 0$ and $n_0 \in \{1, \dots, m-1\}$ be integers such that a and m have no common factors, and define the sequence $(n_i)_{i\in\mathbb{N}}$ recursively by

$$n_i = an_{i-1} \quad [m], \qquad \text{for all } i \geq 1,$$

where $[m]$ means that the equality is taken modulo m. It is clear that for any $i \geq 1$, n_i lies in the set $\{1, 2, \dots, m-1\}$, and that the sequence (n_i) is periodic with period smaller (or equal) than $m-1$. We shall say that the sequence has *full period* if its period is exactly equal to $m-1$. It can be shown that the sequence has a full period if and only if m is a prime number, $a^{m-1}-1$ is divisible by m and for any $j = 1, \dots, m-2$, $a^j - 1$ is not divisible by m. Let us now define $x_i := n_i/m$ for any $i \geq 0$. Then clearly $x_i \in (0,1)$ for all $i \geq 0$. The sequence $(x_i)_{i\geq 0}$ is called a *pseudo-$U_{[0,1]}$ random number sequence* if and only if the sequence $(n_i)_{i\geq 0}$ has full period. We can alternatively define the random sequence $(x_i)_{i\geq 1}$ recursively by

$$x_{i+1} = ax_i - \lfloor ax_i \rfloor, \qquad \text{with } x_0 \in (0,1), \tag{2.1.1}$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than x.

Remark 2.1.1. The general formulation of a linear congruential generator is actually $n_i = (an_{i-1} + b) \quad [m]$, for all $i \geq 1$, where b is also an integer, and where m is not necessarily a prime number. Knuth [44] gave conditions under the generalised linear congruential generator in order to obtain a sequence with full period. As shown in [49], taking $b = 0$ does not entail much loss of generality, but does make the computations faster. This explains why we consider it to be null here, as is usually done in practice.

The table below shows some popular examples of couples (a, m) used in practice.

Modulus m	Multiplier a	Reference
$2^{31}-1$	16807	Lewis, Goodman, Miller (1969)
$2^{31}-1$	39373	Fishman, Moore (1986)
$2^{31}-1$	742938285	Fishman, Moore (1986)
2147483399	40692	L'Ecuyer (1988)

Remark 2.1.2. Microsoft Visual Basic (see `http://support.microsoft.com/kb/231847`) uses a linear congruential generator for the random sequence $(n_i)_{i\geq 1}$ defined recursively by the relation $n_{i+1} = an_i + b \quad [m]$, with $n_0 = 327680$, $a = 1140671485$, $b = 12820163$ and $m = 2^{24}$.

Remark 2.1.3. In order to implement the linear congruential algorithm, one has to make sure that the numbers computed falls with the precision of the computer. In order to generate the random numbers, one has to compute ax_i in(2.1.1). Given the large value of a used in practice (see the table above), this product might go beyond the computer precision. We shall not investigate this issue, but we refer the interested reader to [30, pages 44-46] for more details.

Remark 2.1.4. Even though linear congruential generators are widely used today, one has to be aware of their limitations. In particular a common feature to every such algorithm is the so-called *lattice structure*. Consider a sequence of numbers $(x_1, x_2, \dots, x_p)$ generated by linear congruence, where p indicates the period. Figure 2.1 represents all the overlapping pairs (x_i, x_{i+1}) for $i = 1, \dots, p-1$ in the unit square. This effect is called the lattice structure and Marsaglia [49] has precisely characterised the subspace of the unit hypercube in $\mathbb{R}^d$, covered by all the d-uples $(x_i, \dots, x_{i+d})$ for $i = 1, \dots, p-d$.

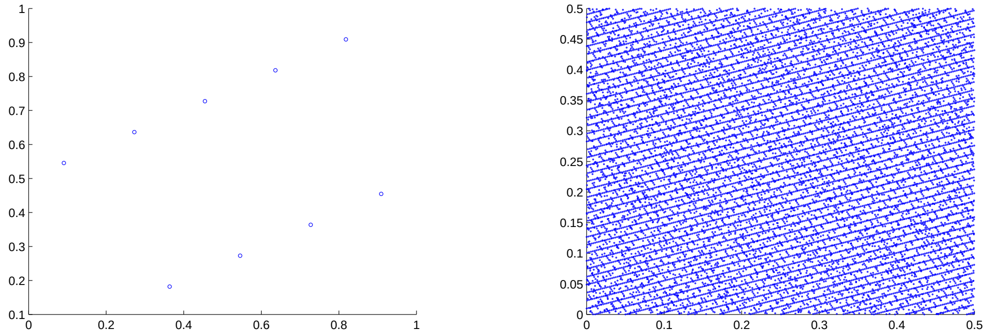

Figure 2.1: Evidence of lattice structure on the square. Left: $a = 6$, $m = 11$ and the period is 10. Right: $a = 1277$, $m = 131072$ and the period is 32768. On the right-hand side, we have zoomed on the smaller square $((0,0),(0.5,0.5))$.

2.1.2 Normally distributed random variables and correlation

Let us now consider a Gaussian random variable X with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$. Recall that the corresponding density reads

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad \text{for all } x \in \mathbb{R}.$$

We now wish to construct an algorithm capable of generating samples from this distribution. As in the uniform case, using the straightforward scaling property $\mathcal{N}(\mu, \sigma^2) \overset{\Delta}{=} \mu + \sigma\mathcal{N}(0,1)$ we may reduce our study to the case $\mu = 0$ and $\sigma = 1$.

Convolution method

From the central limit theorem (see Appendix A.1.4), a Gaussian random variable with zero mean and unit variance can easily be approximated by summing (and norming) many independent uniform random variables. Note however that the central limit theorem ensures convergence in distribution, but not almost sure convergence (see Appendix A.1). Also, from a computational point of view, the convolution method above requires the simulation of many (iid) uniform random variables in order to achieve accurate convergence.

Exercise 8. Let $(X_i)_{i\geq 1}$ be a sequence of iid uniform random variables on the interval $[-1,1]$. Define the sequence $(Z_n)_{i\geq 1}$ as in (A.1.1). Compute the expectations and variances of X_n and Z_n for any $n \geq 1$ and plot the density of the Gaussian approximation Z_n for different values of n.

Box-Muller method

As pointed out above, the convolution method may require a large number of uniform random variables to compute in order to achieve accurate precision. The Box-Muller method—stated in the following proposition—only requires the evaluation of two such random variables.

Proposition 2.1.5. *Let X_1 and X_2 be two independent uniform random variables on $(0,1)$. Define*

$$Z_1 := \sqrt{-2\log(X_1)}\cos(2\pi X_2) \qquad \text{and} \qquad Z_2 := \sqrt{-2\log(X_1)}\sin(2\pi X_2).$$

Then Z_1 and Z_2 are two independent Gaussian random variables with zero mean and unit variance.

Proof. We shall use the notations $\mathbf{x} := (x_1, x_2) \in \mathbb{R}^2$ and $\mathbf{z} := (z_1, z_2) \in \mathbb{R}^2$ throughout this proof. Define the function $h : (\mathbf{x}) \in [0,1]^2 \mapsto \left(\sqrt{-2\log(x_1)}\cos(2\pi x_2), \sqrt{-2\log(x_1)}\sin(2\pi x_2)\right)$. Its inverse $h^{-1} = \left(h_1^{-1}, h_2^{-1}\right)$ has the representation

$$h^{-1}(\mathbf{z}) = \left(\exp\left(-\frac{z_1^2+z_2^2}{2}\right), \frac{1}{2\pi}\text{atan}\left(\frac{z_2}{z_1}\right)\right), \quad \text{for any } (z_1,z_2) \in \mathbb{R}^2 \setminus \{(0,0)\},$$

and $h^{-1}(\mathbf{0}) = (1,0)$. For fixed $z_2 \neq 0$ and $z_1 = 0$, the second component is understood as the limit as z_1 tends to zero. The Jacobian matrix then reads

$$J = \begin{pmatrix} \partial_{z_1} h_1^{-1} & \partial_{z_2} h_1^{-1} \\ \partial_{z_1} h_2^{-1} & \partial_{z_2} h_2^{-1} \end{pmatrix}(z_1,z_2) = \begin{pmatrix} -z_1 \exp\left(-\frac{z_1^2+z_2^2}{2}\right) & -z_2 \exp\left(-\frac{z_1^2+z_2^2}{2}\right) \\ -\frac{1}{2\pi}\frac{z_2}{z_1^2+z_2^2} & \frac{1}{2\pi}\frac{z_1}{z_1^2+z_2^2} \end{pmatrix},$$

(recall that $\text{atan}(x)' = (1+x^2)^{-1}$), and its determinant simplifies to

$$\det(J) = \frac{1}{2\pi}\exp\left(-\frac{z_1^2+z_2^2}{2}\right).$$

Let f be the density of the couple (X_1, X_2) (i.e. $f \equiv 1$). For any subset $B \subset \mathbb{R}^2$, we have

$$\begin{aligned} \mathbb{P}(\mathbf{z} \in B) &= \mathbb{P}(h(\mathbf{x}) \in B) = \mathbb{P}\left(\mathbf{x} \in h^{-1}(B)\right) \\ &= \int_{h^{-1}(B)} f(\mathbf{x})\mathrm{d}x = \int_B f(h^{-1}(\mathbf{z}))\left|\det(J)\right| \mathrm{d}\mathbf{z} \\ &= \int_B \frac{1}{2\pi}\exp\left(-\frac{z_1^2+z_2^2}{2}\right)\mathrm{d}z_1\mathrm{d}z_2, \end{aligned}$$

which is the cumulative distribution function of two independent Gaussian random variables. □

Correlated Gaussian random variables

The Box-Muller method above gives us a way to construct two independent Gaussian random variables X_1 and X_2 (say with zero mean and unit variance). We now wish to generate a third Gaussian random variable, which is correlated with X_1. Recall that the covariance and the correlation between two random variables X and Y are defined by

$$\operatorname{cov}(X,Y) := \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \qquad \text{and} \qquad \rho(X,Y) := \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$

The following exercise (to do) shows how to construct such a correlated random variable.

Exercise 9. Let X_1 and X_2 be two independent Gaussian random variables with mean μ_1 and μ_2, and variance σ_1^2 and σ_2^2. Define the random variable $X_3 = \rho X_1 + \sqrt{1-\rho^2}X_2$ for some $\rho \in [-1,1]$. Determine the distribution of X_3 and its correlation with X_1 and X_2.

Solution. *It is well known that the sum of two independent Gaussian random variables is Gaussian, so that X_3 is Gaussian. By linearity of the expectation operator, we have $\mathbb{E}(X_3) = \rho\mathbb{E}(X_1) + \sqrt{1-\rho^2}\mathbb{E}(X_2) = \rho^2\mu_1 + (1-\rho^2)\mu_2$. Since the two random variables X_1 and X_2 are independent, the variance of any linear combination is the linear sum of the variances, i.e. $\mathbb{V}(X_3) = \rho^2\mathbb{V}(X_1) + (1-\rho^2)\mathbb{V}(X_2) = \rho^2\sigma_1^2 + (1-\rho^2)\sigma_2^2$. Similar computations show that the correlation between X_3 and X_1 (respectively between X_3 and X_2) is $\rho_{1,3} = \rho$ (respectively $\rho_{2,3} = \sqrt{1-\rho^2}$).*

Assume now that we are able to generate $n \geq 1$ independent Gaussian random variables with null mean and unit variance. We will present generic methods to do so in the next section. Let us call $\mathbf{X} := (X_1, \ldots, X_n)^{\mathbf{T}} \in \mathbb{R}^n$ such a vector. We wish to construct a new vector $\mathbf{Y} \in \mathbb{R}^n$ of Gaussian random variable with variance-covariance matrix $\mathbf{\Sigma} = (\sigma_{ij}) \in \mathcal{M}_n(\mathbb{R})$, i.e. such that $\sigma_{ij} := \operatorname{cov}(Y_i, Y_j)$ for any $1 \leq i,j \leq n$. The following properties are immediate:

- the matrix $\mathbf{\Sigma}$ is symmetric, i.e. $\mathbf{\Sigma}^{\mathbf{T}} = \mathbf{\Sigma}$;
- $\sigma_{ii} = 1$ for any $1 \leq i \leq n$ (normalisation);
- the matrix $\mathbf{\Sigma}$ is positive semi-definite, i.e. $\mathbf{x}^{\mathbf{T}}\mathbf{\Sigma}\mathbf{x} \geq 0$, for any $\mathbf{x} \in \mathbb{R}^n$.

Let us first note that the random variable $\overline{X} := \alpha_1 X_1 + \ldots + \alpha_n X_n$ is Gaussian with expectation zero and variance $\mathbb{V}\left(\overline{X}\right) = \alpha_1^2 + \ldots + \alpha_n^2$. Let $\mathbf{A}$ be a matrix in $\mathcal{M}_n(\mathbb{R})$ and define $\mathbf{Y} := \mathbf{A}^{\mathbf{T}}\mathbf{X}$,

Lemma 2.1.6. *The vector $\mathbf{Y}$ is Gaussian with variance-covariance matrix $\mathbf{\Sigma} = \mathbf{A}^{\mathbf{T}}\mathbf{A}$.*

Proof. Let us denote $\mathbf{A} = (\alpha_{i,j})_{1\leq i,j\leq n}$. For $i = 1, \ldots, n$, since $Y_i = \sum_{k=1}^n \alpha_{ki} X_k$ it is clear that

Y_i is Gaussian with mean zero and variance $\sum_{k=1}^{n} \alpha_{ki}^2$. For any $1 \leq i, j \leq n$, we also have

$$\begin{aligned}
\operatorname{cov}(Y_i, Y_j) &= \mathbb{E}(Y_i Y_j) - \mathbb{E}(Y_i)\mathbb{E}(Y_j) \\
&\mathbb{E}\left[\left(\sum_{k=1}^{n} \alpha_{ki} X_k\right)\left(\sum_{l=1}^{n} \alpha_{lj} X_l\right)\right] - \mathbb{E}\left(\sum_{k=1}^{n} \alpha_{ki} X_k\right)\mathbb{E}\left(\sum_{k=1}^{n} \alpha_{kj} X_k\right) \\
&= \sum_{k=1}^{n} \alpha_{ki}\alpha_{kj},
\end{aligned}$$

where we have used the independence properties of vector $\mathbf{X}$ and the fact that its expectation is the zero vector. This sum corresponds exactly to the (i, j) element of the $n \times n$ matrix $\mathbf{A^T A}$, and hence the lemma follows. □

Conversely, this lemma implies that if we want to generate a vector $\mathbf{Y} \in \mathbb{R}^n$ of Gaussian random variable with variance-covariance matrix $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$, then it suffices to determine the matrix $\mathbf{A}$ such that $\mathbf{\Sigma} = \mathbf{A^T A}$. The following theorem gives the solution to the problem.

Theorem 2.1.7. *Let $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive definite matrix. Then there exists an upper triangular matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ such that $\mathbf{\Sigma} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$.*

Proof. Let $\mathbf{\Sigma}$ be as in the theorem. From linear algebra, we know that there exists an upper triangular matrix $\mathbf{U} \in \mathcal{M}_n(\mathbb{R})$ and a diagonal matrix $\mathbf{D} \in \mathcal{M}_n(\mathbb{R})$ such that the so-called *LU decomposition* $\mathbf{\Sigma} = \mathbf{U^T D U}$ holds. We can rewrite this as

$$\mathbf{\Sigma} = \left(\mathbf{U^T}\sqrt{\mathbf{D}}\right)\left(\sqrt{\mathbf{D}}\mathbf{U}\right) = \left(\sqrt{\mathbf{D}}\mathbf{U}\right)^{\mathbf{T}}\left(\sqrt{\mathbf{D}}\mathbf{U}\right),$$

and the theorem follows with $\mathbf{A} = \sqrt{\mathbf{D}}\mathbf{U}$. □

We now give an algorithm to compute the square-root of a symmetric positive definite matrix $\mathbf{\Sigma} = (\sigma_{ij})$. Let us define the matrix $\mathbf{L} \in \mathcal{M}_n(\mathbb{R})$ by the following steps:

(i) start with the first column: $l_{11} := \sqrt{\sigma_{11}}$ and $l_{i1} := \sigma_{i1}/l_{11}$ for all $i = 2, \dots, n$;

(ii) consider now the j-th column: $l_{jj} := \left(\sigma_{jj} - \sum_{k=1}^{j-1} l_{jk}^2\right)^{1/2}$ and $l_{ij} := l_{jj}^{-1}\left(\sigma_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}\right)$ for any $i = j+1, \dots, n$ and $j = 2, \dots, n-1$;

(iii) compute finally the last element in the last column: $l_{nn} := \left(\sigma_{nn} - \sum_{k=1}^{j-1} l_{nk}^2\right)^{1/2}$;

(iv) take the transpose $\mathbf{A} := \mathbf{L^T}$.

Exercise 10. Prove that the above algorithm outputs the correct matrix $\mathbf{\Sigma}$.

Solution. *It follows from the explicit computation of the elements of the matrix Σ from the product* $\mathrm{U^T D U}$.

2.1.3 General methods

We have seen above how to generate Gaussian random variables from uniformly distributed random variables. These methods were specific to the Gaussian case, and we now wish to provide tools valid for any distribution. Let U be uniformly distributed on the interval $[0,1]$ and assume that we wish to generate the random variable X with cumulant distribution function $F : A \to [0,1]$, where A represents the support of the distribution (i.e. the range of possible values the random variable can take). We are looking for a mapping $f : [0,1] \to A$ satisfying the equality $f(U) = X$, i.e.

$$\mathbb{P}\left(f(U) \leq x\right) = F(x), \qquad \text{for all } x \in A. \tag{2.1.2}$$

If the mapping f is bijective and increasing, then $\mathbb{P}\left(f(U) \leq x\right) = \mathbb{P}\left(U \leq f^{-1}(x)\right) = f^{-1}(x)$. This expression suggests to take $f^{-1} \equiv F$ as a candidate. Let us distinguish the following cases:

(i) if the cdf F is continuous and strictly increasing, then the inverse mapping F^{-1} exists, and one simply takes $f \equiv F^{-1}$;

(ii) if the cdf F is continuous but not injective, then we can not necessarily define its inverse F^{-1}. However, since a cumulant distribution function is always right-continuous (see Appendix A.1.1), we may define the generalised inverse $F^{-1} : y \in [0,1] \mapsto \inf\{x \in A : F(x) = y\}$, and the relation (2.1.2) is clearly satisfied;

(iii) if the function F is discontinuous, i.e. the random variable X is discrete, we may represent it as $F(x) = \sum_{n\geq 1: x \leq x_n} \mathbb{P}(x_n)$, where $\{x_1, \ldots, x_m, \ldots\}$ is the set of possible outcomes of X. The generalised inverse can not be defined as in (ii). However, we may define it as $F^{-1}(y) := \inf\{x : F(x) \geq y\}$. The right-continuity of cumulative distribution functions ensures the existence of such an inverse and we leave it to the reader to check that the equality (2.1.2) is satisfied.

Remark 2.1.8. Consider the case where the cumulative distribution function $F : \mathbb{R} \to [0,1]$ is constant on some interval $[a,b]$ and strictly increasing on $\mathbb{R} \setminus [a,b]$. Then we have $\mathbb{P}(a \leq X \leq b) = F(b) - F(a) = 0$.

Example. We have seen in the sections above how to generate Gaussian distributed samples using the convolution or the Box-Muller method. We apply here the inverse transform method to generate sample of a Gaussian random variable with zero mean and unit variance. We know that its cumulative distribution function is given by

$$\mathcal{N}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u, \qquad \text{for any } x \in \mathbb{R}.$$

Since the function $\mathcal{N}$ is clearly smooth and bijective on the whole real line, the inverse mapping function is simply given by $f \equiv \mathcal{N}^{-1}$. From a computational point of view, though, one has to be able to compute this inverse. We shall not push the analysis further here but note that there are fortunately many algorithms and approximate formulae to compute such an inverse function, most notably Peter John Acklam's, available (in many computing languages) at home.online.no/~pjacklam/notes/invnorm.

Exercise 11. Let U be a uniform distribution on the interval $[0,1]$. Determine the mapping f such that $f(U)=X$, where the random variable X is given by the following.

(i) X is exponentially distributed with mean λ, to that $F(x)=1-\mathrm{e}^{x/\lambda}$ for all $x\geq 0$.

(ii) $F(x)=\dfrac{2}{\pi}\mathrm{asin}\left(\sqrt{x}\right)$, for any $x\in[0,1]$.

Solution.

(i) $f(u)=-\lambda\log(u)$*;*

(ii) This example corresponds to the so-called Arcsine law and the cumulative distribution function F corresponds to the distribution of the time at which a standard Brownian motion attains its maximum over the time interval $[0,1]$*. In this case* $f(u)=\dfrac{1}{2}\left(1-\cos(\pi U)\right)$.

Exercise 12. Let U be a uniform distribution on the interval $[0,1]$. Recall that a random variable N is Poisson distributed with parameter $\lambda>0$ if and only if the equality $\mathbb{P}\left(N=n\right)=\mathrm{e}^{-\lambda}\frac{\lambda^n}{n!}$ holds for any integer $n\geq 0$. Let $\left(T_i\right)_{i\geq 1}$ be a sequence of iid exponential random variables with parameter $\lambda>0$. Define the (continuous-time) family of random variables $(N_t)_{t\geq 0}$ by

$$N_t:=\sum_{n\geq 1}n\mathbb{1}_{\{T_1+\ldots+T_n\leq t<T_1+\ldots+T_{n+1}\}},\qquad \text{for any } t\geq 0.$$

(i) Prove that the family $(N_t)_{t\geq 0}$ follows a Poisson distribution with parameter λt.

(ii) Construct an algorithm to generate such a Poisson random variable.

Remark 2.1.9. Note that this method requires the exact same number of uniform random variables than the output random variables one wishes to generate.

2.2 Random paths simulation and option pricing

2.2.1 Simulation and estimation error

This section is a first brief encounter with the simulation of random paths. In Section 1.1.3, we have introduced the Black-Scholes model (see Definition 1.1.11). In this model, the stock price dynamics $\left(S_t\right)_{t\geq 0}$ is such that at any time $t\geq 0$, the random variable S_t is lognormally distributed:

$$S_t=S_0\exp\left(\left(r-\frac{\sigma^2}{2}\right)t+\sigma\sqrt{t}Z\right),\qquad(2.2.1)$$

where Z is a standard Gaussian random variable. We have in particular presented a way to discretise this continuous path by a discrete-time tree. Using the tools developed in Section 2.1, we are now able to either simulate many instances of S_t directly from (2.2.1) or draw paths of this process between zero and t by splitting this time interval as $[0,t] = [t_0, t_1] \cup \ldots \cup [t_{m-1}, t_m]$, for some $m \geq 2$. Between two dates t_{i-1} and t_i ($i = 1, \ldots, m$), we have indeed

$$S_{t_i} = S_{t_{i-1}} \exp\left(\left(r - \frac{\sigma^2}{2}\right)(t_i - t_{i-1}) + \sigma\sqrt{t_i - t_{i-1}} Z_i\right),$$

where all the Gaussian random variables $(Z_i)_{0 \leq i \leq m}$ are independent. Let us now consider the following problem: given such a simulation scheme, estimate the expected value $\theta := \mathbb{E}(S_T)$ of the random variable S_T at time $T > 0$. Consider a vector $\mathbf{S}_T := \left(S_T^i\right)_{1 \leq i \leq n}$ of independent and identically distributed random variables with the same distribution as S_T, and define

$$\widehat{\theta}\left(\mathbf{S}_T\right) := \frac{1}{n}\sum_{i=1}^{n} S_T^i.$$

Straightforward computations show that $\mathbb{E}\left(\widehat{\theta}\left(\mathbf{S}_T\right)\right) = \mathbb{E}(S_T)$ and $\mathbb{V}\left(\widehat{\theta}\left(\mathbf{S}_T\right)\right) = n^{-1}\mathbb{V}(S_T)$, so that the estimator $\widehat{\theta}\left(\mathbf{S}_T\right)$ is *unbiased*. Furthermore, by the central limit theorem, we know that

$$\frac{\widehat{\theta}\left(\mathbf{S}_T\right) - \theta}{\widehat{\sigma}_n/\sqrt{n}} \text{ converges in law to } \mathcal{N}(0,1). \tag{2.2.2}$$

Here the quantity $\widehat{\sigma}$ could be taken as $\mathbb{V}\left(S_T\right)$, but this quantity is in general unknown, so that we consider instead its unbiased estimator

$$\widehat{\sigma}_n := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(S_T^i - \widehat{\theta}\left(\mathbf{S}_T\right)\right)^2}.$$

Note that (2.2.2) tells us that our estimator error $\widehat{\theta}\left(\mathbf{S}_T\right) - \theta$ is distributed as a Gaussian random variable with mean zero and variance $\widehat{\sigma}_n^2/n$.

Digression

In mathematical finance, many (continuous-time) models are represented via the theory of stochastic differential equations (SDEs). In the Black-Scholes case, the logarithm of the stock price dynamics is the unique solution to the SDE

$$\mathrm{d}X_t = \left(r - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad X_0 = x \in \mathbb{R},$$

where $(W_t)_{t \geq 0}$ is a standard Brownian motion, as defined in Definition 1.1.8. In this example, the instantaneous volatility σ is constant. Empirical studies have shown that this is not a realistic feature, and stochastic volatility has to be taken into account. A widely used stochastic volatility model is the so-called Heston model [36], where the log stock price process is the unique solution

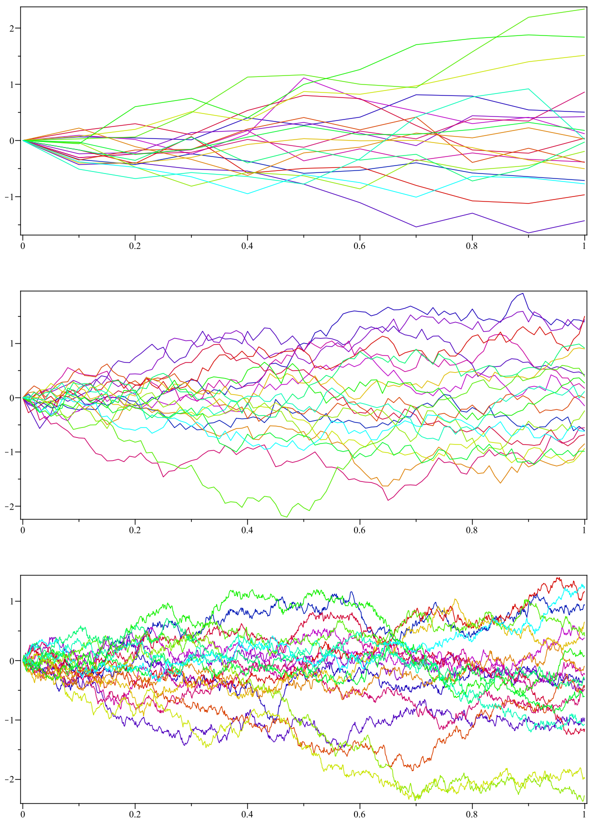

Figure 2.2: We plot here twenty Black-Scholes paths, where the time interval $[0, 1]$ has been split respectively into ten, one hundred and one thousand subintervals.

to

$$\begin{aligned} \mathrm{d}X_t &= \left(r - \frac{V_t}{2}\right)\mathrm{d}t + \sqrt{V_t}\mathrm{d}W_t, \qquad X_0 = x \in \mathbb{R}, \\ \mathrm{d}V_t &= \kappa\left(\theta - V_t\right)\mathrm{d}t + \xi\sqrt{V_t}\mathrm{d}B_t, \qquad V_0 = v_0 > 0, \end{aligned}$$

where $\kappa > 0$ is called the mean-reversion speed, $\theta > 0$, the long-term variance, $\xi > 0$ the volatility of volatility, and where the two Brownian motions $(W_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ are correlated with correlation parameter $\rho \in [-1, 1]$. Simulating such a model can be done as follows:

(i) on a small time interval $[t, t + \Delta]$ discretise the two SDEs as

$$\begin{aligned} X_{t+\Delta} - X_t &= \left(r - \frac{V_t}{2}\right)\Delta + \sqrt{\Delta V_t}Z_t^{(1)}, \\ V_{t+\Delta} - V_t &= \kappa\left(\theta - V_t\right)\Delta + \xi\sqrt{\Delta V_t}Z_t^{(2)}, \end{aligned}$$

where for any $t \geq 0$, $Z_t^{(1)}$ and $Z_t^{(2)}$ are Gaussian with zero mean and unit variance (but correlated with correlation ρ);

(ii) At some time $t \geq 0$, simulate the two-dimensional Gaussian random variable $\left(Z_t^{(1)}, Z_t^{(2)}\right)$. Since the values of X_t and V_t are known, we can deduce the values of $X_{t+\Delta}$ and $V_{t+\Delta}$;

(iii) iterate this procedure.

When following these steps, several problems may arise, in particular a large negative value of $Z_t^{(2)}$ can lead to a negative value of $V_{t+\delta_t}$, and hence taking its square-root is not allowed. This seemingly simple simulation question is in fact a difficult problem and finer results are needed. We shall not however study them here, but refer the interested reader to [30] for complete and thorough results and to [58] for the particular case of the Heston model.

2.2.2 Variance reduction methods

In Section 2.2.1 we have shown that the variance of the error arising from estimating an expectation–and hence option prices—was or order $1/n$, where n is the number of simulations. Variance reductions methods are a tool developed to decrease this variance while keeping the estimator unbiased. This implies that, for a given number of simulations, the estimator will be more robust. There are essentially three types of variance reduction methods and we shall only concentrate on the first two, leaving the third one for a more advanced course on Monte Carlo methods:

(i) Antithetic variables;

(ii) Control variables;

(iii) Importance sampling.

The method of *antithetic variables* is based on the following idea: if the random variable U is uniformly distributed on the closed interval $[0, 1]$, then so is $1 - U$. If one simulates a random path based on U, one might observe large—though rare—movements of U. Simulating the path both with U and with $1 - U$ clearly compensates such large positive occurrences. We shall say that the pair $(U, 1 - U)$ is *antithetic*. Let now f be a monotone function. The transformation method derived in Section 2.1.3 implies that the random variables $f^{-1}(U)$ and $f^{-1}(1-U)$ form an antithetic pair as well. In particular, in the standard Gaussian case, Z and $-Z$ form an antithetic pair. We leave this proof as a guided exercise:

Exercise 13. Let X be a random variable with finite mean $\mu \in \mathbb{R}$, finite variance σ^2 and let $n \in \mathbb{N}$. Define a family $\left(X_i, \widetilde{X}_i\right)_{i=1,\dots,n}$ of iid pairs, where all the X_i and $\widetilde{X}_i$ have the same distribution as X. Note that for fixed $i = 1, \dots, n$, the random variables X_i and $\widetilde{X}_i$ may not be—and in general will not be—independent. Consider the two estimators

$$\widehat{\theta}_n := \frac{1}{2n}\sum_{i=1}^{2n} X_i \qquad \text{and} \qquad \widetilde{\theta}_n := \sum_{i=1}^{n}\left(\frac{X_i + \widetilde{X}_i}{2n}\right).$$

- Using the central limit theorem, determine the estimation error in both the standard case and the antithetic case.

- Give a condition on the covariance $\mathrm{cov}\left(X_i, \widetilde{X}_i\right)$ so that the antithetic method reduces the variance of the estimator.

Solution. *The following quantity are immediate from the definitions of the estimators* $\widetilde{\theta}_n$ *and* $\widehat{\theta}_n$*:*

$$\mathbb{E}\left(\widehat{\theta}_n\right) = \mu, \qquad \mathbb{V}\left(\widehat{\theta}_n\right) = \frac{\sigma^2}{2n},$$
$$\mathbb{E}\left(\widetilde{\theta}_n\right) = \mu, \qquad \mathbb{V}\left(\widetilde{\theta}_n\right) = \frac{1}{n}\mathbb{V}\left(\frac{X_i + \widetilde{X}_i}{2}\right) =: \frac{\widetilde{\sigma}^2}{n}.$$

The central limit theorem therefore implies that

$$\frac{\widehat{\theta}_n - \mu}{\sigma/\sqrt{2n}} \text{ converges in distribution to } \mathcal{N}(0,1);$$
$$\frac{\widetilde{\theta}_n - \mu}{\widetilde{\sigma}/\sqrt{n}} \text{ converges in distribution to } \mathcal{N}(0,1).$$

Note that we can replace $\widetilde{\sigma}$ *by the standard deviation of the* n*-sample* $\left\{\frac{X_1+\widetilde{X}_1}{2}, \ldots, \frac{X_n+\widetilde{X}_n}{2}\right\}$*, and we denote it* $\widetilde{s}_n$*. For any* $\alpha \in (0,1)$*, this provides a* $(1-\alpha)$*-confidence interval of the form*

$$\left[\mu - \frac{\widetilde{s}x_{\alpha/2}}{\sqrt{n}}, \mu + \frac{\widetilde{s}x_{\alpha/2}}{\sqrt{n}}\right],$$

where $x_{\alpha/2}$ *is the unique solution to the equation* $\mathcal{N}\left(x_{\alpha/2}\right) = 1 - \alpha/2$*.*

We now wish to find a condition that ensures that the new estimator $\widetilde{\theta}_n$ *has a lower variance than the original one* $\widehat{\theta}_n$*, i.e.*

$$\frac{1}{n}\mathbb{V}\left(\frac{X_i + \widetilde{X}_i}{2}\right) = \mathbb{V}\left(\widetilde{\theta}_n\right) < \mathbb{V}\left(\widehat{\theta}_n\right) = \frac{1}{2n}\mathbb{V}\left(X_i\right),$$

which is tantamount to $\mathbb{V}\left(X_i + \widetilde{X}_i\right) < 2\mathbb{V}\left(X_i\right)$*. Since*

$$\mathbb{V}\left(X_i + \widetilde{X}_i\right) = \mathbb{V}\left(X_i\right) + \mathbb{V}\left(\widetilde{X}_i\right) + 2\mathrm{cov}\left(X_i, \widetilde{X}_i\right) = 2\mathbb{V}\left(X_i\right) + 2\mathrm{cov}\left(X_i, \widetilde{X}_i\right),$$

the variance of the new estimator will be reduced as soon as $\mathrm{cov}\left(X_i, \widetilde{X}_i\right) < 0$*.*

The method of *control variates* builds upon the antithetic variable methodology. In order to estimate $\theta := \mathbb{E}(g(X))$, where the function $g : \mathbb{R} \to \mathbb{R}$ represents the payoff function, the antithetic variables method suggests to use $\frac{1}{2}\left(\mathbb{E}(g(X)) + \mathbb{E}(g(Y))\right)$ instead where X and Y are negatively correlated. Consider the same structure as before, i.e. we observe (generate) a sample $(X_i)_{1\leq i\leq n}$ of iid random variables with the same distribution as X, where $\mathbb{E}(g(X)) < \infty$. Let us now assume that we can find a functional $f : \mathbb{R} \to \mathbb{R}$ sufficiently close (in some sense to be made precise) to g and such that the quantity $\mathbb{E}\left(f(Y)\right)$ can be computed easily, and assume further that we have another sample $(Y_i)_{1\leq i\leq n}$ of random variables such that the pairs $(g(X_i), f(Y_i))$ form an iid sequence. Consider now the new random variable

$$\widehat{\theta}_i^{\beta} := g(X_i) + \beta\Big(f(Y_i) - \mathbb{E}\left(f(Y_i)\right)\Big),$$

where β is some real number. Compute now the sample mean

$$\widehat{\theta}^{\beta}_{(n)} := \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}^{\beta}_{i}.$$

The control variate estimator $\widehat{\theta}^{\beta}_{(n)}$ is unbiased ($\mathbb{E}(\widehat{\theta}^{\beta}_{(n)}) = \mathbb{E}(g(X))$) and its variance reads

$$\mathbb{V}\left(\widehat{\theta}^{\beta}_{(n)}\right) = \frac{1}{n}\Big(\mathbb{V}\left(g(X)\right) + \beta^2\mathbb{V}\left(f(Y)\right) + 2\beta\operatorname{cov}\left(f(Y), g(X)\right)\Big).$$

Note in particular that

$$\mathbb{V}\left(\widehat{\theta}^{\beta}_{(n)}\right) < \mathbb{V}\left(\widehat{\theta}^{0}_{(n)}\right) \quad \text{if and only if} \quad \beta^2\mathbb{V}\left(f(Y)\right) < -2\beta\operatorname{cov}\left(f(Y), g(X)\right),$$

i.e. the control variate estimator is more robust than the ordinary sample mean $\mathbb{V}\left(\widehat{\theta}^{0}_{(n)}\right)$ as soon as the inequality on the right-hand side is satisfied. The function $\beta \mapsto \mathbb{V}\left(\widehat{\theta}^{\beta}_{(n)}\right)$ is a quadratic with strictly positive dominant coefficient and hence its unique minimum is attained at

$$\beta^* := \operatorname{argmin}\left(\mathbb{V}\left(\widehat{\theta}^{\beta}_{(n)}\right)\right) = -\frac{\operatorname{cov}\left(f(Y), g(X)\right)}{\mathbb{V}\left(f(Y)\right)};$$

therefore the estimator and its variance read

$$\begin{aligned}
\widehat{\theta}^{\beta^*}_{(n)} &= g(X) - \frac{\operatorname{cov}\left(f(Y), g(X)\right)}{\mathbb{V}\left(f(Y)\right)}\left(f(Y) - \mathbb{E}\left(f(Y)\right)\right), \\
\mathbb{V}\left(\widehat{\theta}^{\beta^*}_{(n)}\right) &= \mathbb{V}\left(g(X)\right) - \frac{\operatorname{cov}\left(f(Y), g(X)\right)^2}{\mathbb{V}\left(f(Y)\right)} = \mathbb{V}\left(g(X)\right)\left(1 - \rho\left(f(Y), g(X)\right)^2\right).
\end{aligned}$$

The last equality above clearly implies that the higher the correlation between the two payoffs $f(Y)$ and $g(X)$, the lower the variance of the estimator, and hence the more robust the estimator. One may however question this approach since the optimal coefficient β^* requires the knowledge of the covariance between $f(Y)$ and $g(X)$. In practice, one usually replace the optimal coefficient by its sample mean estimate

$$\widehat{\beta}_n := \frac{\sum_{i=1}^{n}\left(g(X_i) - n^{-1}\sum_{j=1}^{n} g(X_j)\right)\left(f(Y_i) - n^{-1}\sum_{j=1}^{n} f(Y_j)\right)}{\sum_{i=1}^{n}\left(g(X_i) - n^{-1}\sum_{j=1}^{n} g(X_j)\right)^2}.$$

The strong law of large numbers clearly implies that $\widehat{\beta}_n$ converges to β^* as n tends to infinity. Note that this estimate is nothing else but the slope of the least-square regression of the scatter plot $(g(X_i), f(Y_i))$.

We now build upon the previous method to study *Importance sampling.* Consider as before a random variable X, assumed to have a density f (we assume for simplicity that the support of the distribution is the whole real line). We wish to estimate $\theta := \mathbb{E}^f\left[h(X)\right]$, for some function h, which we assume to be positive. The standard Monte Carlo estimate reads $\widehat{\theta}_n := n^{-1}\sum_{i=1}^{n} h(X_i)$, where the family $(X_i)_{1\leq i\leq n}$ is independent and identically distributed as X. Let now g be a strictly positive function. We have

$$\theta = \mathbb{E}^f\left[h(X)\right] = \int h(x)f(x)\mathrm{d}x = \int h(x)\frac{f(x)}{g(x)}g(x)\mathrm{d}x =: \mathbb{E}^g\left[h(X)\frac{f(X)}{g(X)}\right].$$

The two functions f and g are said to be *mutually equivalent*. We then define the g-estimator $\widehat{\theta}_n^g$ by

$$\widehat{\theta}_n^g := \frac{1}{n}\sum_{i=1}^n h(X_i)\frac{f(X_i)}{g(X_i)}.$$

It is clear that this estimator is unbiased. Let us assume for simplicity that $\theta = 0$ (if it is not, then one can subtract θ^2 everywhere in the computations below). We then have

$$\mathbb{V}^g\left[\widehat{\theta}_n^g\right] = \mathbb{E}^g\left[\left(\frac{1}{n}\sum_{i=1}^n h(X_i)\frac{f(X_i)}{g(X_i)}\right)^2\right] = \frac{1}{n}\mathbb{E}^g\left[\left(h(X)\frac{f(X)}{g(X)}\right)^2\right].$$

We therefore obtain

$$\mathbb{V}^f\left[\widehat{\theta}_n\right] - \mathbb{V}^g\left[\widehat{\theta}_n^g\right] = \int h^2(x)\left(1-\frac{f(x)}{g(x)}\right)f(x)\mathrm{d}x.$$

Since we wish to have a lower variance, we need to choose g such that

$$\begin{cases} g(x) > f(x), & \text{when } h^2(x)f(x) \text{ is large,}\\ g(x) < f(x), & \text{when } h^2(x)f(x) \text{ is small.}\end{cases}$$

Example. Consider the integral $I := \int_0^1 \left(1-x^2\right)^{1/2}\mathrm{d}x$. A simple change of variables $x \mapsto \cos(z)$ implies that $I = \pi/4$. We wish to determine a method to approximate accurately π. Note that we can rewrite the integral as $I = \mathbb{E}\left[h(U)\right]$, where U is uniformly distributed on $[0,1]$ and $h : x \mapsto \sqrt{1-x^2}$. The standard Monte Carlo estimator is $\widehat{\theta} := n^{-1}\sum_{i=1}^n h(U_i)$, where the family $(U_i)_{1\le i\le n}$ is iid with common distribution U. Consider now the approximation of the ideal density

$$\widetilde{g}(x) := \frac{h(x)f(x)}{\widehat{\theta}} = \frac{nh(x)}{\sum_{i=1}^n h(X_i)},$$

since $f \equiv 1$. However, the function $\widetilde{g}$ does not integrate to 1, so it is not a proper density function. On each interval $J_k := \left[\frac{k-1}{n}, \frac{k}{n}\right]$, for $k = 1,\ldots,n$, define the midpoint $m_k := (2k-1)/n$, and

$$\xi_k := \frac{h(m_k)}{\sum_{j=1}^n h(m_j)}.$$

it is clear that all the ξ_k are non negative and sum up to one. Hence they correspond to the probabilities of sampling from the interval J_k. Define now the function $g(x) := n\xi_k \mathbb{1}_{\{x\in J_k\}}$ for any $x \in [0,1]$. Compute now the variance of the g-estimator and study (numerically) the convergence differences between this estimator and the standard Monte Carlo estimator.

2.2.3 Option pricing

Recall now that by Theorem 0.2.4: under absence of arbitrage, the value of an option is equal to the discounted expectation of its terminal payoff, where the expectation is taken under the risk-neutral probability. Consider first the case of a European call option in the Black-Scholes model. In Section 2.2.1, we have seen how to generate random paths in the Black-Scholes model and have provided a way to estimate the expectation of a random variable. We hence leave it as an exercise:

Exercise 14. Let $(S_t)_{t\geq 0}$ follow the Black-Scholes model with $r = 5\%$, $\sigma = 20\%$ and $S_0 = 100$. Consider further a European call option with maturity $T = 1$ year and strike $K = 100$, and denote $C_0(K)$ its value (to be determined) at inception. Determine a Monte-Carlo procedure to estimate the value $C_0(K)$ of this call option. Use MATLAB to output the following graph:

- with 10 time subintervals, show the convergence of the estimator with respect to the number of simulated paths;
- with 1000 paths, show the convergence of the estimator with respect to the number of time subintervals;
- with 1000 paths and 100 time subintervals, plot the function $K \mapsto C_0(K)$ for $K = 20, \ldots, 150$.

2.2.4 Application: European down and out barrier option under Black-Scholes

We consider now a more elaborate example, which may reveal some flaws of basic Monte Carlo schemes. We wish to determine the value of a down-and-out European barrier put option and the corresponding call option. The option (with strike K, maturity T and barrier B) is written on the underlying stock price $(S_t)_{t\geq 0}$, which follows the Black-Scholes model (Definition 1.1.11). The payoff of the put and the call respectively read

$$(K - S_T)_+ \mathbb{1}_{\{\inf\{S_t, t\in[0,T]\}>B\}} \qquad \text{and} \qquad (S_T - K)_+ \mathbb{1}_{\{\inf\{S_t, t\in[0,T]\}>B\}}.$$

The simple (without variance reduction methods) Monte Carlo scheme is straightforward to implement. Indeed, we simply need to draw paths according to the simulation scheme presented above. Out of all these paths, we only keep those that have not dropped below the barrier B at some time between the inception and the maturity. It is clear that the level of the barrier must be set above the initial stock price S_0, otherwise the option is always knocked-out. The plot 2.3 below numerically illustrates the convergence of the Monte Carlo method.

2.2.5 Application: Bond pricing with the CIR model

In this example, we would like to highlight some technical difficulties that may arise in the simulation of locally degenerated processes. Let $(X_t)_{t\geq 0}$ be a stochastic process satisfying the following stochastic differential equation:

$$\mathrm{d}X_t = \kappa\left(\theta - X_t\right)\mathrm{d}t + \sigma\sqrt{X_t}\mathrm{d}W_t, \quad \text{with } X_0 > 0, \tag{2.2.3}$$

where κ, θ and σ are strictly positive constants, and W is a standard Brownian motion. This process is known in the probability literature as the Feller process. In finance, it has been (and

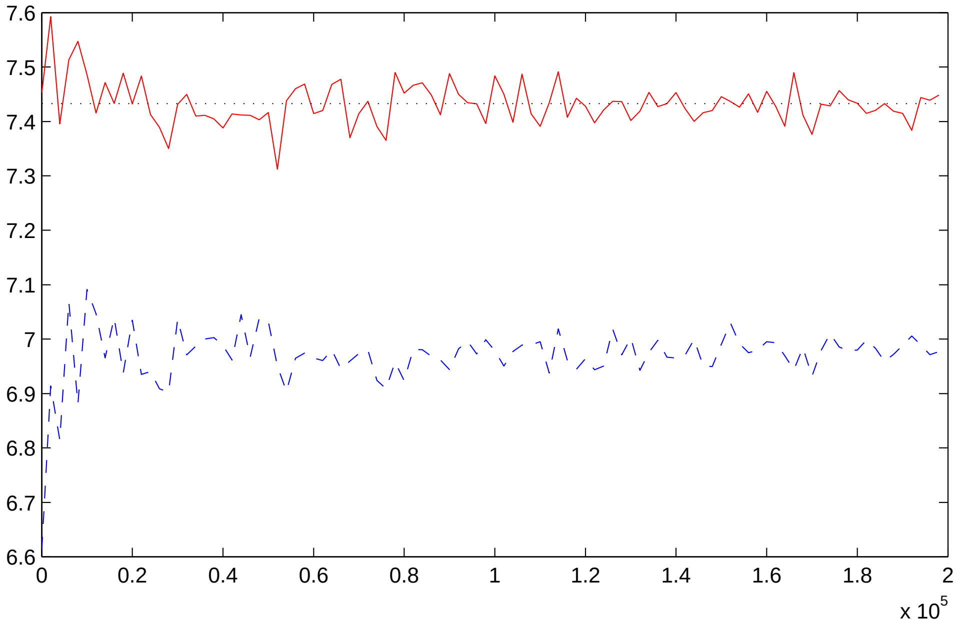

Figure 2.3: Convergence of the European down-and-out put option in the Black-Scholes model under Monte Carlo simulations with 20 time steps. The dotted line corresponds to the standard European Put option price. The barrier is equal to 10 for the solid line and to 60 for the dashed line. The other parameters are: $S_0 = 90$, $K = 100$, $T = 1$ year, $r = 10\%$ and $\sigma = 20\%$. The horizontal axis corresponds to the number of simulations.

still is) used as a standalone interest rate process called the Cox-Ingersoll-Ross process [14]. It is also the basis of the Heston stochastic volatility model [36], where X represents the instantaneous volatility of a stock price process $(S_t)_{t\geq 0}$ satisfying the SDE $\mathrm{d}S_t/S_t = r\mathrm{d}t + \sqrt{X_t}\mathrm{d}Z_t$ with $S_0 > 0$ and Z is another Brownian motion, which may be correlated with W. One can show that the conditional expectation and variance of X_t read

$$\mathbb{E}(X_t|X_0) = \theta + (X_0 - \theta)\mathrm{e}^{-\kappa t} \qquad \text{and} \qquad \mathbb{V}(X_t|X_0) = \frac{\sigma^2}{\kappa} r_0 \mathrm{e}^{-\kappa t}\left(1 - \mathrm{e}^{-\kappa t}\right) + \frac{\sigma^2\theta}{2\kappa}\left(1 - \mathrm{e}^{-\kappa t}\right)^2,$$

so that θ represents its long term mean and κ a mean-reversion strength. Let us start with a brutal discretisation of the SDE (2.2.3). Between two times t and $t + \Delta_t$, we have

$$X_{t+\Delta_t} - X_t = \kappa(\theta - X_t)\Delta_t + \sigma\sqrt{\Delta_t}\sqrt{X_t}\widetilde{n}, \tag{2.2.4}$$

where $\widetilde{n} \sim \mathcal{N}(0, 1)$. For a given $X_t > 0$, since the Gaussian component $\widetilde{n}$ can take arbitrary large values, it can clearly happen that $X_{t+\Delta_t}$ becomes strictly negative. In that case, how does one proceed to the next step, i.e. how does one handle the term $\sqrt{X_{t+\Delta_t}}$? Before actually tackling this issue, one should ask whether the SDE (2.2.3) really makes sense and what happens to the process when it hits the boundary zero. The following proposition answers this question:

Proposition 2.2.1. *When $2\kappa\theta \geq \sigma^2$ the process $(X_t)_{t\geq 0}$ never reaches zero almost surely. If this condition is violated, then the origin is accessible and strongly reflecting.*

The proof of this proposition is outside the scope of these lecture notes, but the interested reader can consult [40, Chapter 15, Section 6] for precise details. The notions of *accessible* and *strongly reflecting* have a precise mathematical meaning, but they essentially imply that when $2\kappa\theta < \sigma^2$, the process can touch the origin, in which case it immediately bounces back to the strictly positive halfspace $\mathbb{R}_+^*$. In particular, the proposition says that the process is well defined and cannot 'become negative' at any point in time. The condition $2\kappa\theta \geq \sigma^2$ is commonly referred to as the *Feller condition.*

Let us now return to the simulation of the Feller diffusion given by (2.2.3). For any $t > 0$, the law of the random variable X_t given X_0 is known exactly as a non-central chi-square distribution:

Proposition 2.2.2. *Let $\chi_{\nu,\lambda}$ be a non-central chi-square random variable with $\nu > 0$ degrees of freedom and non-centrality parameter $\lambda > 0$, i.e. with cumulative distribution function:*

$$\mathbb{P}(\chi_{\nu,\lambda} \leq x) = \mathrm{e}^{-\lambda/2}\sum_{n\geq 0}\frac{(\lambda/2)^n}{n!2^{n+\nu/2}\Gamma(n+\nu/2)}\int_0^x z^{n-1+\nu/2}\mathrm{e}^{-z/2}\mathrm{d}z.$$

Define the two following quantities:

$$d := \frac{4\kappa\theta}{\sigma^2} \qquad \text{and} \qquad \zeta_t := \frac{4\kappa\mathrm{e}^{-\kappa t}}{\sigma^2\left(1 - \mathrm{e}^{-\kappa t}\right)}.$$

Then conditional on X_0, the random variable X_t is distributed as $\frac{\chi_{\nu,\lambda}}{\zeta_t}\mathrm{e}^{-\kappa t}$.

One could therefore sample the law of X_t from the chi-square distribution. However, this is quite time-consuming, and we shall not dive into this part here. Going back to the Euler scheme in (2.2.4), one solution suggested in the literature [43] is to replace the term $\sqrt{X_t}$ by $\sqrt{|X_t|}$. Another suggestion would be to replace it by $\sqrt{\max(X_t, 0)}$. The latter scheme means that if for some $t \geq 0$, X_t becomes negative, then the remaining process between t and $t + \Delta_t$ becomes deterministic with an upward drift equal to $\kappa\theta$. Many other refinements have been proposed and we refer the interested reader to [2]. By no-arbitrage arguments, the price at time zero of a zero-coupon bond B_t with maturity $t \geq 0$ reads

$$B_t = \mathbb{E}\left[\exp\left(-\int_0^t r_s \mathrm{d}s\right)\right].$$

One can show (see [14]) that the price has the exact closed-form solution $B_t = \exp\left(m_t + n_t r_0\right)$, where $\gamma := \frac{1}{2}\sqrt{\kappa^2 + 2\sigma^2}$ and

$$m_t := \frac{2\kappa\theta}{\sigma^2} \log\left(\frac{\gamma \mathrm{e}^{\kappa t/2}}{\gamma\cosh(\gamma t) + \kappa\sinh(\gamma t)/2}\right) \quad \text{and} \quad n_t := \frac{\sinh(\gamma t)}{\gamma\cosh(\gamma t) + \kappa\sinh(\gamma t)/2}.$$

Exercise 15. Implement an Euler scheme for the Feller diffusion (2.2.3) and plot the convergence of the price of the bond as the number of paths / time steps become large. One can take the following values for the parameters: $\kappa = 2.1$, $\theta = 0.09$, $v_0 = 0.07$, $\sigma = 0.1$ and $t = 2$.

Chapter 3

Finite difference methods for PDEs

3.1 Reminder on PDEs and the Black-Scholes heat equation

3.1.1 Review of PDEs and their classification

A partial differential equation (PDE) is a functional equation that contains both a function and some of its derivatives. As opposed to an ordinary differential equation (ODE) in which the function to determine depends on one variable, the unknown function in a PDE depends on several variables. In mathematical finance, these variables are usually the time t and a state variable x that lies in some subset of $\mathbb{R}^n$ ($n \geq 1$). For a given function $f : \mathbb{R} \to \mathbb{R}$, we shall use interchangeably the notations $\dfrac{\partial f}{\partial x}$ and $\partial_x f$ to denote the derivative with respect to the (one-dimensional) variable x. Let now Ω be a subset of $\mathbb{R}^n$, $\mathrm{u} = (u_1, \ldots, u_m)$ a multidimensional function from Ω to $\mathbb{R}^m$. For $\alpha = (\alpha_1, \ldots, \alpha_n) \in (\mathbb{N} \cup \{0\})^n$, with $|\alpha| = \alpha_1 + \ldots + \alpha_n$, we denote by $\mathrm{D}^\alpha\mathrm{u}$ the partial derivative

$$\mathrm{D}^\alpha \mathrm{u} := \frac{\partial^{|\alpha|}\mathrm{u}}{\partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}} = \left(\frac{\partial^{|\alpha|} u_1}{\partial x_1^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}}, \ldots, \frac{\partial^{|\alpha|} u_m}{\partial x_1^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}} \right),$$

and for any $k \geq 1$, $\mathrm{D}^k\mathrm{u} := \{\mathrm{D}^\alpha\mathrm{u} : |\alpha| = k\}$ the set of all partial derivatives of order k. For example, when $|\alpha| = 1$, Du represents the gradient matrix

$$\mathrm{Du} = \begin{pmatrix} \partial_{x_1} u_1 & \ldots & \partial_{x_n} u_1 \\ \vdots & \ddots & \vdots \\ \partial_{x_1} u_m & \ldots & \partial_{x_n} u_m \end{pmatrix}.$$

For a given integer $k \in \mathbb{N}$, $\mathrm{D}^k\mathrm{u}$ represents the tensor of all partial derivatives of order k, namely the collection of all partial derivatives $\mathrm{D}^\alpha\mathrm{u}$ such that $|\alpha| = k$. We now let F be a function from

$\mathbb{R}^{n^k m} \times \mathbb{R}^{n^{k-1} m} \times \ldots \times \mathbb{R}^m \times \Omega \to \mathbb{R}^q$, and consider the following equation:

$$F\left(\mathrm{D}^k \mathrm{u}, \mathrm{D}^{k-1}\mathrm{u}, \ldots, \mathrm{Du}, \mathrm{u}, \mathrm{x}\right) = 0. \tag{3.1.1}$$

where the higher-order term in $\mathrm{D}^k\mathrm{u}$ is not null.

Definition 3.1.1. The equation (3.1.1) is called a partial differential equation in u of order k.

For a given open subset $\Omega \in \mathbb{R}^n$, let $\mathrm{u} : \Omega \to \mathbb{R}^m$ be a k-times differentiable function. It is a solution of (3.1.1) if it satisfies $F\left(\mathrm{D}^k\mathrm{u}(\mathrm{x}), \mathrm{D}^{k-1}\mathrm{u}(\mathrm{x}), \ldots, \mathrm{Du}(\mathrm{x}), \mathrm{u}(\mathrm{x}), \mathrm{x}\right) = 0$, for all $\mathrm{x} \in \Omega$.

Remark 3.1.2. There is in general no guarantee that a solution to a given PDE of the form (3.1.1) will exist. The PDE $(\partial_x u)^2 + 1 = 0$, with $m = n = 1$ has no real solution, for instance.

Definition 3.1.3. The PDE in (3.1.1) is called

(i) linear if it can be written as $\sum_{i=0}^k \alpha_i(\mathrm{x})\mathrm{D}^i\mathrm{u}(\mathrm{x}) = f(\mathrm{x})$ for some functions a_i $(i \leq k)$ and some function f. It is further called homogeneous if $f \equiv 0$;

(ii) semilinear if it can be written as $\alpha_k(\mathrm{x})\mathrm{D}^k\mathrm{u}(\mathrm{x}) + \alpha_0\left(\mathrm{D}^{k-1}\mathrm{u}(\mathrm{x}), \ldots, \mathrm{Du}(\mathrm{x}), \mathrm{u}(\mathrm{x}), \mathrm{x}\right) = 0$;

(iii) quasilinear if it has the form

$$\alpha_k\left(\mathrm{D}^{k-1}\mathrm{u}(\mathrm{x}), \ldots, \mathrm{Du}(\mathrm{x}), u(\mathrm{x}), \mathrm{x}\right)\mathrm{D}^k\mathrm{u}(\mathrm{x}) + \alpha_0\left(\mathrm{D}^{k-1}\mathrm{u}(\mathrm{x}), \ldots, \mathrm{Du}(\mathrm{x}), u(\mathrm{x}), \mathrm{x}\right) = 0;$$

(iv) fully non-linear if it is not quasi-linear.

Example. Check the following examples (with $m = 1$; Δ denotes here the Laplace operator):

- $F \equiv \partial_x u + x\partial_y u$ is linear of order one;
- $F \equiv \partial_x u + u\partial_y u$ is not linear of order one;
- $F \equiv \partial_t u - \partial_{xx} + 1$ is linear inhomogeneous of order two;
- $F \equiv \partial_t u - \partial_{xxt} u + u\partial_x u + 1$ is nonlinear inhomogeneous of order three;
- the heat equation on $\mathbb{R}^n$: $F \equiv \partial_t u - \Delta u$;
- the eikonal equation on $\mathbb{R}^n$: $F \equiv |\mathrm{D}u| - 1$;
- the wave equation on $\mathbb{R}^n$: $F \equiv \partial_{tt} u - \Delta u$.

Among all PDEs, we shall be interested in inhomogeneous linear second-order PDEs in the case $m = 1$, namely equations of the form $\mathcal{L}f = 0$, where the operator $\mathcal{L}$ has the following form:

$$\mathcal{L} := a_{11}\partial_{xx} + 2a_{12}\partial_{xy} + a_{22}\partial_{yy} + a_1\partial_x + a_2\partial_y + a_0. \tag{3.1.2}$$

The following proposition serves as a definition of the type of a PDE.

Proposition 3.1.4. *The operator $\mathcal{L}$ in* (3.1.2) *can be reduced to one of the following three forms:*

- *Elliptic form: if $a_{12}^2 < a_{11}a_{22}$, then $\mathcal{L} = \partial_{xx} + \partial_{yy} + \mathcal{L}_1$;*

- *Hyperbolic form: if $a_{12}^2 > a_{11}a_{22}$, then $\mathcal{L} = \partial_{xx} - \partial_{yy} + \mathcal{L}_1$;*

- *Parabolic form: if $a_{12}^2 = a_{11}a_{22}$, then $\mathcal{L} = \partial_{xx} + \mathcal{L}_1$,*

where $\mathcal{L}_1$ is an operator of order at most one.

Proof. Without loss of generality, we consider $a_0 = a_1 = a_2 = 0$ (otherwise add a quadratic term of the form $\alpha_x x^2 + \alpha_y y^2 + \beta_x x + \beta_y$, where α_x, α_y, β_x, β_y are constant). Assuming that $a_{11} \neq 0$, and denoting $\widetilde{a}_{12} := a_{12}/a_{11}$ and likewise for the other parameters, we can write

$$\mathcal{L} = \left(\partial_x + \widetilde{a}_{12}\partial_y\right)^2 + \left(\widetilde{a}_{22} - \widetilde{a}_{12}^2\right)\partial_{yy}.$$

In the elliptic case $a_{12}^2 < a_{11}a_{22}$ (equivalently $\widetilde{a}_{12}^2 < \widetilde{a}_{22}$), the quantity $\beta := \left(\widetilde{a}_{22} - \widetilde{a}_{12}^2\right)^{1/2}$ is well defined and non zero. With the new variables $\gamma := x + \widetilde{a}_{12}y$ and $\xi := \beta y$, the operator $\mathcal{L}$ reads

$$\mathcal{L} = \partial_{\gamma\gamma} + \partial_{\xi\xi} \qquad \left(\text{Laplace operator on } \mathbb{R}^2\right).$$

The other cases (hyperbolic and parabolic) are treated similarly. □

Remark 3.1.5. In the above proposition / definition, we have assumed the coefficients of the operator $\mathcal{L}$ in (3.1.2) to be constant. We could make them functions of (x, y), and the definitions would remain the same, i.e. the operator $\mathcal{L}$ is elliptic at the point (x, y) if $a_{12}(x, y)^2 < a_{11}(x, y)a_{22}(x, y)$ and is elliptic everywhere if the inequality holds for all (x, y).

Example.

- Laplace equation on $\mathbb{R}^n$, $\Delta f = 0$, is a linear elliptic PDE: $a_{11} = a_{22} = 1$;

- Heat equation on $\mathbb{R}^n$, $\partial_t f - \Delta f = 0$, is a linear parabolic PDE: $a_{11} = a_{22} = -1$, $a_1 = 1$;

- Wave equation on $\mathbb{R}$, $\partial_{tt} f - \partial_{xx} f = 0$, is a linear hyperbolic PDE: $a_{11} = -a_{22} = 1$;

- Eikonal equation on $\mathbb{R}^n$, $|\nabla u| = 1$, is a non-linear first-order PDE.

Partial differential equations are normally defined together with boundary conditions. The classical types of boundary conditions are the Dirichlet boundary conditions, i.e. u(x) is specified when x lies at the boundary $\partial\Omega$ of the domain, the Neumann condition when the derivative of u is set on $\partial\Omega$, and mixed boundary conditions, which are a combination of Dirichlet and Neumann conditions.

3.1.2 The Black-Scholes heat equation

The aim of this section is to present the canonical financial model, namely the Black-Scholes-Merton model, and to use it as a backbone to introduce the different numerical methods we shall see in this course. This model was introduced in 1973 by Fischer Black and Myron Scholes [7] and by Robert Merton [50] to represent the dynamics of an asset price. Merton and Scholes were later (1997) awarded the Nobel Prize in Economics for this result[1].

Derivation of the Black-Scholes PDE

This paragraph is intended to provide a rigorous derivation of the so-called Black-Scholes partial differential equation, but may be omitted in a first reading. Let $(W_t)_{t\geq 0}$ be a Brownian motion and $S := (S_t)_{t\geq 0}$ the asset price process. This model assumes the following dynamics under the so-called historical (observed) probability measure $\mathbb{Q}$:

$$\frac{\mathrm{d}S_t}{S_t} = \mu \mathrm{d}t + \sigma \mathrm{d}W_t, \qquad \text{with } S_0 > 0, \tag{3.1.3}$$

where $\mu \in \mathbb{R}$ is called the drift and $\sigma > 0$ is the instantaneous volatility. The question we are interested in here is the following: assuming (3.1.3), what is the value today (at time $t = 0$) of a European option with payoff $f(S_T)$ at maturity $T > 0$? For clarity, we shall denote V_t the value at time $t \in [0, T]$ of such a financial derivative. The first step is to obtain a probabilistic representation for the option price. As discussed before, under absence of arbitrage there exists a probability $\mathbb{P}$ under which we can write

$$V_0 = B(0, T)\, \mathbb{E}_{\mathbb{P}}[f(S)],$$

where $B(0, T)$ represents the value at time zero of a risk-free zero-coupon bond (discounting factor) paying 1 at time T. An application of Itô's lemma yields that the option price satisfies the stochastic differential equation

$$\mathrm{d}V_t = \left(\mu S_t \partial_S V_t + \partial_t V_t + \frac{\sigma^2}{2} S_t^2 \partial_{SS}^2 V_t\right) \mathrm{d}t + \sigma S_t \partial_S V_t \mathrm{d}W_t, \tag{3.1.4}$$

at any time t between inception and maturity with appropriate boundary conditions. Consider now a portfolio Π consisting at time t of a long position in the option V and a long position in Δ_t shares S, i.e. $\Pi_t = V_t + \Delta_t S_t$. On a small time interval $[t, t + \mathrm{d}t]$, the profit and loss of such a portfolio is $\mathrm{d}\Pi_t = \mathrm{d}V_t + \Delta_t \mathrm{d}S_t$. Using (3.1.4), we obtain

$$\mathrm{d}\Pi_t = \left\{\mu\left(\Delta_t + \partial_S V_t\right) S_t + \partial_t V_t + \frac{\sigma^2}{2} S_t^2 \partial_{SS}^2 V_t\right\} \mathrm{d}t + \left(\Delta + \partial_S V_t\right) \sigma S_t \mathrm{d}W_t.$$

This expression makes it clear that the only way to eliminate the risk—solely present in the form of the Brownian perturbations—is to set $\Delta_t = -\partial_S V_t$. This is called delta hedging. Now, since we

[1] Black (1938-1995) did not get the Nobel prize as the latter is not awarded posthumously

assume absence of arbitrage, the returns of the portfolio Π_t over the period $[t, t+\mathrm{d}t]$ is necessarily equal to the risk-free rate $r \geq 0$. Otherwise (assume the returns are higher than the risk-free one), it is possible to construct an arbitrage, for instance by borrowing money at time t to buy the portfolio, invest it at rate r and sell it at time $t+\mathrm{d}t$. This implies $\mathrm{d}\Pi_t = r\Pi_t \mathrm{d}t$ and hence

$$\partial_t V_t + rS\partial_S V_t + \frac{\sigma^2}{2} S^2 \partial_{SS}^2 V_t = rV_t.$$

This equation is called the Black-Scholes differential equation, associated with the boundary condition given by the payoff $V_T = f(S_T)$. We have already proved, in Theorem 1.1.20 in Chapter 1 that the European call price converges to the (unique) solution of this partial differential equation in a binomial tree model when the time increment tends to zero.

Reduction of the Black-Scholes PDE to the heat equation

Before trying to solve a partial differential equation, it may sound sensible to simplify it. Recall the Black-Scholes parabolic PDE:

$$\partial_t V_t + rS\partial_S V_t + \frac{\sigma^2}{2} S^2 \partial_{SS}^2 V_t = rV_t, \tag{3.1.5}$$

with boundary condition $V_T(S)$ (for instance for a European call option with maturity $T > 0$ and strike $K > 0$, we have $V_T(S) = (S_T - K)_+ := \max(S_T - K, 0)$). Define $\tau := T - t$ and the function $g_\tau(S) := V_t(S)$, then $\partial_t V_t(S) = -\partial_\tau g_\tau(S)$ and hence

$$-\partial_\tau g_\tau + rS\partial_S g_\tau + \frac{\sigma^2}{2} S^2 \partial_{SS}^2 g_\tau = rg_\tau,$$

with boundary condition $g_0(S)$. Define now the function f by $f_\tau(S) := \mathrm{e}^{r\tau} g_\tau(S)$, and we obtain

$$-\partial_\tau f_\tau + rS\partial_S f_\tau + \frac{\sigma^2}{2} S^2 \partial_{SS}^2 f_\tau = 0,$$

with boundary condition $f_0(S)$. Consider a further transformation $x := \log(S)$ and the function $\psi_\tau(x) := f_\tau(S)$. Since $S\partial_S f_\tau(S) = \partial_x \psi_\tau(x)$ and $S^2 \partial_{SS}^2 f_\tau(S) = \partial_{xx}^2 \psi_\tau(x) - \partial_x \psi_\tau(x)$, we obtain

$$-\partial_\tau \psi_\tau + \left(r - \frac{\sigma^2}{2}\right) \partial_x \psi_\tau + \frac{\sigma^2}{2} \partial_{xx}^2 \psi_\tau = 0, \tag{3.1.6}$$

with boundary condition $\psi_0(x)$. Finally, define the function ϕ_τ via $\psi_\tau(x) =: \mathrm{e}^{\alpha x + \beta\tau} \phi_\tau(x)$, so that

$$\begin{aligned}
\partial_x \psi_\tau(x) &= \left(\alpha\phi_\tau(x) + \partial_x \phi_\tau(x)\right) \mathrm{e}^{\alpha x + \beta\tau}, \\
\partial_{xx}^2 \psi_\tau(x) &= \left(\alpha^2 \phi_\tau(x) + 2\alpha\partial_x \phi_\tau(x) + \partial_{xx}^2 \phi_\tau(x)\right) \mathrm{e}^{\alpha x + \beta\tau}, \\
\partial_\tau \psi_\tau(x) &= \left(\beta\phi_\tau(x) + \partial_\tau \phi_\tau(x)\right) \mathrm{e}^{\alpha x + \beta\tau}.
\end{aligned}$$

With the parameters

$$\alpha := -\frac{1}{\sigma^2}\left(r - \frac{\sigma^2}{2}\right) \qquad \text{and} \qquad \beta := -\frac{1}{2\sigma^2}\left(r - \frac{\sigma^2}{2}\right)^2,$$

Equation (3.1.6) becomes the so-called *heat equation*

$$\partial_\tau \phi_\tau(x) = \frac{\sigma^2}{2}\partial_{xx}^2 \phi_\tau(x), \tag{3.1.7}$$

for all real number x with (Dirichlet) boundary condition $\phi_0(x) = \mathrm{e}^{-\alpha x}\psi_0(x)$.

Direct solution of the heat equation

In the following sections, we shall use the heat equation as the fundamental example when deriving finite-difference algorithms. In this very particular case though, one can determine an exact solution using Fourier transform methods. Let us rewrite the problem: we wish to solve the parabolic PDE $\partial_\tau \phi_\tau(x) = \frac{1}{2}\sigma^2\partial_{xx}^2\phi_\tau(x)$, for $x \in \mathbb{R}$ with boundary condition $\phi_0(x) = f(x)$ for some function f. Define the Fourier transform $\widehat{\phi}_\tau$ of the function ϕ_τ by

$$\widehat{\phi}_\tau(z) := \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}zx}\phi_\tau(x)\mathrm{d}x, \qquad \text{for any } z \in \mathbb{R}.$$

A double integration by parts shows that

$$\begin{aligned}
\widehat{\partial_{xx}\phi_\tau}(z) &= \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}zx}\partial_{xx}\phi_\tau(x)\mathrm{d}x \\
&= \frac{1}{2\pi}\left[\mathrm{e}^{\mathtt{i}zx}\partial_x\phi_\tau(x)\right]_{\mathbb{R}} - \frac{\mathtt{i}z}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}zx}\partial_x\phi_\tau(x)\mathrm{d}x \\
&= \frac{1}{2\pi}\left[\mathrm{e}^{\mathtt{i}zx}\partial_x\phi_\tau(x)\right]_{\mathbb{R}} - \frac{\mathtt{i}z}{2\pi}\left[\mathrm{e}^{\mathtt{i}zx}\phi_\tau(x)\right]_{\mathbb{R}} - \frac{z^2}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}zx}\phi_\tau(x)\mathrm{d}x \\
&= -z^2\widehat{\phi_\tau}(z),
\end{aligned}$$

where we have made the standing assumption that the functions ϕ_τ and $\partial_x\phi_\tau$ converge to zero at infinity. We also have $\widehat{\partial_\tau\phi}(z) = \partial_\tau\widehat{\phi}_\tau(z)$ by Fubini's theorem (see Theorem A.3.1 in Appendix A.3). The heat equation therefore becomes $\partial_\tau\widehat{\phi}_\tau(z) + \frac{1}{2}\sigma^2 z^2\widehat{\phi}_\tau(z) = 0$ in the Fourier variable z, with boundary condition $\widehat{\phi}_0(z) = \widehat{f}(z)$. Standard results from ODE theory imply that

$$\widehat{\phi}_\tau(z) = \widehat{f}(z)\exp\left(-\frac{\sigma^2 z^2\tau}{2}\right),$$

and hence, inverting the Fourier transform leads to

$$\begin{aligned}
\phi_\tau(x) &= \int_{\mathbb{R}} \mathrm{e}^{-\mathtt{i}xz}\widehat{\phi}_\tau(z)\mathrm{d}z = \int_{\mathbb{R}} \mathrm{e}^{-\mathtt{i}xz}\widehat{f}(z)\mathrm{e}^{-\frac{1}{2}\sigma^2 z^2\tau}\mathrm{d}z \\
&= \int_{\mathbb{R}} \mathrm{e}^{-\mathtt{i}xz}\left(\frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}z\xi}f(\xi)\mathrm{d}\xi\right)\mathrm{e}^{-\frac{1}{2}\sigma^2 z^2\tau}\mathrm{d}z \\
&= \frac{1}{2\pi}\int_{\mathbb{R}} f(\xi)\left(\int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}z(\xi-x)}\mathrm{e}^{-\frac{1}{2}\sigma^2 z^2\tau}\mathrm{d}z\right)\mathrm{d}\xi \\
&= \frac{1}{\sigma\sqrt{2\pi\tau}}\int_{\mathbb{R}} f(\xi)\exp\left(-\frac{(x-\xi)^2}{2\sigma^2\tau}\right)\mathrm{d}\xi,
\end{aligned}$$

the third line follows by Fubini's theorem, and the last line relies on the following equality:

$$\left.\widehat{\mathrm{e}^{-\alpha z^2}}\right|_\omega := \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\alpha z^2 + \mathtt{i}\omega z}\mathrm{d}z = \frac{1}{2\sqrt{\pi\alpha}}\exp\left(-\frac{\omega^2}{4\alpha}\right),$$

with $\omega := x - \xi$ and $\alpha := \sigma^2\tau/2$. The left-hand side represents the Fourier transform of $\mathrm{e}^{-\alpha z^2}$ evaluated at ω.

3.2 Digression: why are we interested in PDEs?

Option pricing problems can be solved in many ways, in particular by diffusing a random process either along a tree or according to some simulation algorithm, and then evaluate the financial derivative backward. The underlying structure was a probability space and the tools were borrowed from probability theory. Let us (temporarily) forget about finance and switch to physics. In the 1940s, Richard Feynman[2] and Mark Kac[3] (both in Cornell University, but respectively in the Physics and in the Mathematics departments), established a link between stochastic processes and parabolic partial differential equations, originally with a view towards solutions of the heat equation and the Schrödinger equation. We now state the so-called *Feynman-Kac representation theorem* that makes this link precise.

Assumption 3.2.1. For any $t \geq 0$, the functions $\mu(\cdot, t)$ and $\sigma(\cdot, t)$ are globally Lipschitz, with at most linear growth, i.e. there exists a strictly positive constant C such that

$$\begin{aligned} |\mu(x,t)) - \mu(y,t))| &\leq C|x-y|, \text{ for any } t,x,y, \\ |\sigma(x,t)) - \sigma(y,t))| &\leq C|x-y|, \text{ for any } t,x,y, \\ |\mu(t,x)| &\leq C|x|, \text{ for any } t,x, \\ |\sigma(t,x)| &\leq C|x|, \text{ for any } t,x. \end{aligned}$$

Theorem 3.2.2 (Feynmac-Kac theorem)**.** *Consider the parabolic partial differential equation*

$$\partial_t u + \mu(x,t)\partial_x u + \frac{\sigma^2(x,t)}{2}\partial_{xx} u - ru = 0, \tag{3.2.1}$$

for all $(x,t) \in \mathbb{R} \times [0,T)$, with boundary condition $u(T,x) = \phi(x)$, where ϕ is continuous and satisfies $\phi(x) = \mathcal{O}(|x|)$ as $|x|$ tends to infinity. Then, under Assumption 3.2.1, the representation $u(t,x) = \mathbb{E}_{x,t}(\phi(X_T))$ holds, where $(X_t)_{t\geq 0}$ is the unique strong solution of the stochastic differential equation $\mathrm{d}X_s = \mu(s,X_s)\mathrm{d}s + \sigma(s,X_s)\mathrm{d}W_s$ starting at $X_t = x$.

An immediate corollary is of fundamental importance here:

Corollary 3.2.3. *Under the assumptions of Theorem 3.2.2, Equation* (3.2.1) *has a unique solution given by $u(t,x) = \mathbb{E}_{x,t}(\phi(X_T))$.*

Note that, modulo a change of variable $\tau := T - t$, the PDE here reduces to the heat equation (3.1.7) simply by taking $\sigma \equiv 0$ and $\mu \equiv 0$. More general versions of the theorem exist, and a

[2] http://turnbull.mcs.st-and.ac.uk/ history/Biographies/Feynman.html

[3] http://www-history.mcs.st-andrews.ac.uk/Biographies/Kac.html

precise and rigorous justification is outside the scope of these lectures. It however motivates the study of partial differential equations that we shall deal with now.

3.3 Discretisation schemes

We now focus on building up accurate numerical schemes to solve the partial differential equation

$$\partial_\tau u(\tau, x) = \frac{\sigma^2}{2}\partial^2_{xx} u(\tau, x), \tag{3.3.1}$$

for $\tau > 0$ and x in some interval $[x_L, x_U] \in \mathbb{R}$, with (Dirichlet) boundary conditions $u(0, x) = f(x)$ (payoff at maturity), $u(\tau, x_L) = f_L(\tau)$ and $u(\tau, x_U) = f_U(\tau)$. The last two boundary conditions allows one to compute the price of options such as up-and-out or down-and-out options as presented in Section 0.2.3. The two state-boundary points x_L and x_U may be infinite. The idea of finite-difference methods is to approximate each derivative by its first or second-order approximation, and then run a recursive algorithm starting from the time-boundary point. Before making this more precise, let us recall some basic facts on Taylor series. Let $g : \mathbb{R} \to \mathbb{R}$ be a three times continuously differentiable function. For any $x \in \mathbb{R}$, Taylor's series formula gives (with $\varepsilon > 0$)

$$g(x + \varepsilon) = g(x) + \varepsilon g'(x) + \frac{\varepsilon^2}{2} g''(x) + \frac{\varepsilon^3}{6} g'''(x) + \mathcal{O}\left(\varepsilon^4\right), \tag{3.3.2}$$

$$g(x - \varepsilon) = g(x) - \varepsilon g'(x) + \frac{\varepsilon^2}{2} g''(x) - \frac{\varepsilon^3}{6} g'''(x) + \mathcal{O}\left(\varepsilon^4\right). \tag{3.3.3}$$

Subtracting (3.3.3) from (3.3.2) gives

$$g'(x) = \frac{g(x + \varepsilon) - g(x - \varepsilon)}{2\varepsilon} + \mathcal{O}\left(\varepsilon^2\right). \tag{3.3.4}$$

The expression $\frac{1}{2\varepsilon}\left[g(x + \varepsilon) - g(x - \varepsilon)\right]$ is therefore an approximation of the derivative $g'(x)$ with an error of order ε^2. This is called a *central difference approximation* of g' at the point x. Note that Equations (3.3.2) and (3.3.3), taken separately, give the following approximations:

$$g'(x) = \frac{g(x + \varepsilon) - g(x)}{\varepsilon} + \mathcal{O}(\varepsilon), \tag{3.3.5}$$

$$g'(x) = \frac{g(x) - g(x - \varepsilon)}{\varepsilon} + \mathcal{O}(\varepsilon). \tag{3.3.6}$$

The first approximation is called the *forward difference* and the second approximation is the *backward difference*. They both have an error of order ε and are therefore less accurate than the central difference approximation. Concerning the second derivative of the function g, summing Equations (3.3.2) and (3.3.3) gives

$$g''(x) = \frac{g(x + \varepsilon) - g(x) + g(x - \varepsilon)}{\varepsilon^2} + \mathcal{O}\left(\varepsilon^2\right). \tag{3.3.7}$$

Note that this is the reason why we apply Taylor's formula in (3.3.2) and (3.3.3) up to order ε^4. If we had not done so, one could have erroneously concluded that the approximation for the second derivative would hold with an error of order ε instead of ε^2.

Exercise 16. Prove Equality (0.1.1) on Page 9.

We shall study three types of methods to solve the heat equation (3.3.1). Each of them relies on one of the above discretisation scheme for the approximation of the time derivative ∂_τ, while the space-derivative ∂_{xx} (and ∂_x whenever needed) is always approximated by central differences:

- the *implicit method* uses a backward difference scheme, leading to an error of order ε;
- the *explicit method* uses a forward difference scheme, leading to an error of order ε;
- the *Crank-Nicolson method* uses a central difference scheme, leading to an error of order ε^2.

Let us first start by constructing the time-space grid on which we will build the approximation scheme. The time boundaries are 0 and $T > 0$ (the maturity of the option) and the space boundaries are x_L and x_U. Let m and n be two integers. We consider a uniform grid, i.e. we split the space axis into m intervals and the time axis into n intervals, and we denote $\mathcal{I} := \{0, 1, \dots, n\}$ and $\mathcal{J} := \{0, 1, \dots, m\}$. This means that each point on the grid has coordinates $(i\delta_T, x_L + j\delta_x)$ for $i \in \mathcal{I}$ and $j \in \mathcal{J}$, where $\delta_T := \frac{T}{n}$ and $\delta_x := \frac{x_U - x_L}{m}$. At each node, we let $u_{i,j} := u\left(i\delta_T, x_L + j\delta_x\right)$ denote the value of the function u. Note in particular that the boundary conditions imply

$$u_{0,j} = f(x_L + j\delta_x), \qquad u_{i,0} = f_L(i\delta_T), \qquad u_{i,m} = f_U(i\delta_T).$$

From now on, we shall use the following (non standard) notation for ease of clarity.

Notation 3.3.1. For a tridiagonal matrix $\mathrm{T} \in \mathcal{M}_m(\mathbb{R})$, i.e.

$$\mathrm{T} := \begin{pmatrix} a_1 & c_1 & 0 & 0 \\ b_2 & a_2 & \ddots & 0 \\ 0 & \ddots & \ddots & c_{m-1} \\ 0 & 0 & b_m & a_m \end{pmatrix}, \tag{3.3.8}$$

we shall use the short-hand notation $\mathrm{T} = \mathrm{T}_m(\mathrm{a}, \mathrm{b}, \mathrm{c})$ for some $\mathbb{R}^m$-valued vectors a, b and c, or simply $\mathrm{T} = \mathrm{T}_m(a, b, c)$ when the entries in the vectors are all the same.

Remark 3.3.2. The heat equation (3.3.1) is an example of a convection-diffusion equation $-\partial_\tau + \gamma\partial_{xx} + \mu\partial_x = 0$, where $\gamma > 0$ is the diffusion coefficient and μ the convection coefficient. The schemes we shall study below are efficient (up to some precision) to solve this parabolic partial differential equation. However, when γ is very small (the so-called single perturbation of PDEs), these schemes are usually not accurate. In the limit, the structure of the PDE is indeed degenerate and reads $-\partial_\tau + \mu\partial_x = 0$. Other methods have been proposed in the literature, and we refer the interested reader to [23] for an overview of these.

3.3.1 Explicit scheme

In the explicit scheme, the time derivative ∂_τ is evaluated using the forward difference scheme (3.3.5), while the space second derivative ∂_{xx} is approximated with a central difference scheme. More precisely we consider the following approximations

$$\begin{aligned}\partial_\tau u(\tau,x) &= \frac{u(\tau+\delta_T,x)-u(\tau,x)}{\delta_T}+\mathcal{O}\left(\delta_T\right),\\ \partial_{xx}u(\tau,x) &= \frac{u(\tau,x+\delta_x)-2u(\tau,x)+u(\tau,x-\delta_x)}{\delta_x^2}+\mathcal{O}\left(\delta_x^2\right).\end{aligned}$$

Ignoring the terms of order δ_T and δ_x^2, the heat equation (3.3.1) at the node $(i\delta_T, x_L+j\delta_x)$ becomes

$$\frac{u_{i+1,j}-u_{i,j}}{\delta_T}+\mathcal{O}\left(\delta_T\right)=\frac{\sigma^2}{2}\frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{\delta_x^2}+\mathcal{O}\left(\delta_x^2\right), \tag{3.3.9}$$

which we can rewrite

$$u_{i+1,j}=\frac{\delta_T}{\delta_x^2}\frac{\sigma^2}{2}u_{i,j+1}+\left(1-\frac{\delta_T}{\delta_x^2}\sigma^2\right)u_{i,j}+\frac{\delta_T}{\delta_x^2}\frac{\sigma^2}{2}u_{i,j-1}, \tag{3.3.10}$$

for all $i=0,\dots,n-1$, $j=1,m-1$. Let us rewrite this in matrix form. To do so, define for each $i=0,\dots,n$ the vectors $\mathrm{u}_i\in\mathbb{R}^{m-1}$, $\mathrm{b}_i\in\mathbb{R}^{m-1}$ and the matrix $\mathrm{A}\in\mathcal{M}_{m-1}(\mathbb{R})$ by

$$\mathrm{u}_i:=\left(u_{i,1},\dots,u_{i,m-1}\right)^T,\qquad \mathrm{b}_i:=\left(u_{i,0},0,\dots,0,u_{i,m}\right)^T,\qquad \mathrm{A}:=\mathrm{T}_{m-1}\left(1-\alpha\sigma^2,\frac{\alpha\sigma^2}{2},\frac{\alpha\sigma^2}{2}\right),$$

where we introduce the quantity $\alpha:=\delta_T/\delta_x^2$. The recursion (3.3.9) thus becomes

$$\mathrm{u}_{i+1}=\mathrm{A}\mathrm{u}_i+\frac{\alpha\sigma^2}{2}\mathrm{b}_i,\qquad \text{for each } i=0,\dots,n-1,$$

where the time boundary condition reads $\mathrm{u}_0=\left(u_{0,1},\dots,u_{0,m-1}\right)^T=\left(f(x_L+\delta_x),\dots,f(x_L+(m-1)\delta_x)\right)^T$.

Example. Consider the heat equation with $\sigma=\sqrt{2}$, $x_L=0$, $x_U=1$ and boundary conditions $u(0,x)=f(x)=2x\mathbb{1}_{\{0\le x\le 1/2\}}+2\left(1-x\right)\mathbb{1}_{\{1/2\le x\le 1\}}$ and $f_L(\tau)=f_U(\tau)=0$ for any $\tau\in[0,T]$. The explicit solution to this initial value problem is given by

$$u(\tau,x)=\frac{8}{\pi^2}\sum_{n\ge 1}\frac{\sin(n\pi x)}{n^2}\sin\left(\frac{n\pi}{2}\right)\mathrm{e}^{-n^2\pi^2\tau}.$$

Implement the explicit scheme above and study the behaviour of the computed solution with respect to the discretisation parameters.

Remark 3.3.3. The explicit scheme—as well as the other schemes that will follow—computes the value of the function u at some points on the grid. In the case of the Black-Scholes model, we have performed quite a few changes of variables from the stock price S to the space variable x. Fix some time $t\ge 0$ (or remaining time τ). If one wants to compute the option value at some point S, it is not obvious to have the grid match the corresponding x value exactly. In that case one can perform some (linear) interpolation between the two points that are the closest to x.

Exercise 17. Comment on the link between the explicit scheme (3.3.10) and the trinomial tree scheme in Theorem 1.2.1. How are the up, down and middle probabilities expressed in terms of the discretisation parameters and of σ?

Remark 3.3.4. Note that as soon as the inequality $\sigma^2\delta_T/\delta_x^2 \leq 1$ is satisfied, Equation (3.3.10) implies that $u_{i+1,j}$ is a convex combination of the neighbouring three nodes at the previous time $i\delta_T$. In particular, if the initial datum $u_{0,\cdot}$ is bounded, say $\underline{u} \leq u_{0,j} \leq \overline{u}$, for all $j \in \mathcal{J}$ and for some constants $\underline{u}$ and $\overline{u}$, then the inequalities $\underline{u} \leq u_{i,j} \leq \overline{u}$ remain true for all $j \in \mathcal{J}$ and all $i \in \mathcal{I}$. This condition on δ_T and δ_x is called the CFL condition (after Richard Courant, Kurt Friedrichs, and Hans Lewy, who introduced for finite difference schemes of some classes of partial differential equations in 1928, see [13]) and clearly prevents the solution from unbounded oscillations. This stability of the discretisation scheme will be made rigorous in Section 3.3.7 below.

Exercise 18. As an example of Remark 3.3.4, consider the initial data $u_{0,j} = (-1)^j$ for each $j = 0, \ldots, m$. Compute explicitly the value of $u_{i,j}$ for all $(i,j) \in \mathcal{I} \times \mathcal{J}$. What can you conclude on the stability of the scheme?

3.3.2 Implicit scheme

In the implicit scheme, the time derivative ∂_τ is evaluated using the backward difference scheme (3.3.6), while the space second derivative ∂_{xx} is approximated with a central difference scheme. Ignoring the errors or orders δ_T and δ_x, the heat equation (3.3.1) therefore becomes

$$\frac{u(\tau,x) - u(\tau-\delta_T,x)}{\delta_T} = \frac{\sigma^2}{2}\frac{u(\tau,x+\delta_x) - 2u(\tau,x) + u(\tau,x-\delta_x)}{\delta_x^2},$$

which, at the node $(i\delta_T, x_L + j\delta_x)$, reads

$$\frac{u_{i,j} - u_{i-1,j}}{\delta_T} = \frac{\sigma^2}{2}\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\delta_x^2}.$$

Similarly as for the explicit scheme, we can reorganise the equality and we obtain

$$u_{i-1,j} = -\alpha\frac{\sigma^2}{2}u_{i,j+1} + \left(1+\alpha\sigma^2\right)u_{i,j} - \alpha\frac{\sigma^2}{2}u_{i,j-1}, \tag{3.3.11}$$

where as before we set $\alpha := \delta_T/\delta_x^2$. As in the explicit scheme, define for each $i = 1, \ldots, n$ the vectors $\mathrm{u}_i \in \mathbb{R}^{m-1}$, $\mathrm{b}_i \in \mathbb{R}^{m-1}$ and the matrix $\mathrm{A} \in \mathcal{M}_{m-1}(\mathbb{R})$ by

$$\mathrm{u}_i := (u_{i,1}, \ldots, u_{i,m-1})^T, \qquad \mathrm{b}_i := (u_{i,0}, 0, \ldots, 0, u_{i,m})^T, \qquad \mathrm{A} := \mathrm{T}_{m-1}\left(1+\alpha\sigma^2, -\frac{\alpha\sigma^2}{2}, -\frac{\alpha\sigma^2}{2}\right).$$

The recursion (3.3.11) becomes

$$\mathrm{u}_{i-1} = \mathrm{A}\mathrm{u}_i - \frac{\alpha\sigma^2}{2}\mathrm{b}_i, \qquad \text{for each } i = 0, \ldots, n-1,$$

with boundary condition $\mathrm{u}_0 = (u_{0,1}, \ldots, u_{0,m-1})^T = \Big(f(x_L+\delta_x), \ldots, f(x_L+(m-1)\delta_x)\Big)^T$.

3.3.3 Crank-Nicolson scheme

The Crank-Nicolson scheme uses the central difference approximation for the first-order time derivative. It was first described in 1947 (see [16]), and was subsequently fully developed in Los Alamos National Laboratory. Let us consider the point $\left(i\delta_T + \frac{1}{2}\delta_T, x_L + j\delta_x\right)$, and perform a Taylor series expansion around the point $i\delta_T + \frac{1}{2}\delta_T$. Equation (3.3.4) (with $\varepsilon = \frac{1}{2}\delta_T$) gives

$$\partial_\tau u \Big|_{i\delta_T + \frac{1}{2}\delta_T, x_L + j\delta_x} = \frac{u_{i+1,j} - u_{i,j}}{\delta_T} + \mathcal{O}\left(\delta_T^2\right).$$

For the space-derivative, we average the central differences between the points (i, j) and $(i + 1, j)$:

$$\partial_{xx} u \Big|_{\left(i+\frac{1}{2}\right)\delta_T, x_L + j\delta_x} = \frac{1}{2}\frac{u_{i+1,j+1} - 2u_{i+1,j} + u_{i+1,j-1}}{\delta_x^2} + \frac{1}{2}\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\delta_x^2} + \mathcal{O}\left(\delta_x^2\right).$$

The heat equation (3.3.1) thus becomes, after some reorganisation,

$$-\frac{\alpha\sigma^2}{4}u_{i+1,j+1} + \left(1 + \frac{\alpha\sigma^2}{2}\right)u_{i+1,j} - \frac{\alpha\sigma^2}{4}u_{i+1,j-1} = \frac{\alpha\sigma^2}{4}u_{i,j+1} + \left(1 - \frac{\alpha\sigma^2}{2}\right)u_{i,j} + \frac{\alpha\sigma^2}{4}u_{i,j-1}.$$

In matrix form, this reads

$$\mathrm{C}\mathrm{u}_{i+1} = \mathrm{D}\mathrm{u}_i + \frac{1}{2}\alpha\sigma^2\mathrm{b}_i, \qquad \text{for } i = 0, \dots, n-1,$$

where

$$\mathrm{C} := \mathrm{T}_{m-1}\left(1 + \frac{\alpha\sigma^2}{2}, -\frac{\alpha\sigma^2}{4}, -\frac{\alpha\sigma^2}{4}\right), \quad \mathrm{D} := \mathrm{T}_{m-1}\left(1 - \frac{\alpha\sigma^2}{2}, \frac{\alpha\sigma^2}{4}, \frac{\alpha\sigma^2}{4}\right),$$
$$\mathrm{b}_i := \left(\frac{u_{i,0} + u_{i+1,0}}{2}, 0, \dots, 0, \frac{u_{i,m} + u_{i+1,m}}{2}\right)^T \in \mathbb{R}^{m-1}.$$

Exercise 19. Although computing option prices is fundamental, computing the Greeks is a key ingredient in order to properly monitor the risks of the (portfolio of) options. Suppose we are approximating the PDE of an option using central differences for the space variable and forward differences in time, determine the order of accuracy for the Greeks defined in (1.1.13) on page 28.

Solution. *The different Greeks can be approximated by*

$$\Delta_{i,j} := \partial_x u(t,x)|_{(ij)} = \frac{u_{i,j+1} - u_{i,j-1}}{2\delta_x} + \mathcal{O}(\delta_x^2),$$
$$\Gamma_{i,j} := \partial_{xx}^2 u(t,x)\big|_{(ij)} = \frac{u_{i,j+1} - 2u_{ij} - u_{i,j-1}}{\delta_x^2} + \mathcal{O}(\delta_x^2),$$
$$\Theta_{i,j} := \partial_t u(t,x)|_{(ij)} = \frac{u_{i+1,j} - u_{i-1,j}}{2\delta_T} + \mathcal{O}(\delta_T^2).$$

Since the three schemes are accurate to order $\mathcal{O}\left(\delta_x^2\right)$, then the Delta is accurate to order $\mathcal{O}(\delta_x)$ and the Gamma is accurate to order $\mathcal{O}(1)$. Likewise, the Theta is accurate to order $\mathcal{O}(\delta_T)$ in the Crank-Nicolson scheme and to order $\mathcal{O}(1)$ in the implicit and explicit schemes.

Remark 3.3.5. Exercise 19 above highlighted the fact that accuracy is lost when computing the Greeks (i.e. the sensitivities of the option price with respect to its parameters). This is indeed a well-known drawback of Crank-Nicolson type schemes, as outlined for instance in [60]. More refined schemes have been proposed in the literature, and we refer the interested reader to the very good monograph [22].

Remark 3.3.6. The PDE we have studied so far in (3.3.1) was associated with Dirichlet boundary conditions. One could also consider Neumann boundary conditions of the type $\partial_x u(\tau, x_L) = a_L$ or $\partial_x u(\tau, x_U) = a_U$ (for any $\tau \geq 0$), for some real constants a_L, a_U. A first-order approximation (with error $\mathcal{O}(\delta_x)$) gives

$$a_L = \partial_x u(\tau, x_L) = \frac{u(\tau, x_L + \delta_x) - u(\tau, x_L)}{\delta_x} \qquad \text{and} \qquad a_U = \partial_x u(\tau, x_U) = \frac{u(\tau, x_U) - u(\tau, x_U - \delta_x)}{\delta_x}.$$

We can therefore 'eliminate' both $u(\tau, x_L)$ and $u(\tau, x_U)$ and we are left with $m - 1$ points to compute. However, since the three schemes above are of order $\mathcal{O}(\delta_x^2)$, accuracy is lost at the boundary. A second-order approximation of these boundary conditions is therefore needed:

$$\partial_x u(\tau, x_L) = \frac{u(\tau, x_L + \delta_x) - u(\tau, x_L - \delta_x)}{2\delta_x} \quad \text{and} \quad \partial_x u(\tau, x_U) = \frac{u(\tau, x_U + \delta_x) - u(\tau, x_U - \delta_x)}{2\delta_x}$$

where the error is of order $\mathcal{O}(\delta_x^2)$. We have however introduced here two fictitious points $x_L - \delta_x$ and $x_U + \delta_x$. Imposing that the PDE is satisfied at the boundaries, however, allows us to remove these fictitious points, and we are hence left (for each τ) with $m + 1$ points to evaluate.

3.3.4 Generalisation to θ-schemes

One may wonder why these three schemes (implicit, explicit and Crank-Nicolson) lead to a similar recurrence relation. Let us cast a new look at the heat equation (3.3.1) with $\sigma = \sqrt{2}$. A Taylor series expansion at some point (t, x) in the t direction gives $u(\tau + \delta_T, x) = \left(\mathrm{e}^{\delta_T \partial_\tau} u\right)(\tau, x)$, where we write the operator $\mathrm{e}^{\delta_T \partial_\tau}$ as a compact version of $1 + \delta_T \partial_\tau + \frac{1}{2}\delta_T^2 \partial_{\tau\tau}^2 + \ldots$. This implies that

$$u(\tau + \delta_T, x) - u(\tau, x) = \left(\mathrm{e}^{\delta_T \partial_\tau} - 1\right) u(\tau, x) =: \Delta_\tau u(\tau, x),$$

and hence $\partial_\tau = \delta_T^{-1} \log(1 + \Delta_\tau)$; Δ_τ is therefore a one-sided difference operator in the time variable. A Taylor expansion leads to

$$\partial_\tau = \frac{1}{\delta_T}\Delta_\tau - \frac{1}{2\delta_T}\Delta_\tau^2 + \mathcal{O}\left(\Delta_\tau^3\right). \tag{3.3.12}$$

For the central difference scheme, let Δ_x be the central difference operator defined by $\Delta_x u(\tau, x) := u\left(\tau, x + \frac{1}{2}\delta_x\right) - u\left(\tau, x - \frac{1}{2}\delta_x\right)$, which we can also write as

$$\Delta_x = \exp\left(\frac{\delta_x}{2}\partial_x\right) - \exp\left(-\frac{\delta_x}{2}\partial_x\right) = 2\sinh\left(\frac{\delta_x \partial_x}{2}\right)$$

in terms of the operator ∂_x. Therefore $\partial_x = \frac{2}{\delta_x}\operatorname{asinh}\left(\frac{1}{2}\Delta_x\right)$ and Taylor expansions give

$$\begin{aligned}\partial_x &= \frac{1}{\delta_x}\left(\Delta_x - \frac{\Delta_x^3}{24} + \mathcal{O}\left(\Delta_x^5\right)\right),\\ \partial_{xx}^2 &= \frac{1}{\delta_x^2}\left(\Delta_x^2 - \frac{\Delta_x^4}{12} + \mathcal{O}\left(\Delta_x^6\right)\right),\end{aligned} \tag{3.3.13}$$

where we recall that $\operatorname{asinh}(z) = z - \frac{1}{6}z^3 + \mathcal{O}(z^5)$ for z close to zero. Note further that $\Delta_x^2 u(\tau, x) = u(\tau, x+\delta_x) - 2u(\tau, x) + u(\tau, x-\delta_x)$. Between two points $(i\delta_T, x_L + j\delta_x)$ and $((i+1)\delta_T, x_L + j\delta_x)$ on the grid, the heat equation then reads, in operator form:

$$u_{i+1,j} = \mathrm{e}^{\delta_T \partial_\tau} u_{i,j} = \mathrm{e}^{\frac{1}{2}\sigma^2 \delta_T \partial_{xx}^2} u_{ij},$$

where the first equality follows by Taylor expansion (in time) and the second one holds since the function u solves the heat equation. Consider for instance $\tau = \theta i\delta_T + (1-\theta)(i+1)\delta_T$, where θ is some fixed real number in $[0,1]$, and denote u_{ij}^{θ} the value of the function u at this point. Assuming they exist, a forward Taylor expansion gives $u_{ij}^{\theta} = \mathrm{e}^{(1-\theta)\delta_T\partial_\tau} u_{i,j} = \mathrm{e}^{\frac{1}{2}\sigma^2(1-\theta)\delta_T\partial_{xx}^2} u_{i,j}$ and a backward Taylor expansion implies $u_{ij}^{\theta} = \mathrm{e}^{-\theta\delta_T\partial_\tau} u_{i+1,j} = \mathrm{e}^{-\frac{1}{2}\sigma^2\theta\delta_T\partial_{xx}^2} u_{i+1,j}$, so that we can write

$$\mathrm{e}^{-\frac{1}{2}\sigma^2\theta\delta_T\partial_{xx}^2} u_{i+1,j} = \mathrm{e}^{\frac{1}{2}\sigma^2(1-\theta)\delta_T\partial_{xx}^2} u_{i,j}.$$

From the Taylor expansions for the differential operators ∂_τ and ∂_{xx}^2 derived in (3.3.12) and in (3.3.13), we obtain—after tedious yet straightforward algebra—the so-called *θ-recurrence scheme*:

$$\left(1 + \alpha\theta\sigma^2\right) u_{i+1,j} - \frac{\alpha\theta\sigma^2}{2}\left(u_{i+1,j+1} + u_{i+1,j-1}\right) = \Big(1 - \alpha\sigma^2\left(1-\theta\right)\Big) u_{i,j} + \frac{\alpha\sigma^2}{2}\left(1-\theta\right)\left(u_{i,j+1} + u_{i,j-1}\right),$$

where we recall that $\alpha := \delta_T/\delta_x^2$. In matrix form, this can be rewritten as

$$\left(\mathrm{I} - \frac{\alpha\theta\sigma^2}{2}\mathrm{A}\right)\mathrm{u}_{i+1} = \left(\mathrm{I} + \frac{\alpha\sigma^2\left(1-\theta\right)}{2}\mathrm{A}\right)\mathrm{u}_i + \mathrm{b}_i, \qquad \text{for } i = 0, \ldots, n-1, \tag{3.3.14}$$

where

$$\mathrm{A} := \begin{pmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & -2 & 1 \\ 0 & \dots & 0 & 1 & -2 \end{pmatrix},$$

and where b_i represents the vector of boundary conditions. The explicit, implicit and Crank-Nicolson schemes are fully recovered by taking $\theta = 0$, $\theta = 1$ and $\theta = 1/2$. Other schemes are available, corresponding to other values of $\theta \in [0,1]$, but they fall beyond the scope of this course.

3.3.5 Multi-step schemes

The three schemes above are the most common schemes used in practice; it is however possible to construct more elaborate schemes, providing better approximations (with additional computational cost, of course). For some function $g \in \mathcal{C}^3(\mathbb{R} \to \mathbb{R})$ and a point $x \in \mathbb{R}$, consider the Taylor series expansions, for some $\varepsilon > 0$:

$$f(x-\varepsilon) = f(x) - \varepsilon f'(x) + \frac{\varepsilon^2}{2} f''(x) + \mathcal{O}(\varepsilon^3),$$
$$f(x-2\varepsilon) = f(x) - 2\varepsilon f'(x) + 2\varepsilon^2 f''(x) + \mathcal{O}(\varepsilon^3),$$

so that the derivative of f can be approximated by

$$f'(x) = \frac{f(x-2\varepsilon) - 4f(x-\varepsilon) + 3f(x)}{2\varepsilon} + \mathcal{O}(\varepsilon^2).$$

Applying this to the heat equation at some point $((i+2)\delta_T, x_l + j\delta_x)$ on the grid, together with a central order difference for the second derivative in space, we obtain

$$\frac{u_{i,j} - 4u_{i+1,j} + 3u_{i+2,j}}{2\delta_T} + \mathcal{O}(\delta_T^2) = \frac{u_{i+2,j+1} - 2u_{i+2,j} + u_{i+2,j-1}}{\delta_x^2} + \mathcal{O}(\delta_x^2),$$

for any $i = 0, \ldots, n-2$ and $j = 1, \ldots, m-1$, which we can rewrite

$$-\alpha\sigma^2 u_{i+2,j+1} + \left(3 + 2\alpha\sigma^2\right) u_{i+2,j} - \alpha\sigma^2 u_{i+2,j-1} = 4u_{i+1,j} - u_{i,j},$$

or, in vector notations,

$$\mathrm{A}\mathrm{u}_{i+2} = 4\mathrm{u}_{i+1} - \mathrm{u}_i + \alpha\sigma^2 \mathrm{b}_{i+2},$$

where $\mathrm{A} := \mathrm{T}_{m-1}(3 + 2\alpha\sigma^2, -\alpha\sigma^2, -\alpha\sigma^2)$ and again $\mathrm{b}_i := (u_{i,0}, 0, \ldots, 0, u_{i,m}) \in \mathbb{R}^{m-1}$. Note than, in order to compute the first iteration (at time $2\delta_T$), we need to know, not only the value function u_0 at the boundary, but also at the first time step δ_T. This can be done by including an initialisation step using a one-point in time discretisation between time zero and time δ_T.

3.3.6 Non-uniform grids

We have so far considered uniform grids (or meshes). It may be useful, for instance when considering barrier options, to construct the scheme with a mesh taking into account of the singularity of the payoff.

Direct approach

Let us consider a uniform mesh in time as above, and the following mesh in space: we fix an integer $k \in \{1, \ldots, m-1\}$ such that the mesh is uniform on $[x_0, x_k]$ and uniform on $[x_k, x_m]$, but with spacing $\delta_{x,1}$ on the first interval and $\delta_{x,2}$ on the second one. For some (smooth enough) function g around x_k, we can rewrite the Taylor expansions (3.3.2) and (3.3.3) as

$$\begin{aligned} g(x_k - \delta_{x,1}) &= g(x_k) - \delta_{x,1} g'(x_k) + \frac{\delta_{x,1}^2}{2} g''(x_k) - \frac{\delta_{x,1}^3}{6} g'''(x_k) + \mathcal{O}\left(\delta_{x,1}^4\right), \\ g(x_k + \delta_{x,2}) &= g(x_k) + \delta_{x,2} g'(x_k) + \frac{\delta_{x,2}^2}{2} g''(x_k) + \frac{\delta_{x,2}^3}{6} g'''(x_k) + \mathcal{O}\left(\delta_{x,2}^4\right). \end{aligned}$$

However, we cannot simply subtract the two expansions any longer in order to obtain an approximation of the derivatives since $\delta_{x,1} \neq \delta_{x,2}$. Multiplying the first equation by $-\delta_{x,2}^2$ and the second one by $\delta_{x,1}^2$, though, and adding them, we obtain, after simplifications

$$g'(x_k) = -\frac{\delta_{x,2} g(x_k - \delta_{x,1})}{\delta_{x,1}\left(\delta_{x,1} + \delta_{x,2}\right)} + \frac{\delta_{x,1} g(x_k + \delta_{x,2})}{\delta_{x,2}\left(\delta_{x,1} + \delta_{x,2}\right)} - \frac{\delta_{x,1} - \delta_{x,2}}{\delta_{x,1}\delta_{x,2}} g(x_k) + \mathcal{O}\left(\delta_{x,1}\delta_{x,2}\right).$$

For the second derivative, we can do similar computations and obtain

$$g''(x_k) = \frac{2g(x_k - \delta_{x,1})}{\delta_{x,1}(\delta_{x,1} + \delta_{x,2})} + \frac{2g(x_k + \delta_{x,2})}{\delta_{x,2}(\delta_{x,1} + \delta_{x,2})} - \frac{2g(x_k)}{\delta_{x,1}\delta_{x,2}} + \mathcal{O}(\delta_{x,1} + \delta_{x,2}).$$

Note, however, that the approximation for the second derivative here has an error of order one only. Obtaining an order two is possible by using a four-point approximation, but this would break the tridiagonal structure of the iteration matrix.

Coordinate transformation

We are now interested in constructing a non-uniform grid (in space) while preserving the second-order accuracy of the finite-difference scheme. The idea is as follows: we construct a uniform grid on the closed interval $[0, 1]$, and then map it in a non-linear way to the interval $[\underline{S}, \overline{S}]$, the non-linearity of the map breaking the uniformity of the mesh. More precisely, let $S : [0, 1] \to [\underline{S}, \overline{S}]$ be an increasing continuous map such that $S(0) = \underline{S}$ and $S(1) = \overline{S}$. Consider now the Black-Scholes PDE for the function V, as written in (3.1.5), and let us define $\psi(\tau, \xi) \equiv V_t(S)$ for $\xi \in [0, 1]$, with again $\tau := T - t$. Since

$$\partial_t V = -\partial_\tau \psi, \qquad \partial_S V = \frac{\partial_\xi \psi}{S'(\xi)} \qquad \text{and} \qquad \partial_{SS} V = \frac{\partial_{\xi\xi}\psi}{S'(\xi)^2} - \frac{S'(\xi)^2}{S'(\xi)^3}\partial_\xi \psi,$$

the Black-Scholes PDE (3.1.5) then reads

$$\partial_\tau \psi(\tau, \xi) = \frac{\sigma^2}{2}\frac{S(\xi)^2}{S'(\xi)^2}\partial_{\xi\xi}\psi + \left(r - \frac{\sigma^2}{2}\frac{S(\xi)S''(\xi)}{S'(\xi)^2}\right)\frac{S(\xi)}{S'(\xi)}\partial_\xi \psi - r\psi(\tau, \xi),$$

with appropriate boundary conditions.

Exercise 20. (Notebook *BSPDE_NonUniformGrid*)
Consider the map

$$S(\xi) := B + \alpha \sinh(c_1\xi + c_2(1 - \xi)), \qquad \xi \in [0, 1].$$

Given B and α, compute c_1 and c_2 in order to ensure that $S(0) = \underline{S}$ and $S(1) = \overline{S}$. Discuss the influence of the parameters B and α on the behaviour of the mesh.

3.3.7 Stability and convergence analysis

We start this section with a simple example. Suppose we are interested in solving the heat equation (3.3.1) with an explicit difference scheme, as developed in Section 3.3.1. For simplicity, we shall assume that $\sigma = \sqrt{0.2}$. We also consider the following boundary conditions:

$$u(0, x) = x^2 \mathbb{1}_{\{x\in[0,1/2]\}} + (1 - x)^2 \mathbb{1}_{\{x\in[1/2,1]\}} \qquad \text{and} \qquad u(\tau, 0) = u(\tau, 1) = 0, \text{ for all } \tau \geq 0.$$

We plot in Figure 3.1 the outputs of the numerical implementation, where we consider $(\delta_x, \delta_T) = (0.1, 0.001)$ and $(\delta_x, \delta_T) = (0.1, 0.1)$. The grid mesh is $(\delta_x)^{-1}$ and $(\delta_T)^{-1}$. It is clear that increasing the time step in the second figure leads to high numerical instability. We now move on to a rigorous justification of this fact.

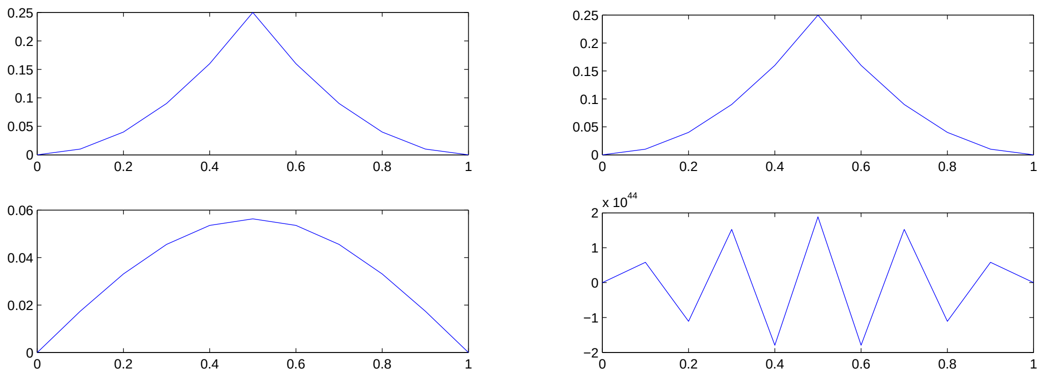

Figure 3.1: Explicit finite difference method for the heat equation with $\delta_x = 0.1$ and $\delta_T = 0.001$ (left) and $\delta_T = 0.1$ (right). The upper plot corresponds to the solution at time 0 (initial condition) and the lower plot to the solution at time 1.

A Fourier transform approach

Let us start with a few preliminary definitions and notations. We shall restrict ourselves here—mainly for notational reasons—as before to the one-dimensional case, even though most of the discussion below carries over to more general spaces. With a view towards more generality, let Φ denote a (differential) operator acting on the space of smooth real functions $C^\infty(\mathbb{R}_+, \mathbb{R})$. In the case of the heat equation (3.3.1), $\Phi := \mathcal{L} - \partial_\tau$, where $\mathcal{L} = \frac{1}{2}\sigma^2\partial_{xx}$ is the rescaled Laplace operator. In the following, we shall denote ψ^* the (unique up to boundary point specification) solution to the equation $\Phi\left(\psi^*\right) = 0$. Define further $\widetilde{\Phi} := (\widetilde{\Phi}_{ij})_{i\in\mathcal{I},j\in\mathcal{J}}$ as the finite difference approximation of Φ on the grid $\mathcal{I} \times \mathcal{J}$ (see Exercise 21 for an example). Note that the equation $\Phi(\psi) = 0$, $\psi \in C^\infty(\mathbb{R}_+, \mathbb{R})$, is a partial differential equation, whereas for any $(i, j) \in \mathcal{I} \times \mathcal{J}$, $\widetilde{\Phi}_{ij}(\psi) = 0$ is a so-called difference equation.

Definition 3.3.7. For any smooth real function $\psi \in C^\infty(\mathbb{R}_+, \mathbb{R})$, the *truncation error* E_{ij} at the node (i, j) is defined as

$$E_{ij}(\psi) := \widetilde{\Phi}_{ij}(\psi) - \Phi(\psi)|_{(i\delta_T, x_L + j\delta_x)}\,.$$

Since ψ^* solves the PDE $\Phi(\psi) = 0$, we immediately see that $E_{ij}(\psi^*) = \widetilde{\Phi}_{ij}(\psi^*)$, and we shall call this quantity the *local truncation error* at the lattice point (i, j).

Definition 3.3.8. The finite difference scheme $(\widetilde{\Phi}_{ij})_{i\in\mathcal{I},j\in\mathcal{J}}$ is said to be *consistent* if for any $(i, j) \in \mathcal{I} \times \mathcal{J}$, the truncation error $E_{ij}(\psi)$ converges to zero as δ_T and δ_x tend to zero for any smooth real function ψ.

Exercise 21. Consider the equation defined by $\Phi(\psi) = (\partial_t + \partial_x)(\psi) \equiv 0$. Assume a forward-time forward-space discretisation scheme:

$$\widetilde{\Phi}_{i,j}(\psi) := \frac{\psi_{i+1,j} - \psi_{i,j}}{\delta_T} + \frac{\psi_{i,j+1} - \psi_{i,j}}{\delta_x}.$$

Prove that the scheme is consistent.

Solution. *A Taylor expansion gives*

$$\begin{aligned} \psi_{i+1,j} &= \psi_{i,j} + \delta_T \partial_t \psi_{i,j} + \frac{1}{2}\delta_T^2 \partial_{tt}^2 \psi_{i,j} + \mathcal{O}(\delta_T^3), \\ \psi_{i,j+1} &= \psi_{i,j} + \delta_x \partial_x \psi_{i,j} + \frac{1}{2}\delta_x^2 \partial_{xx}^2 \psi_{i,j} + \mathcal{O}(\delta_x^3). \end{aligned}$$

Plugging these expansions into the definition of the $\widetilde{\Phi}$, we can compute the truncation error for any $(i,j) \in \mathcal{I} \times \mathcal{J}$ as

$$\begin{aligned} E_{i,j}(\psi) &= \frac{\psi_{i+1,j} - \psi_{i,j}}{\delta_T} + \frac{\psi_{i,j+1} - \psi_{i,j}}{\delta_x} - \partial_t \psi_{ij} - \partial_x \psi_{ij} \\ &= \left(\partial_t \psi_{i,j} + \frac{1}{2}\delta_T \partial_{tt}^2 \psi_{i,j} + \mathcal{O}(\delta_T^2)\right) + \left(\partial_x \psi_{i,j} + \frac{1}{2}\delta_x \partial_{xx}^2 \psi_{i,j} + \mathcal{O}(\delta_x^2)\right) - \partial_t \psi_{ij} - \partial_x \psi_{ij} \\ &= \frac{1}{2}\delta_T \partial_{tt}^2 \psi_{i,j} + \mathcal{O}(\delta_T^2) + \frac{1}{2}\delta_x \partial_{xx}^2 \psi_{i,j} + \mathcal{O}(\delta_x^2), \end{aligned}$$

which clearly converges to zero as δ_x and δ_T tend to zero.

Definition 3.3.9. We define the orders of accuracy p^* (in δ_T) and q^* (in δ_x) as

$$(p^*, q^*) := \sup_{(i,j) \in \mathcal{I} \times \mathcal{J}} \left\{(p,q) \in \mathbb{N}^2 : |E_{ij}(\psi)| \leq C\left(\delta_T^p + \delta_x^q\right), \text{ for some } C > 0\right\}.$$

Exercise 22. What are the orders of accuracy in δ_T and in δ_x of the explicit scheme and the Crank-Nicolson scheme? Are these schemes consistent with the original heat equation?

Solution. *For the explicit scheme, $p^* = 1$ and $q^* = 2$, whereas in the Crank-Nicolson scheme, $p^* = q^* = 2$.*

Let $\widetilde{\psi} := (\widetilde{\psi}_{ij})_{i \in \mathcal{I}, j \in \mathcal{J}}$ be the exact solution of the finite difference approximation scheme $\widetilde{\Phi}$ (i.e. $\widetilde{\Phi}(\widetilde{\psi}) = 0$), and define the *discretisation error* ε by

$$\varepsilon_{ij} := \widetilde{\psi}_{ij} - \psi_{ij}^*, \qquad \text{for each } i \in \mathcal{I}, j \in \mathcal{J},$$

namely the difference, at the point (i,j) on the grid, between the solution of the difference equation and that of the differential equation.

Definition 3.3.10. The scheme is said to converge in the $\|\cdot\|$ norm if

$$\lim_{(\delta_T, \delta_x) \downarrow (0,0)} \sup_{i \in \mathcal{I}, j \in \mathcal{J}} \|\varepsilon_{ij}\| = 0.$$

The story is however not complete yet: first, the truncation error in Definition 3.3.7 is a local concept whereas convergence in norm is a global concept. Then we have to make it clear what is meant by the double limit in (δ_T, δ_x) in the definition of convergence? On a grid with spacing δ and dimension m, let us define the norm $\|\cdot\|_\delta$ by

$$\|v\|_\delta := \left(\delta \sum_{j=1}^{m} v_j^2\right)^{1/2},$$

for any $\mathbb{R}^m$-valued vector $v = (v_j)_{j=1,\dots,m}$. It is an l^2-norm on the grid and measures the size of the solution on the grid. For a matrix $(\psi_{ij})_{i\in\mathcal{I},j\in\mathcal{J}}$ defined on the grid $\mathcal{I}\times\mathcal{J}$, we define $\psi_{i\cdot}$, for any $i \in \mathcal{I}$, as the vector corresponding to the i-th line, so that $\psi_{i\cdot} \in \mathbb{R}^m$, where $m := \dim(\mathcal{J})$.

Definition 3.3.11. The finite difference scheme $\widetilde{\Phi}$ is said to be *stable* in some stability region $\mathcal{S}$ if there exists $n_0 \in \mathbb{N}$ such that for all $T > 0$, there is a strictly positive constant C_T satisfying

$$\left\|\widetilde{\psi}_{n\cdot}\right\|_{\delta_x}^2 \le C_T \sum_{i=0}^{n_0} \left\|\widetilde{\psi}_{i\cdot}\right\|_{\delta_x}^2, \qquad \text{for all } 0 \le n\delta_T \le T \text{ and } (\delta_x, \delta_T) \in \mathcal{S}.$$

The stability inequality above expresses the idea that the norm of the solution vector, at any point in time, is limited by the sums of the norms of the vector solution up to time n_0.

Example. We illustrate this definition with a scheme for the equation introduced in Example 21. Consider a general forward-time forward-space discretisation scheme of the form

$$\psi_{i+1,j} = \alpha\psi_{i,j} + \beta\psi_{i,j+1},$$

for some α and β. Then

$$\begin{aligned}
\|\psi_{i+1,\cdot}\|_{\delta_x}^2 &= \delta_x \sum_j |\psi_{i+1,j}|^2 = \delta_x \sum_j |\alpha\psi_{i,j} + \beta\psi_{i,j+1}|^2 \\
&\le \delta_x \sum_j \left(\alpha^2 |\psi_{i,j}|^2 + \beta^2 |\psi_{i,j+1}|^2 + 2|\alpha||\beta|\,|\psi_{i,j}|\,|\psi_{i,j+1}|\right) \\
&\le \delta_x \sum_j \left(\alpha^2 |\psi_{i,j}|^2 + \beta^2 |\psi_{i,j+1}|^2 + |\alpha||\beta|\left(|\psi_{i,j}|^2 + |\psi_{i,j+1}|^2\right)\right),
\end{aligned}$$

since $x^2 + y^2 \ge 2xy$ for any $(x, y) \in \mathbb{R}^2$. Splitting the indices j and $j + 1$, we obtain

$$\begin{aligned}
\|\psi_{i+1,\cdot}\|_{\delta_x}^2 &\le \delta_x \left(\alpha^2 + |\alpha||\beta|\right) \sum_j |\psi_{i,j}|^2 + \delta_x \left(\beta^2 + |\alpha||\beta|\right) \sum_j |\psi_{i,j+1}|^2 \\
&\le \delta_x \sum_j \left(\alpha^2 + 2|\alpha||\beta| + \beta^2\right) |\psi_{i,j}|^2 \\
&= (|\alpha| + |\beta|)^2 \delta_x \sum_j |\psi_{i,j}|^2 .
\end{aligned}$$

Therefore $\|\psi_{i+1,\cdot}\|_{\delta_x} \le (|\alpha| + |\beta|)^2 \|\psi_{i\cdot}\|_{\delta_x}$. Note that we have here omitted the boundary terms arising from the second sum in the first line. We shall assume here that they are not relevant. Repeating this, we obtain

$$\|\psi_{i,\cdot}\|_{\delta_x}^2 \le (|\alpha| + |\beta|)^{2i} \|\psi_{0\cdot}\|_{\delta_x}^2,$$

and hence the scheme is stable if and only if $|\alpha| + |\beta| \le 1$

Recall now that the L^2-norm of a function $\psi : \mathbb{R} \to \mathbb{R}$ is defined by $\|\psi\|_{L^2(\mathbb{R})} := \left(\int_{\mathbb{R}} \psi(x)^2 \mathrm{d}x\right)^{1/2}$.

Definition 3.3.12. The partial differential equation $\Phi\psi = 0$ is said to be *well-posed* if for each $T > 0$ there exists $C_T > 0$ such that

$$\|\psi(t, \cdot)\|_{L^2(\mathbb{R})} \leq C_T \|\psi(0, \cdot)\|_{L^2(\mathbb{R})},$$

for all $t \in [0, T]$ and for any solution ψ.

We can now state the main result of this section, which provides a full characterisation of the convergence of a scheme.

Theorem 3.3.13 (Lax Equivalence theorem)**.** *A consistent finite difference scheme of a well-posed linear initial-valued problem converges if and only if it is stable .*

Remark 3.3.14. One may wonder how this result changes when studying an inhomogeneous PDE such as $\Phi\psi = g$ for some function g. *Duhamel's principle* says that the solution to such a problem can be written as the superposition of solutions to the homogeneous PDE $\Phi\psi = 0$. Therefore the concepts of well-posedness and stability follow from the homogeneous case.

The importance of this theorem stems from the fact that convergence is not easy to check directly from the definition. However, well-posedness and stability are much easier to check in practice. By means of Fourier methods, we shall now give a precise and easy-to-check condition for the stability of the finite difference scheme. Let $v = (v_{i,j})$ be a function defined on a grid with space increment δ_x and time increment δ_T. We assume for now that there is no restriction in the space domain (i.e. $x \in \mathbb{R}$), and we fix some time $n\delta_T$. The discrete Fourier transform of the vector v_n is defined as

$$\widehat{v}_n(\xi) := \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} \delta_x \mathrm{e}^{-\mathtt{i} m \delta_x \xi} v_{n,m}, \qquad \text{for } \xi \in \Pi_x := \left[-\frac{\pi}{\delta_x}, \frac{\pi}{\delta_x}\right], \tag{3.3.15}$$

and we have the inverse Fourier transform formula

$$v_{n,m} = \frac{1}{\sqrt{2\pi}} \int_{-\pi/\delta_x}^{\pi/\delta_x} \mathrm{e}^{\mathtt{i} m \delta_x \xi} \widehat{v}_n(\xi) \mathrm{d}\xi. \tag{3.3.16}$$

Assume now that we have a (one-step in time) finite difference scheme which we write as

$$v_{n+1,m} = \sum_{j=-d}^{u} \alpha_j(\delta_T, \delta_x) v_{n,m+j}. \tag{3.3.17}$$

This means that at each grid point $((n+1)\delta_T, m\delta_x)$ we can write the scheme using some of the grid points at time $n\delta_T$. The positive integers d and u represent how far up and down we have to go along the grid at time $n\delta_T$. Applying the inverse Fourier transform formula (3.3.16) to the finite difference scheme (3.3.17), we have

$$
\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{-\pi/\delta_x}^{\pi/\delta_x} \mathrm{e}^{\mathtt{i}m\delta_x\xi}\widehat{v}_{n+1}(\xi)\mathrm{d}\xi = v_{n+1,m} &= \sum_{j=-d}^{u}\alpha_j(\delta_T,\delta_x)v_{n,m+j}\\
&= \sum_{j=-d}^{u}\alpha_j(\delta_T,\delta_x)\frac{1}{\sqrt{2\pi}}\int_{-\pi/\delta_x}^{\pi/\delta_x} \mathrm{e}^{\mathtt{i}(m+j)\delta_x\xi}\widehat{v}_n(\xi)\mathrm{d}\xi\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\pi/\delta_x}^{\pi/\delta_x} \mathrm{e}^{\mathtt{i}m\delta_x\xi}\sum_{j=-d}^{u}\alpha_j(\delta_T,\delta_x)\mathrm{e}^{\mathtt{i}j\delta_x\xi}\widehat{v}_n(\xi)\mathrm{d}\xi,
\end{aligned}
$$

which implies, by unicity of the Fourier transform, that

$$
\widehat{v}_{n+1}(\xi) = \sum_{j=-d}^{u}\alpha_j(\delta_T,\delta_x)\mathrm{e}^{\mathtt{i}j\delta_x\xi}\widehat{v}_n(\xi) =: \zeta\left(\xi,\delta_T,\delta_x\right)\widehat{v}_n(\xi),
$$

for all $\xi \in \Pi_x$, where the function ζ is defined in an obvious way, and does not depend on n. Iterating this equality, we obtain

$$
\widehat{v}_n(\xi) = \zeta\left(\xi,\delta_T,\delta_x\right)^n \widehat{v}_0(\xi). \tag{3.3.18}
$$

Recall Parseval identity:

$$
\|\widehat{v}_n\|^2_{L^2(\Pi_x)} := \int_{-\pi/\delta_x}^{\pi/\delta_x}\left|\widehat{v}_n(\xi)\right|^2\mathrm{d}\xi = \sum_{m=-\infty}^{\infty}\delta_x\left|v_{n,m}\right|^2 = \|v_{n,\cdot}\|^2_{\delta_x}. \tag{3.3.19}
$$

Recalling the definition of stability (Definition 3.3.11) of a finite difference scheme, we see that it should be possible to express it simply in terms of the function ζ. The following theorem makes this precise and provides us with a ready-to-use condition to check stability.

Theorem 3.3.15. *A (one-step in time) finite difference scheme is stable if and only if there exist $K > 0$ (independent of ξ, δ_x, δ_T) and $\left(\delta_T^0, \delta_x^0\right)$ such that*

$$
|\zeta(\xi,\delta_T,\delta_x)| \leq 1 + K\delta_T^0,
$$

for all ξ, $\delta_T \in (0, \delta_T^0]$ and $\delta_x \in (0, \delta_x^0]$. In particular, when the function ζ does not depend on δ_T and δ_x, the condition $|\zeta(\xi)| \leq 1$ is sufficient.

The factor ζ is called the *amplification factor* and this analysis is called *von Neumann analysis*, in memory of its founder. Applying Theorem 3.3.15 to (3.3.19) and using (3.3.18), we now see that

$$
\begin{aligned}
\|v_{n,\cdot}\|^2_{\delta_x} = \|\widehat{v}_n\|^2_{L^2(\Pi_x)} &\leq \int_{-\pi/\delta_x}^{\pi/\delta_x}\left|\zeta(\xi,\delta_T,\delta_x)\right|^{2n}\left|\widehat{v}_0(\xi)\right|^2\mathrm{d}\xi\\
&= \left(1+K\delta_T^0\right)^{2n}\|\widehat{v}_0\|^2_{L^2(\Pi_x)}\\
&= \left(1+K\delta_T^0\right)^{2n}\|v_{0,\cdot}\|^2_{\delta_x}.
\end{aligned}
$$

We now apply this result to the three finite difference schemes developed above to the heat equation $\partial_\tau u = \gamma\partial_{xx}^2 u$, with $\gamma > 0$.

Application to θ-schemes

Let us first consider the von Neumann analysis of the explicit scheme. The explicit finite difference (3.3.10) can be rewritten as

$$u_{n+1,m} = u_{n,m} + \alpha\gamma\left(u_{n,m+1} - 2u_{n,m} + u_{n,m-1}\right), \tag{3.3.20}$$

where we recall that $\alpha := \delta_T/\delta_x^2$ and $\gamma = \sigma^2/2$. Writing $u_{n,m}$ in terms of its Fourier transform (3.3.16) and using the relation (3.3.18), the amplification factor reads

$$\zeta(\xi,\delta_T,\delta_x) = 1 + \alpha\gamma\left(\mathrm{e}^{\mathrm{i}\xi\delta_x} - 2 + \mathrm{e}^{-\mathrm{i}\xi\delta_x}\right) = 1 + 2\alpha\gamma\left(\cos(\xi\delta_x) - 1\right) = 1 - 4\alpha\gamma\sin\left(\frac{\xi\delta_x}{2}\right)^2.$$

Hence $|\zeta(\xi,\delta_T,\delta_x)| \leq 1$ if and only if $\alpha\gamma \leq 1/2$. The scheme is hence *conditionally stable.*

Remark 3.3.16. Note that we could argue more directly here: if the inequality $\alpha\gamma \leq 1/2$ is satisfied then using the maximum norm, the scheme (3.3.20) implies

$$\left\|u_{n+1,\cdot}\right\|_\infty \leq \left\|u_{n,\cdot}\right\|_\infty,$$

and we can conclude on its stability.

The implicit finite difference (3.3.11) can be rewritten as

$$u_{n,m} = u_{n-1,m} + \alpha\gamma\left(u_{n,m+1} - 2u_{n,m} + u_{n,m-1}\right). \tag{3.3.21}$$

The amplification factor is

$$\zeta(\xi,\delta_T,\delta_x) = \left(1 - \alpha\gamma\left(\mathrm{e}^{\mathrm{i}\xi\delta_x} - 2 + \mathrm{e}^{-\mathrm{i}\xi\delta_x}\right)\right)^{-1} = \left(1 + 4\alpha\gamma\sin\left(\frac{\xi\delta_x}{2}\right)^2\right)^{-1},$$

and the inequality $|\zeta(\xi,\delta_T,\delta_x)| \leq 1$ always holds. The scheme is therefore *unconditionally stable.*

Exercise 23. Prove that the amplification factor for the Crank-Nicolson scheme is

$$\zeta(\xi,\delta_T,\delta_x) = \frac{1 - 2\alpha\gamma\sin\left(\frac{\xi\delta_x}{2}\right)^2}{1 + 2\alpha\gamma\sin\left(\frac{\xi\delta_x}{2}\right)^2},$$

and conclude on the stability of the scheme.

3.3.8 Convergence analysis via matrices

We review here the convergence analysis from a different—albeit equivalent—point of view.

A crash course of matrix norms

We recall here some basic facts about vector and matrix norms, which shall be useful for a full understanding of the matrix approach of convergence of the finite difference schemes. We let $\mathrm{x} := (x_1, \dots, x_n)$ be a vector in $\mathbb{C}^n$ for some fixed $n \in \mathbb{N}^*$. The norm $\|\cdot\|$ of a vector is a real non-negative number that gives a measure of its size. It has to satisfy the following properties:

- $\|\mathrm{x}\| > 0$ if $\mathrm{x} \neq 0$ and $\|\mathrm{x}\| = 0$ if $\mathrm{x} = 0$;
- $\|\alpha\mathrm{x}\| = |\alpha|\|\mathrm{x}\|$, for any $\alpha \in \mathbb{C}$;
- $\|\mathrm{x} + \mathrm{y}\| \leq \|\mathrm{x}\| + \|\mathrm{y}\|$, for any $\mathrm{x}, \mathrm{y} \in \mathbb{C}^n$.

The most common norms are

- the L^1-norm (or taxicab norm): $\|\mathrm{x}\|_1 := \sum_{i=1}^n |x_i|$;
- the L^p-norm ($p \geq 1$): $\|\mathrm{x}\|_p := \left(\sum\limits_{i=1}^n |x_i|^p\right)^{1/p}$;
- the infinity-norm: $\|\mathrm{x}\|_\infty := \max\limits_{i=1,\dots,n} |x_i|$.

For a matrix A in $\mathcal{M}_n(\mathbb{C})$, we define its norm as follows:

Definition 3.3.17. Let $\|\cdot\|$ be a vector norm on $\mathbb{C}^n$. We then define the *subordinate* matrix norm (and by a slight abuse of language, use the same notation $\|\cdot\|$) by

$$\|A\| := \sup_{\mathrm{x}\in\mathbb{C}^n\setminus\{0\}} \frac{\|A\mathrm{x}\|}{\|\mathrm{x}\|}.$$

The following lemma, the proof of which is straightforward and left to the reader, gathers some immediate properties of subordinate matrix norms:

Lemma 3.3.18. *Let $\|\cdot\|$ be a subordinate matrix norm on $\mathbb{C}^n$.*

- *the equalities $\|A\| = \sup_{\mathrm{x}\in\mathbb{C}^n,\|\mathrm{x}\|=1} \|A\mathrm{x}\| = \sup_{\mathrm{x}\in\mathbb{C}^n,\|\mathrm{x}\|\leq 1} \|A\mathrm{x}\|$ hold;*
- *the identity matrix* I *satisfies $\|\mathrm{I}\| = 1$;*
- *for two matrices* A *and* B *in $\mathcal{M}_n(\mathbb{C})$, the inequality $\|\mathrm{AB}\| \leq \|\mathrm{A}\|\|\mathrm{B}\|$ always holds.*

Remark 3.3.19. The Euclidean matrix norm defined by $\|\mathrm{A}\| := \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}$ for $\mathrm{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ defines a matrix norm, but $\|\mathrm{I}\| = \sqrt{n}$. It is therefore not a subordinate matrix norm.

Exercise 24. Let $\mathrm{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$. Prove the following statements:

- $\|\mathrm{A}\|_2 = \|A^*\|_2$, where A^* denotes the conjugate transpose of A;
- $\|\mathrm{A}\|_2$ is equal to the largest singular value of A, or equivalently to the square root of the largest eigenvalue of the positive semidefinite matrix A^*A;

- $\|\mathrm{A}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$;
- $\|\mathrm{A}\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$;

Norms of matrices *measure* in some sense their sizes. It is hence natural that they will play a role in the behaviour of expressions such as A^p as p tends to infinity. The right tool to study this is the spectral radius of a matrix, which we define as follows.

Definition 3.3.20. For a matrix $\mathrm{A} \in \mathcal{M}_n(\mathbb{C})$, the spectral radius $\rho(\mathrm{A})$ is defined as the maximum modulus of the eigenvalues of A.

Example. Consider the matrix

$$\mathrm{A} = \begin{pmatrix} -1 & 1 \\ 3 & -2 \end{pmatrix}.$$

Compute its 2-norm and its spectral radius.

Example. Consider the two matrices

$$\mathrm{A} = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathrm{B} = \begin{pmatrix} 0 & 2 \\ 0 & 1 \end{pmatrix}.$$

Then clearly $\rho(\mathrm{A}) = 1$ and $\rho(\mathrm{B}) = 1$ and $\rho(\mathrm{A}+\mathrm{B}) = 3$. Therefore the inequality $\rho(\mathrm{A}+\mathrm{B}) \leq \rho(\mathrm{A})\rho(\mathrm{B})$ does not hold, and hence the spectral radius is in general not a norm.

The following lemma highlights an important and useful property of the spectral radius. Its proof is left as an exercise.

Lemma 3.3.21. *Let $\mathrm{U} \in \mathcal{M}_n(\mathbb{C})$ be a unitary matrix (i.e. $\mathrm{U}^* = \mathrm{U}^{-1}$) and $\mathrm{A} \in \mathcal{M}_n(\mathbb{C})$. Then $\|\mathrm{UA}\|_2 = \|\mathrm{AU}\|_2 = \|\mathrm{A}\|_2$. In particular, if A is normal (i.e. $\mathrm{AA}^* = \mathrm{A}^*\mathrm{A}$), then $\rho(\mathrm{A}) = \|\mathrm{A}\|_2$.*

The spectral radius is not a norm over the Hilbert space $\mathcal{M}_n(\mathbb{C})$. However, as shown in the lemma, it is so on the restriction of $\mathcal{M}_n(\mathbb{C})$ to normal matrices, and in particular on its restriction to symmetric matrices. Furthermore the spectral radius is in general not equal to the 2-norm, but for any subordinate norm the following bound always holds:

Lemma 3.3.22. *Let $A \in \mathcal{M}_n(\mathbb{C})$ and $\|\cdot\|$ a subordinate norm. For any $k \in \mathbb{N}^*$, $\rho(\mathrm{A}) \leq \|\mathrm{A}^k\|^{1/k}$.*

Proof. Let λ be an eigenvalue of A with associated eigenvector $\mathrm{u} \neq 0$, then

$$|\lambda|^k \|\mathrm{u}\| = \|\lambda^k \mathrm{u}\| = \|A^k \mathrm{u}\| \leq \|A^k\| \|\mathrm{u}\|.$$

Therefore $|\lambda|^k \leq \|A^k\|$ and the lemma follows. □

The following theorem, proved by Gelfand [28] highlights the importance of the spectral radius, and its relationship with the asymptotic growth rate of the matrix norm.

Theorem 3.3.23. *For any matrix norm $\|\cdot\|$ and any matrix $A \in \mathcal{M}_n(\mathbb{C})$, $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$.*

Convergence analysis

We have seen above that θ-schemes have the general representation $\mathrm{C}\mathrm{u}_{n+1} = \mathrm{A}\mathrm{u}_n + \mathrm{b}_n$, where C is the identity matrix in the implicit and explicit schemes, and the vector b represents boundary conditions. Using the notation (3.3.8), let us define the matrix $\mathrm{T} \in \mathcal{M}_{m-1}(\mathbb{R})$ by $\mathrm{T} := \mathrm{T}_{m-1}(-2, 1, 1)$, and rewrite the recurrence matrix equation as

$$\mathrm{u}_{n+1} = \mathrm{M}\mathrm{u}_n + \widetilde{\mathrm{b}}_n, \tag{3.3.22}$$

where the matrix M takes the following form:

- Explicit scheme: $\mathrm{M} = \mathrm{I} + \alpha\gamma\mathrm{T}$;
- Implicit scheme: $\mathrm{M} = (\mathrm{I} - \alpha\gamma\mathrm{T})^{-1}$;
- Crank-Nicolson scheme: $\mathrm{M} = \left(\mathrm{I} - \frac{1}{2}\alpha\gamma\mathrm{T}\right)^{-1}\left(\mathrm{I} + \frac{1}{2}\alpha\gamma\mathrm{T}\right)$,
- General θ-scheme: $\mathrm{M} = \left(\mathrm{I} - \frac{\alpha\theta\sigma^2}{2}\mathrm{T}\right)^{-1}\left(\mathrm{I} + \frac{\alpha\sigma^2(1-\theta)}{2}\mathrm{T}\right)$,

where I denotes the identity matrix in $\mathcal{M}_{m-1}(\mathbb{R})$ and where the matrix $\widetilde{\mathrm{b}}_n$ of modified boundary conditions is straightforward to compute. We have used here $\gamma := \sigma^2/2$ for notational convenience. The following theorem is a matrix reformulation of the von Neumann analysis above:

Theorem 3.3.24. *If* $\|\mathrm{M}\|_2 \leq 1$*, then the scheme is convergent.*

Here the norm $\|\cdot\|_2$ represents the spectral radius of a real symmetric matrix, i.e. its largest absolute eigenvalue. This theorem can be understood with the following argument: let the vector e_0 denote a small perturbation of the initial condition u_0, and define $(\widetilde{\mathrm{u}}_n)_n$ as the perturbed solution. By (3.3.22), we can write

$$\widetilde{\mathrm{u}}_n = \mathrm{M}\widetilde{\mathrm{u}}_{n-1} + \widetilde{\mathrm{b}}_{n-1} = \mathrm{M}^2\widetilde{\mathrm{u}}_{n-2} + \mathrm{M}\widetilde{\mathrm{b}}_{n-2} + \widetilde{\mathrm{b}}_{n-1} = \ldots = \mathrm{M}^n\widetilde{\mathrm{u}}_0 + \sum_{i=0}^{n-1}\mathrm{M}^{n-1-i}\widetilde{\mathrm{b}}_i.$$

Therefore the error satisfies $\mathrm{e}_n := \mathrm{u}_n - \widetilde{\mathrm{u}}_n = \mathrm{M}^n\mathrm{e}_0$, and hence $\|\mathrm{e}_n\|_2 = \|\mathrm{M}^n\mathrm{e}_0\|_2 \leq \|\mathrm{M}^n\|_2\,\|\mathrm{e}_0\|_2$. Since we want the error to remain bounded, we need to find a constant $\overline{M} > 0$ such that $\|\mathrm{e}_n\|_2 \leq \overline{M}\,\|\mathrm{e}_0\|_2$. It is clear that this will be satisfied as soon as $\|\mathrm{M}\|_2 \leq 1$.

Exercise 25. Show that the eigenvalues of a real symmetric matrix matrix are real.

Exercise 26. For $p \in \mathbb{N}^*$, consider the matrix $\mathrm{T}_p(a, b, c) \in \mathcal{M}_p(\mathbb{R})$, where $(a, b, c) \in \mathbb{R}^3$ with $bc > 0$. Prove that it has exactly p eigenvalues $(\lambda_k)_{1\leq k\leq p}$ and that

$$\lambda_k = a + 2\sqrt{bc}\cos\left(\frac{\pi k}{p+1}\right), \qquad \text{for } k = 1, \ldots, p.$$

Solution. *Let $\lambda \in \mathbb{R}$ be an eigenvalue corresponding to the eigenvector $\mathrm{u} = (u_1, \ldots, u_m)^T \neq 0$. This implies that $\mathrm{T}_p(a, b, c)\mathrm{u} = \lambda\mathrm{u}$ and the difference equation*

$$bu_{k-1} + au_k + cu_{k+1} = \lambda u_k, \qquad \textit{for } k = 1, \ldots, m,$$

where we added the auxiliary boundary conditions $u_0 = u_{m+1} = 0$. This can be rewritten as $bu_{k-1} + (a - \lambda)u_k + cu_{k+1} = 0$. The solutions to such an equation can be expressed in terms of the roots of the characteristic polynomial $\mathcal{P} : x \mapsto cx^2 + (a - \lambda)x + b$. Let q_1 and q_2 denote the two (possibly complex) roots of $\mathcal{P}$. There exist two constant α and β such that $u_k = \alpha q_1^k + \beta q_2^k$, for any $k = 0, \ldots, m + 1$. The two boundary conditions at $k = 0$ and $k = m + 1$ imply

$$\alpha = -\beta \qquad \textit{and} \qquad \left(\frac{q_1}{q_2}\right)^{m+1} = 1.$$

We further know that $q_1q_2 = b/c$, which implies

$$q_1^{2(m+1)} = \left(\frac{b}{c}\right)^{m+1} \qquad \textit{and} \qquad q_2^{-2(m+1)} = \left(\frac{b}{c}\right)^{m+1},$$

so that the possible roots of $\mathcal{P}$ read

$$q_{1,k} = \left|\frac{b}{c}\right|^{1/2} \exp\left(\frac{\mathtt{i}\pi k}{m+1}\right) \qquad \textit{and} \qquad q_{2,k} = \left|\frac{b}{c}\right|^{1/2} \exp\left(\frac{-\mathtt{i}\pi k}{m+1}\right),$$

for $k = 0, \ldots, m$. For each $k = 0, \ldots, m$, there is one eigenvalue λ_k given by the equation $q_{1,k} + q_{2,k} = \frac{\lambda_k - a}{c}$ (sum of the roots of a polynomial of order two). This in particular implies that

$$\lambda_k = a + 2\sqrt{bc}\cos\left(\frac{\pi k}{m+1}\right), \qquad \textit{for each } k = 0, \ldots, m.$$

We can further compute the corresponding eigenvectors $\mathrm{u}_k := (\mathrm{u}_{k,1}, \ldots, \mathrm{u}_{k,m})^T$:

$$\begin{aligned} \mathrm{u}_{k,j} &= \alpha\left(q_{1,k}^j - q_{2,k}^j\right) \\ &= 2\mathtt{i}\alpha\left(\frac{b}{c}\right)^{j/2} \sin\left(\frac{\pi jk}{m+1}\right), \end{aligned}$$

for $k = 0, \ldots, m$. Note however that $\mathrm{u}_{0,\cdot} = 0$, so that λ_0 is not an eigenvalue.

Remark 3.3.25. Let $\mathrm{A} \in \mathcal{M}_m(\mathbb{R})$ and $\mathrm{x} \neq 0$ be an eigenvector of A corresponding to the eigenvalue $\lambda \in \mathbb{R}$. For any positive integer p, we can write

$$\mathrm{A}^p\mathrm{x} = \mathrm{A}^{p-1}(\mathrm{Ax}) = \lambda\mathrm{A}^{p-1}\mathrm{x} = \ldots = \lambda^p\mathrm{x},$$

so that λ^p is an eigenvalue (with corresponding eigenvector x) of the matrix A^p. For two polynomials $\mathcal{P}_1$ and $\mathcal{P}_2$, a similar argument shows that $\mathcal{P}_1(\lambda)/\mathcal{P}_2(\lambda)$ is an eigenvalue of $\mathcal{P}_1(\mathrm{A})/\mathcal{P}_2(\mathrm{A})$.

In the θ-schemes above, we can apply Exercise 26 to compute the m eigenvalues of the tridiagonal matrix $\mathrm{T}_{m-1}(-2, 1, 1)$ as

$$\lambda_k^{\mathrm{T}} = -4\sin\left(\frac{k\pi}{2m}\right)^2, \qquad \text{for } k = 1, \ldots, m-1.$$

The eigenvalues of the transition matrix M in the implicit scheme then follow directly from Remark 3.3.25, and we obtain

$$\lambda_k^{\mathrm{M}} = \left(1 + 4\alpha\gamma \sin\left(\frac{k\pi}{2m}\right)^2\right)^{-1}, \qquad \text{for } k = 1, \ldots, m-1.$$

Since the $\|\cdot\|_2$ norm of a normal matrix is equal to its spectral radius, i.e. the largest absolute eigenvalue, we obtain

$$\|\mathrm{M}\|_2 = \max_{k=1,\ldots,m-1} \left| \left(1 + 4\alpha\gamma \sin\left(\frac{k\pi}{2m}\right)^2\right)^{-1} \right| < 1,$$

for any $\alpha > 0$. The implicit scheme is thus unconditionally stable, consistent and hence (by Theorem 3.3.13) convergent.

In the Crank-Nicolson scheme, the eigenvalues of the matrix M read

$$\lambda_k^{\mathrm{M}} = \frac{1 - 2\alpha\gamma \sin\left(\frac{k\pi}{2m}\right)^2}{1 + 2\alpha\gamma \sin\left(\frac{k\pi}{2m}\right)^2},$$

for $k = 1, \ldots, m-1$. Therefore

$$\|\mathrm{M}\|_2 = \max_{k=1,\ldots,m-1} \left|\lambda_k^{\mathrm{M}}\right| < 1, \qquad \text{for any } \alpha > 0.$$

The Crank-Nicolson scheme is therefore unconditionally stable, consistent and hence convergent.

Exercise 27. Perform such an analysis for the explicit scheme and discuss its stability.

Solution. *Using a similar analysis as for the implicit scheme above, we see that the eigenvalues of the transition matrix* M *in the explicit scheme read*

$$\lambda_k^{\mathrm{M}} = 1 - 4\alpha\gamma \sin\left(\frac{k\pi}{2m}\right)^2, \qquad \textit{for } k = 0, \ldots, m-1.$$

Remark 3.3.26. We could have worked from the beginning with general θ schemes, with $\theta \in [0,1]$. In that case, a similar analysis of the eigenvalues of the transition matrix M shows that the scheme is unconditionally stable as soon as $\theta \in [1/2, 1]$.

Exercise 28. Which condition on the parameters α and γ ensures that θ-schemes are convergent when $\theta \in [0, 1/2)$?

To finish with the matrix convergence analysis, let us mention and prove two results, which provide easy-to-check conditions on the modulus of the eigenvalues of the iteration matrix.

Theorem 3.3.27 (Gerschgorin's Theorem)**.** *The modulus of the largest eigenvalue of a square matrix is always smaller or equal to the largest sum of the moduli of the terms along any row or column.*

Proof. Let $\mathrm{A} = (a_{ij})_{1\leq i,j\leq n}$ denote such a matrix and $(\lambda_k, \mathrm{u}^k)_{1\leq k\leq n}$ its eigenvalues and associated eigenvectors: $\mathrm{A}\mathrm{u}^k = \lambda_k \mathrm{u}^k$ for all $k = 1,\ldots,n$. Fix one $k \in \{1,\ldots,n\}$ and let $l := \max\{j \in \{1,\ldots,n\} : |\mathrm{u}_l^k| \geq \mathrm{u}_j^k\}$. From the eigenvalue equation, we can write

$$\lambda_k = \left(\mathrm{u}_l^k\right)^{-1} \sum_{j=1}^n a_{k,j}\mathrm{u}_j^k,$$

so that $|\lambda_k| \leq \sum_{j=1}^n |a_{k,j}|$, and the theorem follows since the eigenvalues of the transposed matrix are the same. □

Theorem 3.3.28 (Brauer's Theorem)**.** *Let* $\mathrm{A} \in \mathcal{M}_n(\mathbb{C})$ *with eigenvalues* $\lambda_1,\ldots,\lambda_n$*. Then, for any* $i \in \{1,\ldots,n\}$, $\Pi_i := \sum_{1\leq j\leq n, j\neq i} |a_{i,j}| \geq |\lambda_i - a_{i,i}|$.

Proof. The proof is immediate in light of the proof of Theorem 3.3.27. □

Exercise 29. Consider the Crank-Nicolson equations, for $j = 0,\ldots,n-1$: $(2\mathrm{I} - \alpha\mathrm{T}_{n-1})\,\mathrm{u}_{j+1} = (2\mathrm{I} - \alpha\mathrm{T}_{n-1})\,\mathrm{u}_j$, which we can rewrite as $\mathrm{B}\mathrm{u}_{j+1} = (4\mathrm{I} - \mathrm{B})\mathrm{u}_j$, or $\mathrm{u}_{j+1} = (4\mathrm{B}^{-1} - \mathrm{I})\mathrm{u}_j$, where $\mathrm{B} = \mathrm{T}_n((2(1+\alpha), -\alpha, -\alpha)$. We know that the system is stable if and only if the modulus of all eigenvalues of $(4\mathrm{B}^{-1} - \mathrm{I})$ is less than one, i.e. $|4/\lambda - 1| \leq 1$, or $\lambda \geq 2$, where λ is an eigenvalue of B. We can also apply Theorem 3.3.28 directly, noting that $a_{ii} = 2(1+\alpha)$ and $\max_i \Pi_i = 2\alpha$, so that Brauer's Theorem yields $|\lambda - 2(1+\alpha)| \leq 2\alpha$, or $\lambda \in [2, 2+4\alpha]$, showing that the scheme is unconditionally stable, for any (non-negative) value of α.

3.4 PDEs for path-dependent options

3.4.1 The American case: Problem class

3.4.2 The Asian case

The payoff of an Asian option written on the underlying S, with maturity T, depends on the average of the asset price over the whole life of the product. This average can be either continuous or discrete, namely

$$\frac{1}{T}\int_0^T S_t \mathrm{d}t \qquad \text{or} \qquad \frac{1}{n}\sum_{i=1}^n S_{t_i},$$

for some partition $0 < t_1 < \cdots < t_n = T$. We assume for simplicity that the underlying stock price evolves according to the Black-Scholes dynamics:

$$\mathrm{d}S_t = S_t\left(r\mathrm{d}t + \sigma\mathrm{d}W_t\right), \qquad S_0 > 0,$$

and we shall be interested in deriving a PDE to evaluate a continuously monitored Asian Call option with strike K, i.e. an option with the following payoff at maturity:

$$V_T := \left(\frac{1}{T}\int_0^T S_t\mathrm{d}t - K\right)_+.$$

By risk-neutral expectation, the price at time $t \in [0, T]$ is given by

$$V_t = \mathbb{E}\left(\mathrm{e}^{-r(T-t)} V_T | \mathcal{F}_t\right).$$

The discounted option price $(\mathrm{e}^{-rt} V_t)_{t \in [0,T]}$ is clearly a martingale; however, it is not Markovian, since it does not only depends on T, but on the whole trajectory. We therefore augment the state space with the process $I_t := \int_0^t S_u \mathrm{d}u$, which clearly satisfies the stochastic differential equation $\mathrm{d}I_t = S_t \mathrm{d}t$, with starting value $I_0 = 0$. Now, the couple (S, Y) forms a Markov process, and the value of the option depends only on its terminal value, making Feynman-Kac theorem amenable to such a problem: the function $u : [0, T] \times \mathbb{R}_+^2$ defined by

$$u(t, x, y) := \mathbb{E}\left(\mathrm{e}^{-r(T-t)} \left(\frac{1}{T} I_T - K\right)_+ | S_t = x, I_t = y\right)$$

satisfies the following:

Theorem 3.4.1. *The function u satisfies the following partial differential equation:*

$$\left(\partial_t + rx\partial_x + x\partial_y + \frac{1}{2}\sigma^2 x^2 \partial_{xx}\right) u(t, x, y) = ru(t, x, y), \quad \textit{for all } (t, x, y) \in [0, T) \times \mathbb{R}_+ \times \mathbb{R},$$

with boundary conditions

$$\begin{array}{rcll} u(t, 0, y) & = & \mathrm{e}^{-r(T-t)} \left(\frac{y}{T} - K\right)_+, & \textit{for } (t, y) \in [0, T) \times \mathbb{R}, \\ \lim\limits_{y \downarrow -\infty} u(t, x, y) & = & 0, & \textit{for } (t, y) \in [0, T) \times \mathbb{R}_+, \\ u(T, x, y) & = & \left(\frac{y}{T} - K\right)_+, & \textit{for } (x, y) \in \mathbb{R}_+ \times \mathbb{R}. \end{array}$$

Proof. Exercise. □

3.5 Solving general second-order linear parabolic partial differential equations

We have so far concentrated our efforts in solving the heat equation, which enabled us to solve the Black-Scholes equation from the set of transformations developed in Section 3.1.2. However the latter are not always available. In view of the Feynmac-Kac theorem (Theorem 3.2.2), one has then to solve a general second-order linear parabolic partial differential equation of the form

$$\partial_t u = \mathcal{L}u \qquad \text{where} \qquad \mathcal{L}u := \mu(x, t)\partial_x u + a(x, t)\partial_{xx} u + c(x, t)u, \tag{3.5.1}$$

with some appropriate boundary conditions, and where the function $a(\cdot, \cdot)$ is strictly positive (equal to $\frac{1}{2}\sigma^2(\cdot, \cdot)$ in the heat equation above). As opposed to the heat equation, we do have here a term in ∂_x. In order to be consistent with the $\mathcal{O}(\delta_x^2)$ order of accuracy of the scheme, central finite differences for this term will be necessary.

Remark 3.5.1. Note that if the boundary condition is of the form $u(T, x) = x$, then the solution to the partial differential equation above is trivially $u(t, x) = x$. Financially speaking, this simply means that the price today of an option that pays out the final value of the stock price at maturity is necessarily equal to the stock price at inception. If the boundary condition reads $u(T, x) = c \in \mathbb{R}$, then the value of the option at any time $t \in [0, T]$ is clearly equal to $c\mathrm{e}^{-r(T-t)}$.

Let us first state some existence results (recall that $\overline{D}$ represents the closure of a set D).

Theorem 3.5.2 (Maximum principle for parabolic equations)**.** *Let D and Γ be some subsets of $\mathbb{R}\times[0, T]$ for some $T > 0$. Assume that the coefficients of* (3.5.1) *are continuous and that $a(x, t) > 0$ for all $(x, t) \in D$. If*

(i) $\mathcal{L}u \le 0$ on $\overline{D} \setminus \Gamma$;

(ii) μ is bounded by some constant on $\overline{D} \setminus \Gamma$;

(iii) $u(x, t) \ge 0$ for all $(x, t) \in \Gamma$.

Then $u(x, t) \ge 0$ on $\overline{D}$.

This theorem is fundamental since it allows us to determine the properties of solutions of parabolic differential equations. In particular it tells us that the solution to the Black-Scholes equation is necessarily positive.

3.5.1 Applications to θ-schemes

Concerning the implicit finite difference scheme, the discretisation of (3.5.1) gives

$$u_{i,j} = -\left(\alpha a_{i+1,j} - \frac{\beta}{2}\mu_{i+1,j}\right) u_{i+1,j-1} + (1 + 2\alpha a_{i+1,j} - c_{i+1,j}\delta_T)\, u_{i+1,j} - \left(\alpha a_{i+1,j} + \frac{\beta}{2}\mu_{i+1,j}\right) u_{i+1,j+1},$$

for any $i = 0, \ldots, n-1$, $j = 1, \ldots, m-1$, where $\alpha := \delta_T/\delta_x^2$ and $\beta := \delta_T/\delta_x$. The notation $a_{i,j}$ denotes as usual the value of the function a evaluated at the node $(i\delta_T, x_L + j\delta_x)$, and likewise for the other functions. In matrix notations, the problem reduces to solving the equation $\mathrm{A}_{i+1}u_{i+1} = u_i + \mathrm{b}_i$ for $i = 1, \ldots, n-1$, where (recall that T stands for the tridiagonal matrix notation (3.3.8))

$$\begin{aligned}
\mathrm{A}_i &= \mathrm{T}_{m-1}\left(\frac{\beta}{2}\mu_{i,j} - \alpha a_{i,j}, 1 + 2\alpha a_{i,j} - \delta_T c_{i,j}, -\left(\alpha a_{i,j} + \frac{\beta}{2}\mu_{i,j}\right)\right) \in \mathcal{M}_{m-1}(\mathbb{R}),\\
\mathrm{b}_i &= \left(\left(\alpha a_{i,1} - \frac{\beta}{2}\mu_{i,1}\right) u_{i,0}, 0, \ldots, 0, \left(\frac{\beta}{2}\mu_{i,m-1} + \alpha a_{i,m-1}\right) u_{i,m}\right)^T \in \mathbb{R}^{m-1}.
\end{aligned}$$

Exercise 30. Write the discretisation of (3.5.1) in the explicit scheme in matrix form.

In the Crank-Nicolson discretisation scheme, we can perform a similar finite-difference scheme and after some tedious but straightforward algebra, we obtain the matrix equation $\mathrm{C}u_{i+1} = \mathrm{D}u_i +$

b_i, for $i = 0, \dots, n-1$, where

$$\begin{aligned}
\mathrm{C} &= \mathrm{T}\left(\frac{\beta}{4}\widetilde{\mu}_{ij} - \frac{\alpha}{2}\widetilde{a}_{ij}, 1 + \alpha\widetilde{a}_{ij} - \frac{\delta_T}{2}\widetilde{c}_{ij}, -\left(\frac{\alpha}{2}\widetilde{a}_{ij} + \frac{\beta}{4}\widetilde{\mu}_{ij}\right)\right),\\
\mathrm{D} &= \mathrm{T}\left(\frac{\alpha}{2}\widetilde{a}_{ij} - \frac{\beta}{4}\widetilde{\mu}_{ij}, 1 - \alpha\widetilde{a}_{ij} + \frac{\delta_T}{2}\widetilde{c}_{ij}, \frac{\alpha}{2}\widetilde{a}_{ij} + \frac{\beta}{4}\widetilde{\mu}_{ij}\right),\\
\mathrm{b}_i &= \begin{pmatrix}
\left(\frac{\alpha}{2}\widetilde{a}_{i,0} - \frac{\beta}{4}\widetilde{\mu}_{i,0}\right) u_{i+1,0} + \left(\frac{\alpha}{2}\widetilde{a}_{i,0} - \frac{\beta}{4}\widetilde{\mu}_{i,0}\right) u_{i,0}\\
0\\
\vdots\\
0\\
\left(\frac{\alpha}{2}\widetilde{a}_{i,m-1} + \frac{\beta}{4}\widetilde{\mu}_{i,m-1}\right) u_{i+1,m} - \left(\frac{\alpha}{2}\widetilde{a}_{i,m-1} + \frac{\beta}{4}\widetilde{\mu}_{i,m-1}\right) u_{i,m}
\end{pmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{a}_{ij} &:= a\left(i\delta_T + \frac{1}{2}\delta_T, x_L + j\delta_x\right),\\
\widetilde{\mu}_{ij} &:= \mu\left(i\delta_T + \frac{1}{2}\delta_T, x_L + j\delta_x\right),\\
\widetilde{c}_{ij} &:= c\left(i\delta_T + \frac{1}{2}\delta_T, x_L + j\delta_x\right).
\end{aligned}$$

3.6 Two-dimensional PDEs

The partial differential equations we have studied so far were one-dimensional (in space). This came from the fact that we were looking at financial derivatives written on a single stock price, as in the one-dimensional Black-Scholes model. Many financial derivatives are actually written on several assets—for instance basket options—and hence the methods above have to be extended to higher dimensions. Even in the case of a single asset, higher-dimensional PDEs can be needed, for instance in the case of stochastic volatility, stochastic interest rates. We shall focus here on the heat equation in two dimensions, which provide us with the canonical model to study such a feature. Let us consider the two-dimensional (in space) partial differential equation

$$\partial_\tau u = b_1 \partial_{xx} u + b_2 \partial_{yy} u, \tag{3.6.1}$$

on a square. From Theorem 3.1.4, we know that this PDE is parabolic if $b_1 > 0$ and $b_2 > 0$, which we assume from now on. In particular when $b_1 = b_2$, the PDE (3.6.1) precisely corresponds to the two-dimensional heat equation. Let us now see how this comes into the financial modelling picture. In Section 3.1.2, we saw how to reduce the Black-Scholes differential equation to the heat equation in one dimension. The two-dimensional Black-Scholes model for the pair $(S_1(t), S_2(t))_{t\geq 0}$ reads

$$\begin{aligned}
S_1(t) &= S_1(0)\exp\left(\left(r - \frac{1}{2}\sigma_1^2\right)t + \sigma_1\sqrt{t}\left(\rho Z + \sqrt{1-\rho^2}W\right)\right),\\
S_2(t) &= S_2(0)\exp\left(\left(r - \frac{1}{2}\sigma_2^2\right)t + \sigma_2\sqrt{t}Z\right),
\end{aligned}$$

where W and Z are two independent Gaussian random variables with zero mean and unit variance. The two volatilities σ_1 and σ_2 are strictly positive, the risk-free interest rate r is non negative and the correlation parameter ρ lies in $(-1, 1)$. In two dimensions (in the space variable), one can show that the Black-Scholes differential equation reads

$$\partial_t V + \mathcal{L}V = 0, \tag{3.6.2}$$

where

$$\mathcal{L} := rS_1\partial_{S_1} + rS_2\partial_{S_2} + \frac{1}{2}\sigma_1^2 S_1^2 \partial_{S_1S_1} + \frac{1}{2}\sigma_2^2 S_2^2 \partial_{S_2S_2} + \rho\sigma_1\sigma_2 S_1S_2\partial_{S_1S_2} - r.$$

For clarity, we do not mention the boundary conditions here, but it is clear that they are fundamental in establishing a unique solution consistent with the pricing problem. Let us consider this equation on a logarithmic scale, i.e. $x_1 := \log(S_1)$ and $x_2 := \log(S_2)$, and define $\nu_1 := r - \frac{1}{2}\sigma_1^2$ and $\nu_2 := r - \frac{1}{2}\sigma_2^2$. The PDE (3.6.2) reduces to

$$\partial_t V + \nu_1\partial_{x_1}V + \nu_2\partial_{x_2}V + \frac{1}{2}\sigma_1^2\partial_{x_1x_1} + \frac{1}{2}\sigma_2^2\partial_{x_2x_2} + \rho\sigma_1\sigma_2\partial_{x_1x_2} - r = 0. \tag{3.6.3}$$

In order to simplify this equation further, we need to remove the mixed second-order derivative. Let us first recall some elementary facts from linear algebra.

Theorem 3.6.1 (Spectral theorem)**.** *Let* $\mathrm{A} \in \mathcal{M}_n(\mathbb{R})$*. If the matrix* A *has* n *linearly independent eigenvectors* $(\mathrm{u}_1, \ldots, \mathrm{u}_n)$ *(i.e. there exists* $(\lambda_1, \ldots, \lambda_n) \neq 0$ *such that* $\mathrm{A}\mathrm{u}_i = \lambda_i \mathrm{u}_i$ *for any* $i = 1, \ldots, n$*), then the decomposition* $\mathrm{A} = \mathrm{U}\Lambda\mathrm{U}^{-1}$ *holds where each column of* U *is an eigenvector and where the matrix* Λ *is diagonal with* $\Lambda_{ii} = \lambda_i$*.*

Remark 3.6.2. In particular, when the matrix A is real and symmetric, the eigenmatrix U is orthogonal, i.e. $\mathrm{U}^{-1} = \mathrm{U}^T$, and hence $\mathrm{A} = \mathrm{U}\Lambda\mathrm{U}^T$.

Proposition 3.6.3. *Let* $\mathrm{A} \in \mathcal{M}_2(\mathbb{R})$*. Then* A *has at most two eigenvalues* λ_- *and* λ_+ *and*

$$\lambda_\pm = \frac{1}{2}\left(\mathrm{Tr}(\mathrm{A}) \pm (\mathrm{Tr}(\mathrm{A}) - 4\mathrm{det}(\mathrm{A}))^{1/2}\right).$$

The proof is left as an exercise. Consider now the covariance matrix related to the PDE (3.6.3):

$$\Sigma := \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

From the proposition above, we can compute explicitly its two eigenvalues λ_- an λ_+ (which we respectively rename λ_1 and λ_2 for convenience). Denote by u_1 and u_2 the corresponding eigenvectors and $\mathrm{U} := (\mathrm{u}_1\mathrm{u}_2)$ as in the spectral theorem. Consider finally the change of variables $(y_1, y_2)^T = \mathrm{U}(x_1, x_2)^T$. The partial differential equation (3.6.3) becomes

$$\partial_t V + \alpha_1\partial_{y_1}V + \alpha_2\partial_{y_2}V + \frac{\lambda_1}{2}\partial_{y_1y_1}V + \frac{\lambda_2}{2}\partial_{y_2y_2}V - rV = 0,$$

where $(\alpha_1, \alpha_2)^T = \mathrm{U}(\nu_1, \nu_2)^T$. As in the one-dimensional case, let us define the transformation

$$V(y_1, y_2, t) := \mathrm{e}^{\beta_1 y_1 + \beta_2 y_2 + \beta_3 t}\Psi(y_1, y_2, t).$$

Upon choosing $\beta_1 := -\alpha_1/\lambda_1$, $\beta_2 := -\alpha_2/\lambda_2$ and $\beta_3 := \frac{\alpha_1^2}{2\lambda_1} + \frac{\alpha_2^2}{2\lambda_2} + r$, we finally obtain

$$\partial_t\Psi + \frac{\lambda_1}{2}\partial_{y_1y_1}\Psi + \frac{\lambda_2}{2}\partial_{y_2y_2}\Psi = 0.$$

We can make a last change of variable $(z_1, z_2) := (y_1, y_2\sqrt{\lambda_1/\lambda_2})$ in order to obtain the standard heat equation in two dimensions:

$$\partial_t\Psi + \frac{\lambda_1}{2}\left(\partial_{z_1z_1} + \partial_{z_2z_2}\right)\Psi = 0.$$

3.6.1 θ-schemes for the two-dimensional heat equation

By a change of time $t \mapsto T-t$, where $T > 0$ is the time boundary of the problem, the heat equation boils down to (modulo some constant factor)

$$\partial_t u = \gamma^2\left(\partial_{xx}u + \partial_{yy}u\right), \tag{3.6.4}$$

and we are interested in solving it for $(x, y, t) \in [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}] \times [0, T]$. We specify now the following boundary conditions:

$$\begin{aligned}
u(x, y, 0) &= u_0(x, y), \text{ for any } (x, y) \in [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}],\\
u(a, y, t) &= f_a(y, t), \text{ for any } (y, t) \in [\underline{y}, \overline{y}] \times [0, T],\\
u(b, y, t) &= f_b(y, t), \text{ for any } (y, t) \in [\underline{y}, \overline{y}] \times [0, T],\\
u(x, c, t) &= f_c(y, t), \text{ for any } (x, t) \in [\underline{x}, \overline{x}] \times [0, T],\\
u(x, d, t) &= f_d(y, t), \text{ for any } (x, t) \in [\underline{x}, \overline{x}] \times [0, T].
\end{aligned}$$

The first boundary condition corresponds to the payoff at maturity, whereas the other boundary conditions account for possible knock-out barriers. The functions $f_{\cdot}$ are assumed to be smooth. For $(n_x, n_y, n_T) \in \mathbb{N}^3$, consider the discretisation steps $\delta_x > 0$, $\delta_y > 0$ and $\delta_T > 0$ defined by

$$\delta_x := \frac{\overline{x} - \underline{x}}{n_x}, \qquad \delta_y := \frac{\overline{y} - \underline{y}}{n_y}, \qquad \delta_T := \frac{T}{n_T}.$$

At some node $(i, j, k) \in [0, n_x] \times [0, n_y] \times [0, n_T]$ (in Cartesian coordinates: $(a + i\delta_x, c + j\delta_y, k\delta_T)$), we use the notation $u_{i,j}^k$ for the function u evaluated at this point.

Explicit scheme

At the node (i, j, k), approximating the time-derivative of the function u using a forward difference scheme $\partial_t u|_{(i,j,k)} = \delta_T^{-1}\left(u_{i,j}^{k+1} - u_{i,j}^k\right) + \mathcal{O}(\delta_T)$, the heat equation (3.6.4) is approximated by

$$u_{i,j}^{k+1} = \Big(1 - 2\left(\alpha_x + \alpha_y\right)\Big)u_{i,j}^k + \alpha_x\left(u_{i+1,j}^k + u_{i-1,j}^k\right) + \alpha_y\left(u_{i,j+1}^k + u_{i,j-1}^k\right),$$

for any $i = 1, \ldots, n_x - 1$, $j = 1, \ldots, n_y - 1$, $k = 0, \ldots, n_T - 1$, and where we define $\alpha_x := \gamma^2\delta_T/\delta_x^2$ and $\alpha_y := \gamma^2\delta_T/\delta_y^2$.

Implicit scheme

At the node (i, j, k), using a backward difference scheme $\partial_t u|_{(i,j,k)} = \delta_T^{-1}\left(u_{i,j}^k - u_{i,j}^{k-1}\right) + \mathcal{O}(\delta_T)$ for the time-derivative, the heat equation (3.6.4) is approximated by

$$\Big(1 + 2\left(\alpha_x + \alpha_y\right)\Big)u_{i,j}^k - \alpha_x\left(u_{i+1,j}^k + u_{i-1,j}^k\right) - \alpha_y\left(u_{i,j+1}^k + u_{i,j-1}^k\right) = u_{i,j}^{k-1},$$

for any $i = 1, \ldots, n_x - 1$, $j = 1, \ldots, n_y - 1$, $k = 0, \ldots, n_T - 1$, and where α_x and α_y are defined as in the explicit scheme.

Crank-Nicolson

If we apply the Crank-Nicolson scheme to the two-dimensional heat equation, we obtain

$$\begin{aligned}\frac{u_{i,j}^{k+1} - u_{i,j}^k}{\delta_T} &= \frac{\gamma^2}{2}\frac{u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}}{\delta_x^2} + \frac{\gamma^2}{2}\frac{u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1}}{\delta_x^2}\\ &+ \frac{\gamma^2}{2}\frac{u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}}{\delta_y^2} + \frac{\gamma^2}{2}\frac{u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1}}{\delta_y^2}.\end{aligned}$$

Using a similar Fourier analysis as above, one can further show that the Crank-Nicolson scheme in two (space) dimensions remains unconditionally stable. Consider the following vector:

$$\mathrm{U}^k := \left(u_{11}^k, \ldots, u_{n_x-1,1}, u_{12}^k, \ldots, u_{n_x-1,2}, \ldots, u_{1,n_y-1}^k, \ldots, u_{n_x-1,n_y-1}\right)^T \in \mathbb{R}^{(n_x-1)(n_y-1)}.$$

The Crank-Nicolson scheme can be written in matrix form as follows

$$\left(\mathrm{I} + \frac{1}{2}\mathrm{C}\right)\mathrm{U}^{k+1} = \left(\mathrm{I} - \frac{1}{2}\mathrm{C}\right)\mathrm{U}^k + \mathrm{b}_k,$$

where the vector b_k represents the boundary conditions and where

$$\begin{aligned}\mathrm{D}_x &:= \mathrm{T}\left(a, -\delta_x, -\delta_x\right) \in \mathcal{M}_{n_x-1}(\mathbb{R}),\\ \mathrm{D}_y &:= -\delta_y\mathrm{I} \in \mathcal{M}_{n_y-1}(\mathbb{R}),\\ \mathrm{C} &:= \mathrm{T}\left(\mathrm{D}_x, \mathrm{D}_y, \mathrm{D}_y\right) \in \mathcal{M}_{(n_x-1)^2(n_y-1)^2}(\mathbb{R}),\\ a &:= 2\gamma^2\delta_T\left(\delta_x^2 + \delta_y^2\right).\end{aligned}$$

We may solve these matrix equations as in the one-dimensional case. However, the matrix to invert is now block-tridiagonal and its inversion is computer intensive and often not tractable for practical purposes. We therefore search for an alternative method, tailored for this multidimensional problem, in particular preserving the tridiagonal structure of the matrix.

3.6.2 The ADI method

Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be linear operators, and assume that we are able to solve the equations $\partial_t u = \mathcal{L}_1 u$ and $\partial_t u = \mathcal{L}_2 u$. Taking $\mathcal{L}_1 = b_1\partial_{xx}$ and $\mathcal{L}_2 = b_2\partial_{yy}$, the equation (3.6.1) becomes $\partial_t u = \mathcal{L}_1 u + \mathcal{L}_2 u$.

We now use a central difference scheme around the point $\left(k+\frac{1}{2}\right)\delta_T$, i.e. perform a Taylor series around this point, and average the central differences in space, so that (3.6.1) becomes

$$\frac{u^{k+1}-u^k}{\delta_T} = \frac{1}{2}\left(\mathcal{L}_1 u^{k+1} + \mathcal{L}_1 u^k\right) + \frac{1}{2}\left(\mathcal{L}_2 u^{k+1} + \mathcal{L}_2 u\right) + \mathcal{O}\left(\delta_T^2\right),$$

which we can rewrite as

$$\left(\mathcal{I} - \frac{\delta_T}{2}\left(\mathcal{L}_1 + \mathcal{L}_2\right)\right) u^{k+1} = \left(\mathcal{I} + \frac{\delta_T}{2}\left(\mathcal{L}_1 + \mathcal{L}_2\right)\right) u^k + \mathcal{O}\left(\delta_T^3\right). \tag{3.6.5}$$

Applying a central-difference scheme for the space variables as in the Crank-Nicolson scheme will eventually lead to the computation of the inverse of the matrix version of the left-hand side, which is not an easy task. However, using the identities

$$\begin{aligned}
(1+z_1)(1+z_2) &= 1 + z_1 + z_2 + z_1 z_2,\\
(1-z_1)(1-z_2) &= 1 - z_1 - z_2 + z_1 z_2,
\end{aligned}$$

we can turn (3.6.5) into (take $z_1 = -\frac{1}{2}\delta_T\mathcal{L}_1$ and $z_2 = -\frac{1}{2}\delta_T\mathcal{L}_2$)

$$\left(\mathcal{I} - \frac{\delta_T}{2}\mathcal{L}_1\right)\left(\mathcal{I} - \frac{\delta_T}{2}\mathcal{L}_2\right) u^{k+1} = \left(\mathcal{I} + \frac{\delta_T}{2}\mathcal{L}_1\right)\left(\mathcal{I} + \frac{\delta_T}{2}\mathcal{L}_2\right) u^k + \frac{\delta_T^2}{4}\mathcal{L}_1\mathcal{L}_2\left(u^{k+1} - u^k\right) + \mathcal{O}\left(\delta_T^3\right).$$

Now, the last two terms on the right-hand side are of order $\mathcal{O}\left(\delta_T^3\right)$, so that this simplifies to

$$\left(\mathcal{I} - \frac{\delta_T}{2}\mathcal{L}_1\right)\left(\mathcal{I} - \frac{\delta_T}{2}\mathcal{L}_2\right) u^{k+1} = \left(\mathcal{I} + \frac{\delta_T}{2}\mathcal{L}_1\right)\left(\mathcal{I} + \frac{\delta_T}{2}\mathcal{L}_2\right) u^k + \mathcal{O}\left(\delta_T^3\right).$$

Let us now use a Crank-Nicolson central difference scheme for the space variables, i.e. let the matrices L_1 and L_2 be the second-order approximations of the operators $\mathcal{L}_1$ and $\mathcal{L}_2$, and u^k the vector of solutions at time $k\delta_T$. We obtain

$$\left(\mathrm{I} - \frac{\delta_T}{2}L_1\right)\left(\mathrm{I} - \frac{\delta_T}{2}L_2\right)\mathrm{u}^{k+1} = \left(\mathrm{I} + \frac{\delta_T}{2}L_1\right)\left(\mathrm{I} + \frac{\delta_T}{2}L_2\right)\mathrm{u}^k, \tag{3.6.6}$$

where we have ignored the terms of order $\mathcal{O}\left(\delta_T^3\right)$, $\mathcal{O}\left(\delta_T\delta_x^2\right)$ and $\mathcal{O}\left(\delta_T\delta_y^2\right)$. Several schemes now exist to solve such an equation. Peaceman and Rachford [51] splits (3.6.6) as follows

$$\begin{aligned}
\left(\mathrm{I} - \frac{1}{2}\delta_T L_1\right)\widetilde{\mathrm{u}}^{k+1/2} &= \left(\mathrm{I} + \frac{1}{2}\delta_T L_2\right)\mathrm{u}^k,\\
\left(\mathrm{I} - \frac{1}{2}\delta_T L_2\right)\mathrm{u}^{k+1} &= \left(\mathrm{I} + \frac{1}{2}\delta_T L_1\right)\widetilde{\mathrm{u}}^{k+1/2},
\end{aligned} \tag{3.6.7}$$

where the notation $\widetilde{u}$ expresses the fact that this is not an approximation of the function u at the time $(k+1/2)\delta_T$ but only an auxiliary quantity. This amounts to introducing an intermediate time step—i.e. splitting the interval $[t, t+k\delta_T]$ into $[t, t+k\delta_T/2]$ and $[t+k\delta_T/2, t+\delta_T]$—and to using an implicit scheme on one subinterval and an explicit scheme on the other subinterval. This set of two equations is furthermore equivalent to the original matrix equation (3.6.6), where we use the fact that, for $i = 1, 2$,

$$\left(\mathrm{I} - \frac{1}{2}\delta_T L_i\right)\left(\mathrm{I} + \frac{1}{2}\delta_T L_i\right) = \left(\mathrm{I} + \frac{1}{2}\delta_T L_i\right)\left(\mathrm{I} - \frac{1}{2}\delta_T L_i\right).$$

As usual, we need to specify (space) boundary conditions for this split scheme at the intermediate time point $t + k\delta_T/2$. These are simply obtained using the (space) boundary conditions at the times t and $t + k\delta_T$ and plugging them into (3.6.7).

Let us now reconsider the PDE $\partial_t u = \mathcal{L}_1 u + \mathcal{L}_2 u$, and use a backward-in-time scheme:

$$\left(\mathcal{I} - \delta_T\mathcal{L}_1 - \delta_T\mathcal{L}_2\right) u^{k+1} = u^k + \mathcal{O}(\delta_T^2),$$

where we use $\mathcal{I}$ as the identity operator. We can rewrite this as

$$\left(\mathcal{I} - \delta_T\mathcal{L}_1 - \delta_T\mathcal{L}_2 + \delta_T^2\mathcal{L}_1\mathcal{L}_2\right) u^{k+1} = \left(\mathcal{I} + \delta_T^2\mathcal{L}_1\mathcal{L}_2\right) u^k + \delta_T^2\mathcal{L}_1\mathcal{L}_2\left(u^{k+1} - u^k\right) + \mathcal{O}(\delta_T^2),$$

which implies that

$$\left(\mathrm{I} - \delta_T L_1\right)\left(\mathrm{I} - \delta_T L_2\right)\mathrm{u}^{k+1} = \left(\mathrm{I} + \delta_T^2 L_1 L_2\right)\mathrm{u}^k,$$

where the matrices L_1 and L_2 are defined as above, I is the identity matrix, and we have again ignored the higher-order terms. The Douglas-Rachford method [19] reads

$$\begin{aligned}\left(\mathrm{I} - \delta_T L_1\right)\widetilde{\mathrm{u}}^{k+1/2} &= \left(\mathrm{I} + \delta_T L_2\right)\mathrm{u}^k,\\ \left(\mathrm{I} - \delta_T L_2\right)\mathrm{u}^{k+1} &= \widetilde{\mathrm{u}}^{k+1/2} - \delta_T L_2\mathrm{u}^k.\end{aligned}$$

As we mentioned in the one-dimensional case, the stability of the scheme is of fundamental importance. In the case of the Douglas-Rachford scheme, we can apply the same Fourier-transform approach and set $u_{i,j}^k \to g^k \mathrm{e}^{\mathrm{i}i\delta_x}\mathrm{e}^{\mathrm{i}j\delta_y}$ and $\widetilde{u}_{i,j}^{k+1/2} \to \widetilde{g}g^k\mathrm{e}^{\mathrm{i}i\delta_x}\mathrm{e}^{\mathrm{i}j\delta_y}$. We therefore obtain

$$\begin{aligned}\left(1 + 4b_1\delta_{x,T}\sin\left(\frac{\delta_x}{2}\right)^2\right)\widetilde{g} &= 1 - 4b_2\delta_{y,T}\sin\left(\frac{\delta_y}{2}\right)^2,\\ \left(1 + 4b_2\delta_{y,T}\sin\left(\frac{\delta_y}{2}\right)^2\right)g &= \widetilde{g} + 4b_2\delta_{y,T}\sin\left(\frac{\delta_y}{2}\right)^2,\end{aligned}$$

where $\delta_{x,T} := \delta_T/\delta_x^2$ and $\delta_{y,T} := \delta_T/\delta_y^2$. Therefore the amplification factor g reads

$$g = \frac{1 + 16b_1b_2\delta_{x,T}\delta_{y,T}\sin\left(\frac{\delta_x}{2}\right)^2\sin\left(\frac{\delta_y}{2}\right)^2}{\left(1 + 4b_1\delta_{x,T}\sin\left(\frac{\delta_x}{2}\right)^2\right)\left(1 + 4b_2\delta_{y,T}\sin\left(\frac{\delta_y}{2}\right)^2\right)} \leq 1,$$

so that the scheme is unconditionally stable. A similar analysis can be done for the Peaceman-Rachford scheme, but we omit it here for brevity.

3.7 Numerical solution of systems of linear equations

In Sections 3.3 and 3.5 above, we have explored different ways to approximate a partial differential equation. In particular, our analysis has boiled down to solving a matrix equation of the form $\mathrm{A}\mathrm{x} = \mathrm{b}$, where $\mathrm{A} \in \mathcal{M}_m(\mathbb{R})$ and x and b are two $\mathbb{R}^m$-valued vectors. An interesting feature outlined above was that the matrix A had a tridiagonal structure, i.e. it can be written as $\mathrm{A} = \mathrm{T}_m(\mathrm{a}, \mathrm{b}, \mathrm{c})$, where we use the tridiagonal notation (3.3.8). By construction, this matrix is invertible, and hence the solution to the matrix equation is simply $\mathrm{x} = \mathrm{A}^{-1}\mathrm{b}$. Classical matrix inversion results—in particular the Gaussian elimination method—are of order $\mathcal{O}(m^3)$ (in the number of operations). Since we may want to have a fine discretisation grid, the dimension m may be very large, and these methods may be too time consuming for high-dimensional problems. We may however use the simplified structure of the matrix A.

3.7.1 Gaussian elimination

Gaussian elimination is a method devised to solve systems of linear equations, and hence to compute the inverse of a matrix. Note that it was invented in the second century BC, but got its name in the 1950s based on the fact that Gauss came up with standard notations. Consider the system

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_m \end{pmatrix}.$$

Assume for now that the coefficient a_{11} is not null. Dividing by it and modifying the last $m-1$ lines, we obtain

$$\begin{pmatrix} 1 & a_{12}/a_{11} & \dots & a_{1m}/a_{11} \\ 0 & a_{22} - a_{12}(a_{21}/a_{11}) & \dots & a_{2m} - a_{1m}(a_{21}/a_{11}) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2} - a_{12}(a_{m1}/a_{11}) & \dots & a_{mm} - a_{1m}(a_{m1}/a_{11}) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1/a_{11} \\ b_2 - b_1(a_{21}/a_{11}) \\ \vdots \\ b_m - b_1(a_{m1}/a_{11}) \end{pmatrix}.$$

If we now repeat this process, we obtain

$$\begin{pmatrix} 1 & \widetilde{a}_{12} & \widetilde{a}_{13} & \dots & \widetilde{a}_{1m} \\ 0 & 1 & \widetilde{a}_{23} & \dots & \widetilde{a}_{2m} \\ 0 & 0 & 1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \widetilde{a}_{m-1,m} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} \widetilde{b}_1 \\ \widetilde{b}_2 \\ \vdots \\ \vdots \\ \widetilde{b}_m \end{pmatrix}.$$

where all the coefficients $\widetilde{a}_{ij}$ and $\widetilde{b}_i$ are determined recursively. The matrix equation is then solved by backward substitution:

$$\begin{cases} x_m = & \widetilde{b}_m, \\ x_{m-k} = & \widetilde{b}_{m-k} - \displaystyle\sum_{j=m-k+1}^{m} \widetilde{a}_{m-k,j} x_j, \qquad \text{for any } k = 1, \dots, m-1. \end{cases}$$

This method has however several drawbacks. The first obvious one occurs as soon as one diagonal element becomes null, in which case, we cannot proceed as above. From a numerical point of view, even if not null, a very small value of the diagonal element can lead to numerical (decimal) truncation, which can get amplified as the scheme goes on.

Exercise 31. Note that a zero on the diagonal does not mean that the matrix is singular!! Consider for example the matrix equation equation:

$$\begin{pmatrix} 0 & 3 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}.$$

What happens when one applies Gauss elimination? Is there however an (obvious) solution?

The way to bypass this issue is to use pivoting. The idea is that one can interchange rows and columns of the matrix A (keeping track of the corresponding modified vectors x and b) without modifying the problem. Interchanging two rows implies interchanging the corresponding two elements of the vector b. Interchanging two columns implies interchanging the corresponding two elements of the vector x.

Exercise 32. Apply pivoting to the system in Exercise 31 to use Gaussian elimination.

The Gaussian elimination method therefore requires the computation of an invertible matrix B such that the matrix BA is upper triangular. Once this is done, all that is left is (i) to compute the product Bb and (ii) to solve the triangular system (BA)x = Bb by backward substitution. The existence of such a matrix B is guaranteed by the following lemma, the proof of which is simply the construction of the Gaussian elimination method itself.

Lemma 3.7.1. *Let* A *be a square matrix. There exists at least one invertible matrix* B *such that* BA *is upper triangular.*

Some remarks are in order here:

- we never compute the matrix B;
- if the original matrix A is not invertible, then one of the diagonal elements of the matrix BA will be null, and the backward substitution will be impossible;
- it is easy to show that det(A) = ±det(BA), the sign depending on the number of permutations needed in order to remove any null pivot.

3.7.2 LU decomposition

In the Gaussian elimination method above, the vector b is modified when solving the matrix equation. This makes the method rather cumbersome when one has to solve a recursive equation, repeating the same procedure at each step. A quicker computation can be achieved by first finding the so-called LU decomposition for the matrix $\mathrm{A} \in \mathcal{M}_m(\mathbb{R})$, i.e. by determining a lower triangular matrix $\mathrm{L} \in \mathcal{M}_m(\mathbb{R})$ and an upper triangular matrix $\mathrm{U} \in \mathcal{M}_m(\mathbb{R})$ such that $\mathrm{A} = \mathrm{LU}$.

Proposition 3.7.2. *Let* $\mathrm{A} = (a_{ij}) \in \mathcal{M}_m(\mathbb{R})$ *such that all the subdiagonal matrices*

$$(a_{11}), \begin{pmatrix} a_{11} & a_{11} \\ a_{21} & a_{22} \end{pmatrix}, \ldots, \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix}, \ldots, \begin{pmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mm} \end{pmatrix},$$

are invertible. Then there exist a unique lower triangular matrix $\mathrm{L} \in \mathcal{M}_m(\mathbb{R})$ *with unit diagonal and upper triangular matrix* $\mathrm{U} \in \mathcal{M}_m(\mathbb{R})$ *such that* $\mathrm{A} = \mathrm{LU}$.

In particular, if the matrix A is positive definite, then the proposition holds. The proof of this proposition is again constructive, and the following practical computation of the decomposition is very similar. Assume that such a decomposition holds. For any $1 \le i, j \le m$ we have

$$a_{ij} = \sum_{k=1}^{m} l_{ik} u_{kj} = \sum_{k=1}^{i \wedge j} l_{ik} u_{kj} = \begin{cases} l_{i1}u_{1j} + \ldots + l_{ii}u_{ij}, & \text{if } i < j, \\ l_{i1}u_{1j} + \ldots + l_{ij}u_{jj}, & \text{if } i > j, \\ l_{i1}u_{1j} + \ldots + l_{ii}u_{jj}, & \text{if } i = j. \end{cases}$$

There are m^2 equations to solve and $m^2 + m$ variables to determine. We therefore have the freedom to choose m of them arbitrarily. Following Proposition 3.7.2, *Crout's algorithm* proceeds as follows:

(i) for each $i = 1, \ldots, m$, set $l_{ii} = 1$;

(ii) for each $j = 1, \ldots, m$, let

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \qquad \text{for } i = 1, \ldots, j, \tag{3.7.1}$$

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \right), \text{ for } i = j+1, \ldots, m. \tag{3.7.2}$$

Note that under the conditions of Proposition 3.7.2, the term u_{jj} can never be null.

Solving the system

We are interested here in solving the matrix equation $\mathrm{Ax} = \mathrm{b}$, where $\mathrm{A} \in \mathcal{M}_m(\mathbb{R})$ and x and b are two $\mathbb{R}^m$-valued vectors. If the matrix A admits an LU-decomposition, then there exist two matrices L and U (respectively lower triangular and upper triangular) such that $\mathrm{A} = \mathrm{LU}$. We have already used this factorisation in Theorem 2.1.7 and in Section 3.6.2 above. The system $\mathrm{Ax} = \mathrm{b}$

can therefore be written as $\mathrm{L}(\mathrm{Ux}) = \mathrm{b}$. Set $\mathrm{z} := \mathrm{Ux} = (z_i)_{1\leq i\leq n}$. The new system then reads

$$\begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{m1} & l_{m2} & \dots & l_{mm} \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ \vdots \\ z_m \end{pmatrix} = \begin{pmatrix} l_{11}z_1 \\ l_{21}z_1 + l_{22}z_2 \\ \vdots \\ l_{m1}z_1 + \ldots + l_{mm}z_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_m \end{pmatrix}.$$

So that, starting from the last line, we can solve this equation by successive forward substitution:

$$z_1 = \frac{b_1}{l_{11}}, \qquad z_k = \frac{1}{l_{kk}}\left(b_k - \sum_{j=1}^{k-1} l_{kj}z_j\right), \quad \text{for } k = 2, \ldots, m.$$

Once this is done, we can then solve the other part $\mathrm{Ux} = \mathrm{z}$, i.e.

$$\begin{pmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ 0 & u_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & u_{m-1,m} \\ 0 & 0 & 0 & u_{mm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} u_{11}x_1 + \ldots + u_{1m}x_m \\ \vdots \\ u_{m-1,n}x_{n-1} + u_{mm}x_m \\ u_{mm}x_m \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ \vdots \\ z_m \end{pmatrix}.$$

Backward substitution hence gives

$$x_m = \frac{z_m}{u_{mm}}, \qquad x_k = \frac{1}{u_{kk}}\left(z_k - \sum_{j=m}^{k+1} u_{jk}x_j\right), \quad \text{for } k = m-1, \ldots, 1.$$

Note that the backward substitution step is exactly the same as in the Gaussian elimination method. The LU decomposition requires $\sum_{j=1}^{m-1}\sum_{i=j+1}^{m}\left(1 + \sum_{k=j+1}^{m}\right)$ operations and the forward-backward substitution requires $2\sum_{i=1}^{m} i$ steps. The total amount of operations is hence of order $m^3/3$ as m becomes large. The determinant of the matrix A is also trivially equal to $u_{11}\cdots u_{mm}$, and also requires an order of $m^3/3$ operations.

Remark 3.7.3. In the particular case where the matrix A is tridiagonal (as in the one-dimensional θ-schemes), we have the obvious decomposition $\mathrm{A} = \mathrm{T}(\mathrm{a}, \mathrm{b}, \mathrm{c}) = \mathrm{T}(1, \mathrm{l}, 0)\mathrm{T}(\mathrm{d}, 0, \mathrm{u})$, where $\mathrm{T}(1, \mathrm{l}, 0)$ is lower triangular and $\mathrm{T}(\mathrm{d}, 0, \mathrm{u})$ is upper triangular (recall the tridiagonal notation (3.3.8)). The vectors d, u and l are computed recursively.

Exercise 33. Give an explicit representation for the vectors d, u and l in Remark 3.7.3.

3.7.3 Cholesky decomposition

We have already seen this decomposition in Theorem 2.1.7 when considering correlation matrices. It indeed solely applies to real symmetric positive definite matrices.

Proposition 3.7.4. *Let* $\mathrm{A} \in \mathcal{M}_m(\mathbb{R})$ *be a real symmetric positive definite matrix. There exists a unique lower triangular* $\mathrm{Z} \in \mathcal{M}_m(\mathbb{R})$ *such that* $\mathrm{ZZ}^T = \mathrm{A}$.

The proof of this proposition is based on the LU decomposition above and the fact that the inverse (when it exists) of a lower triangular matrix is again lower triangular. The construction of the algorithm was detailed in the proof of Theorem 2.1.7 and we shall not repeat it here. From a computational point of view—and this is left as a simple exercise—one can show that the number of operations is of order $m^3/6$ as the size m of the matrix A becomes large, which is twice as fast as the LU decomposition.

3.7.4 Banded matrices

The methods we have presented so far apply in fairly general situations. In the case of the finite difference schemes, the matrix A is tridiagonal and has hence many zeros. It is therefore natural to wonder whether the methods above are quicker for this special type of matrices.

Definition 3.7.5. A matrix $\mathrm{A} = (a_{ij}) \in \mathcal{M}_m(\mathbb{R})$ is called a banded matrix with half-bandwidth $p \in \mathbb{N}$ (equivalently with band size $2p$) if $a_{ij} = 0$ whenever $|i - j| > p$.

In the case of a tridiagonal matrix, for instance, $p = 1$. The following lemma—the proof of which is left as an exercise—shows why this is important.

Lemma 3.7.6. *For a matrix $\mathrm{A} \in \mathcal{M}_m(\mathbb{R})$ with half-bandwidth p, the number of operations is of order mp^2 for the LU decomposition and of order $mp^2/2$ for the Cholesky decomposition.*

3.7.5 Iterative methods

When the problem under consideration requires a fine grid, the iteration matrix becomes high dimensional. The Gaussian elimination method above can become computationally intensive, and one may need to resort to more suitable methods, in particular the so-called *iterative methods.* As before, we are interested in solving the system $\mathrm{Ax} = \mathrm{b}$, where $\mathrm{A} \in \mathcal{M}_m(\mathbb{R})$ and $\mathrm{b} \in \mathbb{R}^m$. We shall assume that the matrix A does not have any zero diagonal elements (if such is the case, we can always interchange some rows and columns in order to satisfy the assumption). In particular, it is clear that this system has a unique solution if and only if the matrix A is invertible, which we shall assume from now on. The essence of iterative methods is to rewrite this matrix equation as a fixed-point iteration

$$\mathrm{x}^{k+1} = \Psi(\mathrm{x}^k, \mathrm{b}), \qquad \text{for any } k \in \mathbb{N}. \tag{3.7.3}$$

A solution x^* of the equation $\mathrm{x}^* = \Psi(\mathrm{x}^*, \mathrm{b})$ is called a fixed-point.

Definition 3.7.7. The fixed-point iteration (3.7.3) is said to be

(i) consistent with the matrix A if for any b, the vector $\mathrm{A}^{-1}\mathrm{b}$ is a fixed-point of (3.7.3);

(ii) convergent if for any $\mathrm{b} \in \mathbb{R}^m$, there exists some vector $\mathrm{x}^* \in \mathbb{R}^m$ such that the sequence $(\mathrm{x}^k)_{k\geq 0}$ defined by (3.7.3) converges to x^*, for any seed x^0;

(iii) linear if Ψ is linear in its two variables: $\Psi(\mathrm{x}, \mathrm{b}) = \mathrm{Hx} + \mathrm{Nb}$, H being the iteration matrix.

The basic idea in these methods is to decompose the matrix A as the sum of three matrices: a diagonal matrix D, a lower triangular matrix L and an upper triangular matrix U, such that the diagonal elements of the two matrices L and U are null. The problem $\mathrm{Ax} = \mathrm{b}$ therefore reads

$$(\mathrm{D} + \mathrm{L} + \mathrm{U})\, \mathrm{x} = \mathrm{b},$$

which we can further rewrite as $\mathrm{x} = \mathrm{D}^{-1}\mathrm{b} - \mathrm{D}^{-1}(\mathrm{L} + \mathrm{U})\,\mathrm{x}$. Note that the inverse matrix D^{-1} is well defined by the assumptions on A. The methods that follow are based on this decomposition. The following lemma—the proof of which is left as an exercise—provides some intuition.

Lemma 3.7.8. *Let the matrix* A *be invertible and the mapping* Ψ *linear. Then the fixed-point iteration is consistent with* A *if and only if*

$$\mathrm{H} = \mathrm{I} - \mathrm{NA} \iff \mathrm{N} = (\mathrm{I} - \mathrm{H})\,\mathrm{A}^{-1} \iff \mathrm{A}^{-1} = (\mathrm{I} - \mathrm{H})^{-1}\,\mathrm{N},$$

where I *is the identity matrix in* $\mathcal{M}_m(\mathbb{R})$.

We shall encounter a simpler version of the following theorem below. Its proof, based on simple yet tedious properties of the spectral radius is rather long and hence left for the avid reader.

Theorem 3.7.9. *If* $\rho(\mathrm{H}) < 1$ *then there exists a unique fixed point* $\mathrm{x}^* = (\mathrm{I} - \mathrm{H})^{-1}\mathrm{Nb}$ *to* (3.7.3) *and the iteration converges to* x^* *for any starting point.*

Proof. Since $\rho(\mathrm{H}) < 1$, then 1 does not belong to the spectrum of H, and therefore the iteration $\Psi(\mathrm{x}^k, \mathrm{b}) = \mathrm{x}^{k+1}$ admits a unique fixed-point $\mathrm{x}^* = (\mathrm{I} - \mathrm{H})^{-1}\mathrm{Nb}$. For any $k \geq 0$, the error $\mathrm{e}_k := \mathrm{x}^k - \mathrm{x}^*$ satisfies $\mathrm{e}_k = \mathrm{H}(\mathrm{x}^{k-1} - \mathrm{x}^*) = \mathrm{He}_{k-1} = \ldots = \mathrm{H}^k\mathrm{e}_0$. It can be proved (using Jordan's decomposition) that for every $\varepsilon > 0$, there exists a norm $\|\cdot\|_\varepsilon$ such that $\|\mathrm{H}\|_\varepsilon \leq \rho(\mathrm{H})$. Therefore, for any $k \geq 0$,

$$\|\mathrm{e}_k\|_\varepsilon = \|\mathrm{H}^{k-1}\mathrm{e}_0\|_\varepsilon \leq \|\mathrm{H}\|_\varepsilon^{k-1}\|\mathrm{e}_0\|_\varepsilon.$$

We then deduce that the sequence $(\mathrm{e}_k)_{k\geq 0}$ converges to zero in the norm $\|\cdot\|_\varepsilon$. Since all norms on the finite-dimensional space $\mathbb{R}^n$ are equivalent, the sequence $(\mathrm{x}^k)_{k\geq 0}$ converges to x^*. □

When the spectral radius is greater than one, we can actually prove that the scheme does not converge:

Proposition 3.7.10. *If* $\rho(\mathrm{H}) \geq 1$, *then there exist two vectors* x^0 *and* b *such that the sequence* $(\mathrm{x}^k)_{k\geq 0}$ *does not converge.*

Proof. Let us choose $\mathrm{b} = 0$ for simplicity, and pick an eigenvalue λ of H such that $|\lambda| \geq 1$; let x^0 be the corresponding (non-zero) eigenvector. If $\lambda = 1$, then the sequence $(\mathrm{x}^k)_{k\geq 0}$ clearly converges since $\mathrm{x}^k = \mathrm{x}^0$ for all $k \geq 0$. However, should one choose another seed, say $\widetilde{\mathrm{x}}^0 = -\mathrm{x}^0$, then the new sequence converges to $-\widetilde{\mathrm{x}}_0$. If $|\lambda| > 1$, then clearly the sequence $(\mathrm{x}^k)_{k\geq 0}$ diverges to $+\infty$ or $-\infty$. In the case $|\lambda| = 1$, but $\lambda \neq 1$, i.e. $\lambda = \mathrm{e}^{\mathrm{i}\theta}$ for some $\theta \in (0, 2\pi)$, we obtain $\mathrm{x}^k = \mathrm{e}^{\mathrm{i}k\theta}\mathrm{x}^0$, so that $\limsup_{m\uparrow\infty} |\mathrm{x}^n - \mathrm{x}^m| = 2|\mathrm{x}^0|$, for any $n \geq 0$. Therefore, the sequence $(\mathrm{x}^k)_{k\geq 0}$ is not a Cauchy sequence, and hence does not converge. □

Jacobi iteration

The Jacobi method—due to Carl Gustav Jacob Jacobi (1804-1851)—proceeds by approximating the solution x^k at step $k \geq 1$ by

$$\mathrm{x}^k = \mathrm{H}\mathrm{x}^{k-1} + \mathrm{D}^{-1}\mathrm{b},$$

where $\mathrm{H} := -\mathrm{D}^{-1}(\mathrm{L} + \mathrm{U})$, and where the seed x^0 can be chosen arbitrarily. In component notations, this can also be written as

$$\mathrm{x}_i^k = \frac{1}{a_{ii}}\left(\mathrm{b}_i - \sum_{j=1, j\neq i}^{m} a_{ij}\mathrm{x}_j^{k-1}\right), \qquad \text{for } i = 1\ldots, m.$$

In the context of PDE solving, the solution vector x corresponds to the function u to be solved for evaluated at some time t_i. A sensible choice for x^0 is to consider the function u at the previous time step t_{i-1}. The scheme is then stopped as soon as some predefined accuracy is reached, i.e. as soon as

$$\left\|\mathrm{x}^k - \mathrm{x}^{k-1}\right\| \leq \varepsilon, \qquad \text{for some predefined tolerance } \varepsilon > 0.$$

One may wonder whether the scheme is actually convergent, i.e. whether or not the family $(\mathrm{x}_k)_{k\geq 0}$ converges to some vector x as k tends to infinity. For ease of notation, let us rewrite the Jacobi iteration as

$$\mathrm{x}^k = \mathrm{H}\mathrm{x}^{k-1} + \beta,$$

where $\beta := \mathrm{D}^{-1}\mathrm{b}$, and note that both β and H are independent of k (the method is *stationary*). We then have

$$\begin{aligned}
\left\|\mathrm{x}^k\right\| &\leq \left\|\mathrm{H}\mathrm{x}^{k-1} + \beta\right\| \\
&\leq \|\mathrm{H}\| \left\|\mathrm{x}^{k-1}\right\| + \|\beta\| \\
&\leq \|\mathrm{H}\|^2 \left\|\mathrm{x}^{k-2}\right\| + \|\mathrm{H}\| \|\beta\| + \|\beta\| \\
&\leq \ldots \leq \|\mathrm{H}\|^k \left\|\mathrm{x}^0\right\| + \left(\sum_{i=0}^{k-1} \|\mathrm{H}\|^i\right) \|\beta\|.
\end{aligned}$$

If $\|\mathrm{H}\| < 1$, the sum above converges to $(1 - \|\mathrm{H}\|)^{-1} \|\beta\|$. One may conclude that this condition suffices to ensure the convergence of the scheme, whatever the initial data x^0 is. While this is indeed true, a stronger result holds. Recall that the spectral radius of the matrix H is given by

$$\rho(\mathrm{H}) := \max_{i=1,\ldots,m} |\lambda_i|,$$

where the λ_i represents the eigenvalues of H. We now know that the condition $\rho(\mathrm{H}) < 1$ ensures that H^k converges to zero as k tends to infinity, and the quantity $\sum_{k\geq 0} \mathrm{H}^k = (\mathrm{I} - \mathrm{H})^{-1}$ is called the *resolvent* of H. Applying this to the scheme, we see that

$$\lim_{k\to\infty} \mathrm{x}^k = \beta(\mathrm{I} - \mathrm{H})^{-1} =: \mathrm{x}^\infty,$$

for any initial guess x^0. Note that this limit is also a fixed point of the algorithm (exercise).

Remark 3.7.11. It can further be shown that the speed of convergence is of order $\rho(\mathrm{H})^k$, i.e.

$$\frac{\left\|\mathrm{x}^k - \mathrm{x}^\infty\right\|}{\left\|\mathrm{x}^0 - \mathrm{x}^\infty\right\|} = \mathcal{O}\left(\rho(\mathrm{H})^k\right).$$

Hence, the smaller the spectral radius, the quicker the convergence.

Computing all the eigenvalues of the matrix H may however be computationally intensive, and we may wish to find a shortcut to determine whether the matrix is convergent or not.

Definition 3.7.12. A matrix $\mathrm{Q} = (q_{ij}) \in \mathcal{M}_m(\mathbb{R})$ is said to be (weakly) *diagonally dominant* if

$$|q_{ii}| \geq \sum_{j=1,\ldots,m, j\neq i} |q_{ij}|, \qquad \text{for any } i = 1, \ldots, m.$$

If the inequality is strict, we actually say that the matrix is strictly diagonally dominant.

Proposition 3.7.14 below shows why such a concept is important. We need a preliminary lemma before though.

Lemma 3.7.13 (Hadamard)**.** *A diagonally dominant matrix is invertible.*

Proof. Let $\mathrm{A} = (a_{ij})$ be such a matrix in $\mathcal{M}_m(\mathbb{C})$ and $\mathrm{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ such that $\mathrm{Ax} = 0$. Define $1 \leq p \leq m$ such that $|x_p| = \max_{1\leq i\leq m} |x_i|$. Therefore

$$a_{pp}x_p + \sum_{i\neq p} a_{pi}x_i = 0.$$

If $\mathrm{x} \neq 0$, then clearly $|x_p| > 0$ and

$$|a_{pp}||x_p| \leq \sum_{i\neq p} |a_{pi}||x_i| \leq |x_p| \sum_{i\neq p} |a_{pi}| < |a_{pp}||x_p|,$$

which yields a contradiction, so that x is indeed null, and the lemma follows. □

Proposition 3.7.14. *Consider the matrix equation* $\mathrm{Ax} = \mathrm{b}$*. If* A *is strictly diagonally dominant matrix then the Jacobi scheme converges.*

Proof. Let $\mathrm{H} := -\mathrm{D}^{-1}(\mathrm{L} + \mathrm{U})$ be the iteration matrix. Since for any $1 \leq i \leq m$, we have $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$, we deduce $\sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1$, and hence, for any $\lambda \in \mathbb{C}$ such that $|\lambda| \geq 1$, the matrix $(\lambda I - \mathrm{H})$ is diagonally dominant, and hence invertible by Lemma 3.7.13. Therefore $\rho(\mathrm{H}) < 1$ and the Jacobi iteration converges. □

Note that the matrix to check here is A. What the proposition essentially says is that if A is strictly diagonally dominant, then $\rho(\mathrm{H}) < 1$. This proposition therefore gives a simple way to determine whether or not the Jacobi scheme converges. When the matrix A is only (weakly) diagonally dominant, one must in principle check its spectral radius. However, since $\rho(\mathrm{H}) \leq \|\mathrm{H}\|$ (exercise), we only need to check the inequality $\|\mathrm{H}\| < 1$ in any convenient norm. The usual one is the sup norm, i.e.

$$\|\mathrm{H}\|_\infty := \sup_{i=1,\ldots,m} \sum_{j=1}^{m} |h_{ij}|. \tag{3.7.4}$$

Exercise 34. For a tridiagonal matrix A, write a simple recursive formula for the Jacobi iteration.

Exercise 35. Consider the matrix equation

$$\begin{pmatrix} 5 & -2 & 3 \\ -3 & 9 & 1 \\ 2 & -1 & -7 \end{pmatrix} \mathrm{x} = \begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix}.$$

Use the Jacobi iteration with starting value $\mathrm{x}^0 = 0$ and write the first three steps.

Exercise 36. Consider the matrix equation

$$\begin{pmatrix} 1 & -5 \\ 7 & -1 \end{pmatrix} \mathrm{x} = \begin{pmatrix} -4 \\ 6 \end{pmatrix}.$$

Compute the first steps of the Jacobi iteration with starting value $\mathrm{x}^0 = 0$ and prove that the scheme diverges.

Exercise 37. Recall that the matrix A in the implicit, explicit and Crank-Nicolson schemes is tridiagonal. Check the convergence of the Jacobi method for these schemes. You may want to check first whether the matrix is strictly diagonally dominant or not.

Gauss-Seidel iteration

This method is a modification of the Jacobi iteration scheme due to Carl Friedrich Gauss (1777-1855) and Philipp L. Seidel (1821-1896). It essentially follows the exact same steps but requires less memory. When applying the Jacobi iteration scheme, the whole vector x remains unchanged until the end of each step. In the Gauss-Seidel method, one uses the new value of the element x_i as soon as it is computed. The scheme therefore reads

$$\mathrm{x}^k = -\left(\mathrm{L} + \mathrm{D}\right)^{-1} \mathrm{U}\mathrm{x}^{k-1} + \left(\mathrm{L} + \mathrm{D}\right)^{-1} \mathrm{b},$$

Generally, the convergence of the Gauss-Seidel scheme follows from that of the Jacobi iteration. Some counterexamples do exist though. If the matrix A is strictly diagonally dominant, the scheme will however be convergent for any initial seed x^0. A general result is stated below in Proposition 3.7.15 without proof.

Exercise 38. For a tridiagonal matrix A, write a simple recursive formula for this scheme.

Exercise 39. Compute the first few steps of the iteration using the same example as in Exercise 35.

Exercise 40. Compare the divergence of Exercise 36 using the Jacobi iteration and the Gauss-Seidel method. Interchange now the rows of the system and show that this leads to a convergent iteration scheme.

Successive Over Relaxation method (SOR)

The SOR method is essentially a linear interpolation between the Gauss-Siedel iteration and the previous approximate solution. Let us consider the Gauss-Seidel iteration scheme. We have

$$\begin{aligned}\mathrm{x}^k &= -\left(\mathrm{L}+\mathrm{D}\right)^{-1}\left[\mathrm{U}\mathrm{x}^{k-1}-\mathrm{b}\right]\\ &= \mathrm{x}^{k-1}-\left(\mathrm{L}+\mathrm{D}\right)^{-1}\left(\mathrm{L}+\mathrm{D}\right)\mathrm{x}^{k-1}-\left(\mathrm{L}+\mathrm{D}\right)^{-1}\left[\mathrm{U}\mathrm{x}^{k-1}-\mathrm{b}\right]\\ &= \mathrm{x}^{k-1}-\left(\mathrm{L}+\mathrm{D}\right)^{-1}\left[\mathrm{A}\mathrm{x}^{k-1}-\mathrm{b}\right]\\ &= \mathrm{x}^{k-1}-\left(\mathrm{L}+\mathrm{D}\right)^{-1}\xi^{k-1}\end{aligned}$$

where the vector ξ^{k-1} represents residual vector. The SOR method corrects this by penalising (or relaxing) the correction term ξ^{k-1}. Let ω be a strictly positive number. The SOR scheme reads $\mathrm{x}^k = \mathrm{x}^{k-1}-\omega\left(\mathrm{L}+\mathrm{D}\right)^{-1}\xi^{k-1}$. It can be shown that the scheme converges only when the relaxation parameter ω lies in $(0,2)$, and we refer the interested reader to [37] for more details and a recent review of the optimal choice of ω. In particular, if ρ_J denotes the spectral radius of the iteration matrix in the Jacobi method, it can be proved that the optimal relaxation parameter reads

$$\omega = \frac{2}{1+\sqrt{1-\rho_J^2}}.$$

Exercise 41. For a tridiagonal matrix A, write a simple recursive formula for the SOR method.

We have here merely scratched the surface of iterative methods for matrix equations. In particular we have not said much—or nothing at all—concerning the complexity of each scheme. We refer the interested reader to [52] for more details on the subtleties of these and for more advanced numerical methods used in scientific computing. The following proposition—which we state without proof—provides the main convergence results for these schemes.

Proposition 3.7.15. *Consider the matrix equation* $\mathrm{Ax}=\mathrm{b}$. *If* A *is symmetric definite positive, then the Gauss-Seidel iteration converges and the SOR converges if* $\omega \in (0,2)$. *The Jacobi iteration converges if we further assume that* A *is tridiagonal.*

Chapter 4

Fourier and integration methods

4.1 A primer on characteristic functions

4.1.1 Fourier transforms and their inverses

Reminder on L^p spaces

Let $(\mathbb{R}^n, \Sigma)$ be a measurable space and associate the Lebesgue measure to it. For $p \in [1, \infty)$, the space of functions $f : \mathbb{R}^n \to \mathbb{R}$ satisfying

$$\|f\|_p := \left(\int_{\mathbb{R}} |f(\mathrm{x})|^p \mathrm{dx}\right)^{1/p} < \infty$$

is called $L^p(\mathbb{R})$. When $p = \infty$, the space $L^\infty(\mathbb{R}^n)$ is the set of functions which are essentially bounded, i.e. bounded everywhere except possibly on sets of (Lebesgue) measure zero. More precisely, we define

$$\|f\|_\infty := \text{ess sup}_{\mathrm{x}\in\mathbb{R}^n} f(\mathrm{x}) := \inf\{K \geq 0 : |f(\mathrm{x})| \leq K \text{ for almost every } \mathrm{x} \in \mathbb{R}^n\}.$$

As an example on the real line, consider the function defined by $f(x) = 1$ if x is a rational and zero otherwise. Since the set of rational numbers has Lebesgue measure zero, it follows that the essential supremum (its infinity norm) is equal to zero whereas its supremum is clearly equal to one. For any $p \in [1, \infty]$ the space $L^p(\mathbb{R}^n)$ thus defined is a Banach space (complete normed vector space). Unless $p = 2$ however, these spaces are not Hilbert spaces (i.e. not complete with respect to the norm associated with its inner product).

Fourier transforms on Schwartz space

Fourier transforms are often introduced on the space $L^1(\mathbb{R}^n)$. However some manipulations are not allowed there due to convergence issues. The natural space to introduce such a transform is the Schwartz space, which we define below after some notations. A *multi-index* α is an ordered

n-tuple $(\alpha_1, \ldots, \alpha_n)$ of non-negative integers. For a smooth function $f \in \mathbb{R}^n$, we shall denote $\partial^\alpha f$ the multiple derivative (whenever it exists) $\partial^{\alpha_1} \ldots \partial^{\alpha_n} f$, and hence $|\alpha| := \alpha_1 + \ldots + \alpha_n$ represents the *total order of differentiation.* If $\mathrm{x} = (x_1, \ldots, x_n)$ is a vector in R^n and α a multi-index, we shall further use the notation $\mathrm{x}^\alpha := x_1^{\alpha_1} \ldots x_n^{\alpha_n}$.

Definition 4.1.1. A function $f \in C^\infty(\mathbb{R}^n)$ is called a Schwartz function if for every pair of multi-indices α and β, there exists a positive constant C—possibly function of α and β—such that

$$\rho_{\alpha,\beta}(f) := \sup_{\mathrm{x}\in\mathbb{R}^n} \left|\mathrm{x}^\alpha \partial^\beta f(\mathrm{x})\right| \leq C.$$

The space of all such functions is called the Schwartz space and is denoted $\mathcal{S}(\mathbb{R}^n)$. The operator $\rho_{\alpha,\beta}$ is called the Schwartz seminorm (i.e. a norm without the positivity property).

Exercise 42. Show that the Schwartz seminorm is a norm on the Schwartz space.

The Schwartz space therefore represents all the (smooth) functions that decay faster than any polynomial at infinity. For example, the function $\mathrm{x} \mapsto \mathrm{e}^{-|\mathrm{x}|^2}$ belongs to $\mathcal{S}(\mathbb{R}^n)$ whereas the function $\mathrm{x} \mapsto \mathrm{e}^{-|\mathrm{x}|}$ is not in $\mathcal{S}(\mathbb{R}^n)$ since it is not differentiable at the origin. A polynomial of order $m \in \mathbb{N}$ is clearly not in $\mathcal{S}(\mathbb{R}^n)$, but any smooth function with compact support is. Convergence of functions in the Schwartz space is defined with respect to the Schwartz seminorm, and is a strong form of convergence as the following proposition shows.

Proposition 4.1.2. *Let f and $(f_k)_{k\in\mathbb{N}}$ be in $\mathcal{S}(\mathbb{R}^n)$. If (f_k) converges to f in $\mathcal{S}(\mathbb{R}^n)$ (i.e. $\lim_{k\uparrow\infty} \rho_{\alpha,\beta}(f - f_k) = 0$) then the family (f_k) converges to f in L^p, for any $p \geq 1$.*

We can now define the Fourier transform

Definition 4.1.3. For a function f in $\mathcal{S}(\mathbb{R}^n)$, its Fourier transform $\widehat{f}$ is defined by

$$\widehat{f}(\xi) := \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\xi\cdot\mathrm{x}} f(\mathrm{x})\mathrm{dx}, \qquad \text{for any } \xi \in \mathbb{R}^n.$$

Remark 4.1.4. In the literature, one can also find the notation $\mathcal{F}f$ for the Fourier transform. The definition $\widehat{f}(\xi) := \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}\xi\cdot\mathrm{x}} f(\mathrm{x})\mathrm{dx}$ can also be found, as well as factors $(2\pi)^{-n}$ or $(2\pi)^{-n/2}$ in front. These notations lead to the same properties and are used differently according to one's preferences.

Exercise 43. Determine the Fourier transform (if it exists) of the function $f : x \in \mathbb{R} \mapsto \mathrm{e}^{-x^2}$.

The following proposition lists some of the most fundamental properties of the Fourier transform. We leave these items to prove as an exercise for the interested reader.

Proposition 4.1.5. *Let f and g be two functions on $\mathcal{S}(\mathbb{R}^n)$, $a \in \mathbb{R}$, α a multi-index, and denote $\widetilde{f}$ the function defined by $\widetilde{f}(\mathrm{x}) := f(-\mathrm{x})$.*

(i) $\|\widehat{f}\|_{L^\infty} \leq \|f\|_{L^1}$.

(ii) *(Linearity)* $\widehat{af+g} = a\widehat{f} + \widehat{g}$.

(iii) $\widetilde{\widehat{f}} = \widehat{\widetilde{f}}$.

(iv) *(Differentiation)* $\widehat{\partial^\alpha f}(\xi) = (-\mathtt{i}\xi)^\alpha \widehat{f}(\xi)$, *for any* $\xi \in \mathbb{R}$.

(iv) $\widehat{f} \in \mathcal{S}(\mathbb{R}^n)$, *i.e. the Fourier transform is an isomorphism in* $\mathcal{S}(\mathbb{R}^n)$.

(v) *(Convolution)* $f * g \in \mathcal{S}(\mathbb{R}^n)$ *and* $\widehat{f * g} = \widehat{f} \cdot \widehat{g}$, *where*

$$(f * g)(\mathrm{x}) := \int_{\mathbb{R}^n} f(\mathrm{x}-\mathrm{y}) g(\mathrm{y}) \mathrm{dy}, \qquad \textit{for any } \mathrm{x} \in \mathbb{R}^n.$$

We can now define the inverse Fourier transform:

Definition 4.1.6. For a Schwartz function f in $\mathbb{R}^n$, we define its inverse Fourier transform as

$$\check{f}(\xi) := (2\pi)^{-n} \widehat{f}(-\xi).$$

We summarise some properties of the inverse Fourier transform in the following theorem, the proof of which is left as a—not easy—exercise.

Theorem 4.1.7. *Let f and g be two functions in $\mathcal{S}(\mathbb{R}^n)$. The following identities hold*

(i) $\int_{\mathbb{R}^n} \widehat{f}(\mathrm{x}) g(\mathrm{x}) \mathrm{dx} = \int_{\mathbb{R}^n} f(\mathrm{x}) \widehat{g}(\mathrm{x}) \mathrm{dx}$.

(ii) $\widehat{\check{f}} = \check{\widehat{f}} = f$.

(iii) *(Plancherel identity)* $\|f\|_{L^2} = (2\pi)^{-n} \|\widehat{f}\|_{L^2} = \|\check{f}\|_{L^2}$.

(iv) $\int_{\mathbb{R}^n} f(\mathrm{x}) g(\mathrm{x}) \mathrm{dx} = \int_{\mathbb{R}^n} \widehat{f}(\mathrm{x}) \check{g}(\mathrm{x}) \mathrm{dx}$.

The Schwartz space is however not very large, and we wish to extend the definition (and the properties) of the Fourier transform to larger spaces such as L^1 and L^2. Definition 4.1.3 clearly makes sense as a convergent integral for functions in $L^1(\mathbb{R}^n)$, and most of the properties above can be checked to hold. However one does not always have $\widehat{f} \in L^1(\mathbb{R}^n)$ (take for example $f(x) \equiv \mathrm{e}^{-x} \mathbb{1}_{\{x \geq 0\}}$), and hence the inverse Fourier transform is not defined (and therefore $\widehat{\check{f}} \neq f$). Things become more subtle on $L^2(\mathbb{R}^n)$ since the integral defining the Fourier transform in Definition 4.1.3 does not converge absolutely (consider for example the function $f \in L^2(\mathbb{R}) \setminus L^1(\mathbb{R})$ defined by $f(x) \equiv (1+x^2)^{-1/2}$). We can however make sense of it as the limit of Fourier transforms of functions in $L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$. We shall not delve into these details here and refer the interested reader to [31] for more details.

Fourier transforms on $L^1(\mathbb{R})$

From now on, we shall only be looking at Fourier transform on the real line, and for any $f \in L^1(\mathbb{R})$, we recall the definition of its Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$:

$$\widehat{f}(\xi) := \int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}\xi x} f(x)\mathrm{d}x, \qquad \text{for all } \xi \in \mathbb{R}. \tag{4.1.1}$$

Clearly $f \in L^1(\mathbb{R}^n)$ implies that $\widehat{f} \in L^\infty(\mathbb{R}^n)$. Furthermore if f is non-negative, then the identity $\|\widehat{f}\|_\infty = \widehat{f}(0) = \|f\|_1$ holds.

Remark 4.1.8. In the previous chapter, we used the Fourier transform of a function on a grid, namely for a function f_δ on a grid with $\delta > 0$ step size, we defined (see (3.3.15)):

$$\widehat{f_\delta}(\xi) := \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} \delta \mathrm{e}^{-\mathtt{i}m\delta\xi} f_\delta(m\delta), \qquad \text{for } \xi \in \left[-\frac{\pi}{\delta}, \frac{\pi}{\delta}\right].$$

This corresponds to a Riemann sum, so that, letting δ tend to zero, we obtain

$$\lim_{\delta \downarrow 0} \widehat{f_\delta}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathtt{i}\xi x} f(x)\mathrm{d}x, \qquad \text{for all } \xi \in \mathbb{R},$$

where the function f is the continuous version of the function f_δ.

The following result is easy to prove:

Lemma 4.1.9. *If $f \in L^1(\mathbb{R}^n)$ then $\widehat{f}$ is continuous.*

Proof. Fix some $\xi \in \mathbb{R}^n$. Then

$$\widehat{f}(\xi + h) = \int_{\mathbb{R}^n} \mathrm{e}^{\mathtt{i}(\xi+h)\cdot \mathrm{x}} f(\mathrm{x})\mathrm{dx}.$$

As h tends to zero, the integrand converges pointwise to $\mathrm{e}^{\mathtt{i}\xi\cdot\mathrm{x}} f(\mathrm{x})$, and hence the dominated convergence theorem implies the lemma. □

We have the following behaviour of the Fourier transform at infinity:

Lemma 4.1.10. *(Riemann-Lebesgue lemma) If f belongs to $L^1(\mathbb{R})$, then* $\lim_{|\xi|\uparrow\infty} |\widehat{f}(\xi)| = 0$.

The proof of the lemma in this generality is outside the scope of these lectures. The following version, with stronger assumptions, is easier to prove:

Lemma 4.1.11. *If $f \in \mathcal{C}^1(\mathbb{R}) \cap L^1(\mathbb{R})$, then* $\lim_{|\xi|\uparrow\infty} |\widehat{f}(\xi)| = 0$.

Proof. Since f belongs to $L^1(\mathbb{R})$, for any $\varepsilon > 0$, there exists $z > 0$ such that $\int_{\mathbb{R}\setminus[-z,z]} |f(x)|\mathrm{d}x < \varepsilon$. Now, for any $\xi \in \mathbb{R}$, an integration by parts yields

$$\int_{[-z,z]} \mathrm{e}^{\mathtt{i}\xi x} f(x)\mathrm{d}x = \left[\frac{\mathrm{e}^{\mathtt{i}\xi x} f(x)}{\mathtt{i}\xi}\right]_{-z}^{z} - \frac{1}{\mathtt{i}\xi} \int_{[-z,z]} \mathrm{e}^{\mathtt{i}\xi x} f'(x)\mathrm{d}x.$$

Therefore

$$\begin{aligned}|\widehat{f}(\xi)| = \left|\int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\xi x} f(x)\mathrm{d}x\right| &\leq \left|\int_{\mathbb{R}\setminus[-z,z]} \mathrm{e}^{\mathrm{i}\xi x} f(x)\mathrm{d}x\right| + \left|\int_{[-z,z]} \mathrm{e}^{\mathrm{i}\xi x} f(x)\mathrm{d}x\right| \\ &\leq \varepsilon + \frac{|f(z)| + |f(-z)|}{|\xi|} + \frac{1}{|\xi|}\int_{-z}^{z} |f'(x)|\mathrm{d}x.\end{aligned}$$

Since z is finite, the second term tends to zero as $|\xi|$ tends to infinity, and so does the last one, and the lemma follows. □

As mentioned above, we can define the inverse Fourier transform of a function f in $L^1(\mathbb{R})$ as

$$\check{f}(\xi) := \frac{1}{2\pi}\widehat{f}(-\xi) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\xi x} f(x)\mathrm{d}x, \qquad \text{for all } \xi \in \mathbb{R}.$$

Note that the identity $\check{\widehat{f}} \equiv f$ does hold in the Schwartz space, but not necessarily in $L^1(\mathbb{R})$. Consider in particular the function $f(x) \equiv \mathbb{1}_{(-1,1)}(x)$. Clearly f is bounded and hence belongs to $L^1(\mathbb{R})$. However for any $\xi \in \mathbb{R}$, we have $\widehat{f}(\xi) = 2\sin(\xi)/\xi$, which clearly does not belong to $L^1(\mathbb{R})$. We finally state the following Fourier inversion result for discontinuous functions, due to P.G. Lejeune Dirichlet (1805-1859) and Camille Jordan (1838-1922). Let $I = [\underline{\zeta}, \overline{\zeta}]$ be an interval on the real line and $\zeta = \{\underline{\zeta} = \zeta_0, \ldots, \zeta_n = \overline{\zeta}\}$ a partition of it. The variation of a function f on I, relative to ζ is defined as $V_I(g, \zeta) := \sum_{k=1}^{n} |f(\zeta_k) - f(\zeta_{k-1})|$. The function f is then said to have bounded variation if the supremum of $V_I(g, \zeta)$ over all partitions ζ of I is bounded.

Theorem 4.1.12. *[Dirichlet-Jordan Theorem] Let f be an integrable function on the real line, with bounded variation in a neighbourhood of some point $x \in \mathbb{R}$. Then*

$$\lim_{R\uparrow\infty} \frac{1}{2\pi}\int_{-R}^{R} \mathrm{e}^{-x\xi}\widehat{f}(\xi)\mathrm{d}\xi = \frac{1}{2}\left(\lim_{z\downarrow x} f(z) + \lim_{z\uparrow x} f(z)\right).$$

Remark 4.1.13. We have defined above Fourier transforms on the real line. It is sometimes convenient to extend this definition to the complex plane—or at least a subset of it—i.e. to define $\widehat{f}(\xi)$ from (4.1.1) for $\xi \in \mathbb{C}$. It is clear that there are no restrictions on $\Re(\xi)$. However, the Fourier transform is not well defined for all $\xi \in \mathbb{C}$, and is so only in a so-called *strip of regularity* $a < \Im(\xi) < b$ which is parallel to the real axis. This extension of the Fourier transform to (part of) the complex plane is called the *generalised Fourier transform*, which we also denote by $\widehat{f}$ and its inverse—whenever it exists—is given by

$$f(x) = \frac{1}{2\pi}\int_{\mathrm{i}z-\infty}^{\mathrm{i}z+\infty} \mathrm{e}^{-\mathrm{i}x\xi}\widehat{f}(\xi)\mathrm{d}\xi,$$

with $z \in (a, b)$, where the integration is carried out along a horizontal contour in the complex plane, parallel to the real axis.

4.1.2 Characteristic functions

For a one-dimensional random variable X taking values of the real line, we define its characteristic function $\phi_X : \mathbb{R} \to \mathbb{C}$ by

$$\phi_X(\xi) := \mathbb{E}\left(\mathrm{e}^{\mathtt{i}\xi X}\right), \qquad \text{for all } \xi \in \mathbb{R},$$

where $\mathtt{i} = \sqrt{-1}$. Since the map $\xi \mapsto \left|\mathrm{e}^{\mathtt{i}\xi X}\right|$ is continuous and bounded, the function ϕ_X is well defined. If the random variable X has a density f_X then

$$\phi_X(\xi) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathtt{i}\xi x} f_X(x)\mathrm{d}x, \qquad \text{for all } \xi \in \mathbb{R},$$

i.e. the characteristic function is the Fourier transform of the density: $\phi_X \equiv \widehat{f}_X$. The following properties are straightforward and we omit their proofs for brevity:

(i) $\phi(0) = 1$;

(ii) let $(\alpha, \beta) \in \mathbb{R}^2$. If $Y = \alpha X + \beta$, then $\phi_Y(\xi) = \mathrm{e}^{\mathtt{i}\beta\xi}\phi_X(\alpha\xi)$ for all $\xi \in \mathbb{R}$;

(iii) if X and Y are two independent random variables then $\phi_{X+Y}(\xi) = \phi_X(\xi)\phi_Y(\xi)$ for all $\xi \in \mathbb{R}$;

(iv) let $n \geq 1$ and assume that the n-th moment of X exists (i.e. $\mathbb{E}(X^n) < \infty$), then

$$\mathbb{E}(X^n) = \frac{1}{\mathtt{i}^n} \left.\frac{\mathrm{d}^n \phi_X(\xi)}{\mathrm{d}\xi^n}\right|_{\xi=0}.$$

(v) the function ϕ is uniformly continuous on $\mathbb{R}$.

The characteristic function therefore completely characterises the distribution. We finish this part with the following fundamental theorem:

Theorem 4.1.14. *Let X be a random variable with distribution F and let $\phi : \mathbb{R} \to \mathbb{C}$ be its characteristic function. If $\int_{\mathbb{R}} |\phi(\xi)|\mathrm{d}\xi < \infty$, then X admits a density f and*

(i) $f(x) = (2\pi)^{-1} \int_{\mathbb{R}} \mathrm{e}^{-\mathtt{i}x\xi}\phi(\xi)\mathrm{d}\xi$;

(ii) $f(x) = F'(x)$ for all $x \in \mathbb{R}$;

(iii) f is uniformly continuous on $\mathbb{R}$.

Remark 4.1.15. For notational simplicity, we have considered here random variables on the real line. All the definitions and properties above generalise directly to the multi-dimensional case. For instance if X is a random variables in $\mathbb{R}^n$ for some $n \in \mathbb{N}$, then we define its characteristic function $\phi_X : \mathbb{R}^n \to \mathbb{C}$ by

$$\phi_X(\xi) := \mathbb{E}\left(\mathrm{e}^{\mathtt{i}\langle \xi, X\rangle}\right), \qquad \text{for all } \xi \in \mathbb{R}^n,$$

where $\langle\cdot,\cdot\rangle$ represents the Euclidean inner product on $\mathbb{R}^n$.

4.1.3 Examples

We review here some basic models used in financial modelling from the point of view of their characteristic functions. The Black-Scholes model is the fundamental model for the dynamics of stock price processes, but assumes continuous paths. However, such a feature is not always realistic and jumps have to be introduced, for instance via Poisson processes.

Black-Scholes

The first example that comes to mind is obviously the Black-Scholes model. In this model, the random variable X_t—representing the logarithm of the stock price—satisfies

$$X_t = X_0 + \left(r - \frac{\sigma^2}{2}\right) t + \sigma W_t,$$

for any $t \geq 0$, where W_t is equal in law to $\sqrt{t}\mathcal{N}(0,1)$. This implies that X_t is Gaussian with mean $X_0 + \left(r - \sigma^2/2\right) t$ and variance $\sigma^2 t$, and hence

$$\begin{aligned} \phi_{X_t}(\xi) := \mathbb{E}\left(\mathrm{e}^{\mathtt{i}\xi X_t}\right) &= \exp\left(\mathtt{i}\xi\left(X_0 + \left(r - \frac{\sigma^2}{2}\right)t\right)\right)\mathbb{E}\left(\mathrm{e}^{\mathtt{i}\xi\sigma W_t}\right) \\ &= \exp\left(\mathtt{i}\xi X_0 + \left(r - \frac{\sigma^2}{2}\right)\mathtt{i}\xi t - \frac{\sigma^2\xi^2 t}{2}\right). \end{aligned}$$

Poisson processes

Another popular model—mainly used as a building block for other processes—in finance is the Poisson process $(N_t)_{t\geq 0}$. It is a counting process, in the sense that at time $t \geq 0$, N_t represents the number of events that have happened up to time t. Such a process has independent increments and is such that for each $t \geq 0$, the random variable N_t is Poisson distributed with parameter λt for $\lambda > 0$ (the intensity), i.e.

$$\mathbb{P}\left(N_t = n\right) = \frac{(\lambda t)^n}{n!}\mathrm{e}^{-\lambda t}, \quad \text{for any } n = 0, 1, \ldots$$

It is easy to compute its characteristic function

$$\phi_{N_t}(\xi) := \left(\mathrm{e}^{\mathtt{i}\xi N_t}\right) = \sum_{n\geq 0} \frac{(\lambda t)^n}{n!}\mathrm{e}^{-\lambda t}\mathrm{e}^{\mathtt{i}\xi n} = \exp\left(\lambda t\left(\mathrm{e}^{\mathtt{i}\xi} - 1\right)\right).$$

Note that the paths of a Poisson process are non-decreasing and discontinuous.

Compound Poisson processes

As we mentioned, Poisson processes are mainly used as building blocks. In particular they are the main ingredients in compound Poisson processes. Such a process $(X_t)_{t\geq 0}$ is defined as

$$X_t := \sum_{k=1}^{N_t} Z_k,$$

where (N_t) is a Poisson process with parameter λt and $(Z_k)_{k\geq 0}$ is a family of independent and identically distributed random variables with common law F. It is then easy to see that

$$\begin{aligned}
\phi_{X_t}(\xi) := \mathbb{E}\left(\mathrm{e}^{\mathtt{i}\xi X_t}\right) = \mathbb{E}\left(\exp\left(\mathtt{i}\xi\sum_{k=1}^{N_t} Z_n\right)\right) &= \mathbb{E}\left(\mathbb{E}\left(\mathrm{e}^{\mathtt{i}\xi\sum_{k=1}^{n} Z_k}\right)\Big| N_t = n\right) \\
&= \sum_{n\geq 0}\mathbb{E}\left(\mathrm{e}^{\mathtt{i}\xi\sum_{k=1}^{n} Z_k}\right)\frac{(\lambda t)^n}{n!}\mathrm{e}^{-\lambda t} = \sum_{n\geq 0}\left(\int_{\mathbb{R}}\mathrm{e}^{\mathtt{i}\xi z}F(\mathrm{d}z)\right)^n\frac{(\lambda t)^n}{n!}\mathrm{e}^{-\lambda t} \\
&= \exp\left(\lambda t\int_{\mathbb{R}}\left(\mathrm{e}^{\mathtt{i}\xi z}-1\right)F(\mathrm{d}z)\right).
\end{aligned}$$

Note that the paths of a compound Poisson process are clearly not continuous.

Affine processes

In [20] and [21], Duffie and co-authors introduced a general class of stochastically continuous, time-homogeneous Markov processes, called affine processes, in finance. These are characterised through their characteristic functions, and include in particular jumps as well as stochastic volatility. We consider here a two-dimensional version: (X, V), where $X := \log(S)$ denotes the logarithm of the stock price and V its instantaneous variance. We assume that there exist two functions φ, ψ : $\mathbb{R}_+ \times \mathbb{C}^2$ such that

$$\Phi_t(u,w) := \log\mathbb{E}\left(\mathrm{e}^{uX_t+wV_t}|X_0,V_0\right) = \varphi(t,u,w) + V_0\psi(t,u,w) + uX_0.$$

Note that 'affine' refers to the fact that the term on the right-hand side is linear (affine) in the state variable. The following theorem is proved in [20] and [42]:

Theorem 4.1.16. *Assume that $|\varphi(\tau,u,\eta)|$ and $|\psi(\tau,u,\eta)|$ are both finite for some $(\tau,u,\eta) \in \mathbb{R}_+\times\mathbb{C}^2$, then, for all $t\in[0,\tau]$ and $w\in\mathbb{C}$ such that $\Re(w)\leq\Re(\eta)$, $|\varphi(t,u,w)|$ and $|\psi(t,u,w)|$ are again finite, and the derivatives*

$$F(u,w) := \lim_{t\downarrow 0}\partial_t\varphi(t,u,w) \qquad \textit{and} \qquad R(u,w) := \lim_{t\downarrow 0}\partial_t\psi(t,u,w)$$

exist, and the functions φ and ψ satisfy the following system of (Riccati) equations:

$$\begin{aligned}
\partial_t\varphi(t,u,w) &= F(u,\psi(t,u,w)), \quad \varphi(0,u,w) = 0, \\
\partial_t\psi(t,u,w) &= R(u,\psi(t,u,w)), \quad \psi(0,u,w) = w.
\end{aligned}$$

Consider for example the case of an exponential Lévy model, for which

$$\varphi(t,u,w) \equiv \phi(u)t \qquad \text{and} \qquad \psi(t,u,w) \equiv 0,$$

the Black-Scholes case (without interest rates, say) being the simplest example with $\phi(u) \equiv \frac{1}{2}u(u-1)\sigma^2$. Then, clearly $F\equiv\phi$ and $R\equiv 0$, and the Riccati equations are trivially solved.

In the course of this chapter, and throughout some examples we will encounter other processes with well-defined characteristic functions that are available in closed form. Note in passing that this approach via characteristic functions allows us to easily combine processes. We can for instance consider the Black-Scholes model for the dynamics of the stock price process and add an independent compound Poisson process. The paths of the stock price are therefore not continuous any more, and these jumps can reflect financial decisions (such as dividends) or unexpected reactions of financial markets to political events.

Before moving on to pricing with characteristic functions, let us issue a warning. Consider the extension of the characteristic function ϕ_X to the complex plane. By definition, the latter is continuous. However, the continuity property does not in general extend from the real line to the complex plane. The most obvious (counter)examples of this are the log and the square root functions, which exhibit branch cut phenomena, e.g. which become multivalued in the complex plane.

4.2 Pricing using characteristic functions

Most models do not allow for a closed-form representation of the density of the stock price process. Black-Scholes does, but as soon as one deviates from this basic assumption, densities are not available any more. This implies that pricing directly using the density is not possible. Likewise, as soon as the dynamics of the process becomes refined and complex, PDE methods may not be available any longer (or at least not in the form we studied them before). However, the characteristic function is sometimes available, and is in particular so in mathematical finance for a large class of models, namely *affine models*, which incorporate jumps and stochastic volatility. We shall not delve into the theory of affine processes here, but we do bear in mind though that most of the tools presented here are applicable to this class of models.

4.2.1 The Black-Scholes formula revisited

This section provides a motivational example for the (inverse) Fourier transform approach. We already determined the fair value of a European Call (or Put) option in the Black-Scholes model in Chapter 1, and we provided numerical schemes to compute it. We now reformulate this option pricing problem in terms of Fourier transform, which shall lay the grounds to develop the method for more advanced models. We shall assume that a risk-neutral probability $\mathbb{P}$ is given, and we denote $\mathbb{E}^{\mathbb{P}}$ the expectation under this probability. The option pricing problem at time zero (with maturity $T > 0$ and strike $K > 0$) written on the underlying S (and correspondingly $X := \log(S)$)

then reads

$$\begin{aligned}
C_T(k) &= \mathrm{e}^{-rT}\mathbb{E}^{\mathbb{P}}\left(S_T - K\right)_+ \\
&= \mathrm{e}^{-rT}\int_k^{\infty}\left(\mathrm{e}^x - K\right)q_T(x)\mathrm{d}x \\
&= \mathrm{e}^{-rT}\int_k^{\infty}\mathrm{e}^x q_T(x)\mathrm{d}x - K\mathrm{e}^{-rT}\int_k^{\infty}q_T(x)\mathrm{d}x \qquad (4.2.1)
\end{aligned}$$

where q_T represents the density of the (Gaussian distributed) logarithmic stock price X_T at time T, and $k := \log(K)$. If ϕ_T denotes the characteristic function of the density q_T, we can write

$$\phi_T(\xi) := \int_{\mathbb{R}}\mathrm{e}^{\mathtt{i}\xi x}q_T(x)\mathrm{d}x.$$

Let us define

$$\Pi_2 := \int_k^{\infty}q_T(x)\mathrm{d}x = \int_k^{\infty}\frac{1}{2\pi}\int_{\mathbb{R}}\mathrm{e}^{-\mathtt{i}x\xi}\phi_T(\xi)\mathrm{d}\xi\mathrm{d}x.$$

It is clear that Π_2 is a probability and is precisely equal to $\mathbb{P}\left(X_T \geq k\right) = \mathbb{P}\left(S_T \geq K\right)$. The following theorem allows us to express this probability in terms of the characteristic function ϕ_T.

Theorem 4.2.1 (Gil-Pelaez inversion theorem [29]). *If F is a one-dimensional distribution function and $\phi : \xi \mapsto \int_{\mathbb{R}}\mathrm{e}^{\mathtt{i}\xi x}F(\mathrm{d}x)$ its characteristic function, then we have the inverse formula for any continuity point x of F:*

$$F(x) = \mathbb{P}(X \leq x) = \frac{1}{2} - \frac{1}{2\pi}\int_{\mathbb{R}}\frac{\mathrm{e}^{-\mathtt{i}x\xi}\phi(\xi)}{\mathtt{i}\xi}\mathrm{d}\xi.$$

Proof. Let us first consider the step function

$$\begin{aligned}
\mathrm{sgn}(y-x) &:= \begin{cases} -1 & \text{if } y < x, \\ 0 & \text{if } y = x, \\ 1 & \text{if } y > x \end{cases} \\
&= \frac{2}{\pi}\int_0^{\infty}\frac{\sin\left((y-x)\xi\right)}{\xi}\mathrm{d}\xi.
\end{aligned}$$

Note further that we have $\int_{\mathbb{R}}\mathrm{sgn}(y-x)F(\mathrm{d}y) = 1 - 2F(x)$. Let now ε and α be two strictly positive real numbers. We can then write

$$\begin{aligned}
\frac{1}{\pi}\int_{\varepsilon}^{\alpha}\frac{\mathrm{e}^{\mathtt{i}x\xi}\phi(-\xi) - \mathrm{e}^{-\mathtt{i}x\xi}\phi(\xi)}{\mathtt{i}\xi}\mathrm{d}\xi &= \frac{1}{\pi}\int_{\varepsilon}^{\alpha}\int_{\mathbb{R}}\frac{\mathrm{e}^{-\mathtt{i}\xi(y-x)} - \mathrm{e}^{\mathtt{i}\xi(y-x)}}{\mathtt{i}\xi}F(\mathrm{d}y)\mathrm{d}\xi \\
&= -\frac{2}{\pi}\int_{\varepsilon}^{\alpha}\int_{\mathbb{R}}\frac{\sin\left((y-x)\xi\right)}{\xi}F(\mathrm{d}y)\mathrm{d}\xi \\
&= -\int_{\mathbb{R}}\frac{2}{\pi}\int_{\varepsilon}^{\alpha}\frac{\sin\left((y-x)\xi\right)}{\xi}\mathrm{d}\xi F(\mathrm{d}y),
\end{aligned}$$

where the last line follows from Fubini's theorem (see Theorem A.3.1 in Appendix A.3). We now let ε tend to zero and α tend to infinity, and we obtain

$$\begin{aligned}
\frac{1}{\pi}\int_0^{\infty}\frac{\mathrm{e}^{\mathtt{i}x\xi}\phi(-\xi) - \mathrm{e}^{-\mathtt{i}x\xi}\phi(\xi)}{\mathtt{i}\xi}\mathrm{d}\xi &= -\int_{\mathbb{R}}\lim_{\varepsilon\downarrow 0,\alpha\uparrow\infty}\frac{2}{\pi}\int_{\varepsilon}^{\alpha}\frac{\sin\left((y-x)\xi\right)}{\xi}\mathrm{d}\xi F(\mathrm{d}y) \\
&= -\int_{\mathbb{R}}\mathrm{sgn}\left(y-x\right)F(\mathrm{d}y) = 2F(x) - 1.
\end{aligned}$$

A change of variable concludes the proof. □

It is easy to show that

$$\Re\left(\phi_T(\xi)\right) = \frac{\phi_T(\xi) + \phi_T(-\xi)}{2}, \quad \text{and} \quad \Im\left(\phi(\xi)\right) = \frac{\phi_T(\xi) - \phi_T(-\xi)}{2\mathtt{i}},$$

so that the function ϕ is even in its real part and odd in its imaginary part. Recalling the fact that Π_2 is a complementary probability, Gil-Pelaez inversion theorem and the symmetry properties of ϕ_T imply that

$$\Pi_2 = \frac{1}{2} + \frac{1}{2\pi}\int_{\mathbb{R}} \frac{\mathrm{e}^{-\mathtt{i}k\xi}\phi_T(\xi)}{\mathtt{i}\xi}\mathrm{d}\xi = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \Re\left(\frac{\mathrm{e}^{-\mathtt{i}k\xi}\phi_T(\xi)}{\mathtt{i}\xi}\right)\mathrm{d}\xi.$$

Concerning the first integral in (4.2.1), let us introduce a new probability measure $\widetilde{\mathbb{P}}$ by

$$\frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}} := \frac{S_T}{\mathbb{E}(S_T)}.$$

We can therefore write

$$\mathbb{E}^{\widetilde{\mathbb{P}}}\left(\mathrm{e}^{\mathtt{i}\xi X_T}\right) = \frac{\mathbb{E}^{\mathbb{P}}\left(\mathrm{e}^{X_T}\mathrm{e}^{\mathtt{i}\xi X_T}\right)}{\mathbb{E}\left(\mathrm{e}^{X_T}\right)} = \frac{\phi_T(\xi - \mathtt{i})}{\phi_T(-\mathtt{i})}.$$

Furthermore

$$\mathrm{e}^{-rT}\int_k^\infty \mathrm{e}^x q_T(x)\mathrm{d}x = \mathrm{e}^{-rT}\int_k^\infty \mathrm{e}^{X_0} q_T(x)\left(\frac{\mathrm{e}^x}{\mathrm{e}^{X_0}}\right)\mathrm{d}x = S_0\widetilde{P}\left(X_t \geq k\right) =: S_0\Pi_1,$$

where we have used the fact that the stock price process $(S_t)_{t\geq 0}$ is a martingale, i.e. $\mathbb{E}(S_T) = \mathrm{e}^{rT}S_0$. Using again Gil-Pelaez inversion theorem for Π_1 and the symmetry properties of ϕ_X, we obtain

$$\Pi_1 = \frac{1}{2} + \frac{1}{2\pi}\int_{\mathbb{R}} \frac{\phi_T(\xi - \mathtt{i})}{\mathtt{i}\xi\phi_T(-\mathtt{i})}\mathrm{e}^{-\mathtt{i}\xi k}\mathrm{d}\xi = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \Re\left(\frac{\phi_T(\xi - \mathtt{i})}{\mathtt{i}\xi\phi_T(-\mathtt{i})}\mathrm{e}^{-\mathtt{i}\xi k}\right)\mathrm{d}\xi.$$

We can therefore rewrite the option pricing problem as $C_T(k) = S_0\Pi_1 - K\mathrm{e}^{-rT}\Pi_2$. This representation expresses the option price as a difference of two probabilities, each under a different probability measure: the risk-neutral one $\mathbb{P}$ and the so-called *Share measure* $\widetilde{\mathbb{P}}$. Note further that the event considered is the probability of ending *in the money*: $S_T \geq K$. In order to derive this representation, we have only assumed (i) that the stock price was a true martingale and (ii) that it had a density. This representation therefore extends beyond the simple Black-Scholes model to any model satisfying these two conditions.

4.2.2 Option pricing with characteristic functions

We consider as above a European call option $C_T(k)$, with strike $K > 0$ and maturity $T > 0$ written on an underlying stock price process $(S_t)_{t\geq 0}$. Note that we also use (interchangeably) the notation $C_T(k)$, with $k := \log(K)$, but this should not create any confusion. The approach outlined in the previous subsection is appealing since it only requires the knowledge of the characteristic function of the (logarithmic) stock price process. However, it requires two separate integrations. A now popular approach developed by Peter Carr and Dilip Madan [10] is somehow a reformulation of the above derivation and leads to a pricing formula involving a single integral. As before, we can write the Call option price as

$$C_T(k) = \mathrm{e}^{-rT}\mathbb{E}^{\mathbb{P}}\left((S_T - K)_+\right) = \mathrm{e}^{-rT}\int_k^\infty \left(\mathrm{e}^x - \mathrm{e}^k\right) q_T(x)\mathrm{d}x$$

From this, it is easy to see that

$$\lim_{k\downarrow-\infty} C_T(k) = \lim_{k\downarrow-\infty} \mathrm{e}^{-rT}\int_k^\infty \left(\mathrm{e}^x - \mathrm{e}^k\right) q_T(x)\mathrm{d}x = \mathrm{e}^{-rT}\mathbb{E}^{\mathbb{P}}\left(\mathrm{e}^{X_T}\right) = S_0,$$

so that the call price function $k \mapsto C_T(k)$ is not in $L^1(\mathbb{R})$, and its Fourier transform does not exist. Let α be a strictly positive real number and define the *dampened* call option price $c_T(k) := \mathrm{e}^{\alpha k}C_T(k)$, such that the function c_T is integrable (in $L^1(\mathbb{R})$). With ψ denoting its Fourier transform, we then have

$$\begin{aligned}\psi(\xi) &:= \int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}\xi k} c_T(k)\mathrm{d}k = \int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}\xi k}\mathrm{e}^{\alpha k}\mathrm{e}^{-rT}\left(\int_k^\infty \left(\mathrm{e}^x - \mathrm{e}^k\right) q_T(x)\mathrm{d}x\right)\mathrm{d}k \\ &= \int_{\mathbb{R}} \mathrm{e}^{-rT} q_T(x)\left(\int_{-\infty}^x \left(\mathrm{e}^x - \mathrm{e}^k\right)\mathrm{e}^{(\alpha+\mathtt{i}\xi)k}\mathrm{d}k\right)\mathrm{d}x, \qquad (4.2.2)\end{aligned}$$

where again we have used Fubini's theorem on the second line. Now

$$\begin{aligned}\int_{-\infty}^x \left(\mathrm{e}^x - \mathrm{e}^k\right)\mathrm{e}^{(\alpha+\mathtt{i}\xi)k}\mathrm{d}k &= \mathrm{e}^x\int_{-\infty}^x \mathrm{e}^{(\alpha+\mathtt{i}\xi)k}\mathrm{d}k - \int_{-\infty}^x \mathrm{e}^{(1+\alpha+\mathtt{i}\xi)k}\mathrm{d}k \\ &= \frac{\mathrm{e}^x}{\alpha+\mathtt{i}\xi}\left[\mathrm{e}^{(\alpha+\mathtt{i}\xi)k}\right]_{-\infty}^x - \frac{1}{1+\alpha+\mathtt{i}\xi}\left[\mathrm{e}^{(1+\alpha+\mathtt{i}\xi)k}\right]_{-\infty}^x \\ &= \frac{\mathrm{e}^{(\alpha+1+\mathtt{i}\xi)x}}{(\alpha+\mathtt{i}\xi)(1+\alpha+\mathtt{i}\xi)},\end{aligned}$$

where we have used the fact that $\alpha > 0$ implies that $\lim_{k\downarrow-\infty}\mathrm{e}^{(\alpha+\mathtt{i}\xi)k} = 0$. So that (4.2.2) becomes

$$\psi(\xi) = \int_{\mathbb{R}} \mathrm{e}^{-rT} q_T(x)\frac{\mathrm{e}^{(\alpha+1+\mathtt{i}\xi)x}}{(\alpha+\mathtt{i}\xi)(1+\alpha+\mathtt{i}\xi)}\mathrm{d}x = \mathrm{e}^{-rT}\frac{\phi_T(\xi-(\alpha+1)\mathtt{i})}{(\alpha+\mathtt{i}\xi)(1+\alpha+\mathtt{i}\xi)}.$$

The final step is to express the dampened call price as the inverse Fourier transform of the function ψ, and to turn it into the original call price function. More precisely, we have

$$\begin{aligned}C_T(k) = \mathrm{e}^{-\alpha k}c_T(k) &= \frac{\mathrm{e}^{-\alpha k}}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\mathtt{i}\xi k}\psi(\xi)\mathrm{d}\xi \\ &= \frac{\mathrm{e}^{-\alpha k}}{\pi}\int_0^\infty \mathrm{e}^{-\mathtt{i}\xi k}\psi(\xi)\mathrm{d}\xi, \qquad (4.2.3)\end{aligned}$$

where again we have used the symmetry properties of the function ϕ_T.

Remark 4.2.2. One might be tempted to take any $\alpha > 0$ so that the dampened call option price decays fast enough as k tends to infinity. However, if one chooses $\alpha > 0$ too large, then the dampened option price might not be integrable any longer on the positive half-axis, i.e. when k tends to infinity. We shall not delve into these subtleties here, but simply mention that an upper bound α_+ for α has been provided by Roger Lee [46] and is given by

$$\alpha_+ := \sup\left\{\alpha > 0 : \mathbb{E}^{\mathbb{P}}\left(S_T^{\alpha+1}\right) < \infty\right\}.$$

Similar considerations can be made for European Put options by symmetry.

The above result, though very useful in practice, is however limited to the case of vanilla Call options. We now investigate an extension of it to more general European payoff functions. We shall denote by h a payoff function, and by H the value of the option at inception of the contract: $H(T) := \mathbb{E}[h(X_T)]$, where we assume for simplicity that interest rates are null. As mentioned above, integrability of the payoff is a necessary property, and we therefore introduce the dampened payoff function h_α ($\alpha \in \mathbb{R}$) defined by $h_\alpha(x) \equiv \mathrm{e}^{-\alpha x} h(x)$. In particular, for a Call option, one needs to consider $\alpha > 1$, and for a Put option, $\alpha < 0$. We shall state and prove below two results, depending on whether the payoff function h is continuous or not.

Theorem 4.2.3. *Suppose that there exists $\alpha \in \mathbb{R}$ such that both h_α and $\hat{h}_\alpha$ are integrable, and that $\mathbb{E}(S_T^\alpha)$ is finite, then*

$$H(T) = \frac{1}{\pi} \int_0^{+\infty} \phi_T(-(\xi + \mathtt{i}\alpha)) \hat{h}_\alpha(\xi) \mathrm{d}\xi.$$

Proof. We first prove that the integral in the theorem is well defined. Note that

$$\hat{h}(\xi + \mathtt{i}\alpha) = \int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}(\xi+\mathtt{i}\alpha)x} h(x) \mathrm{d}x = \int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}\xi x} \mathrm{e}^{-\alpha x} h(x) \mathrm{d}x = \int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}\xi x} h_\alpha(x) \mathrm{d}x = \hat{h}_\alpha(\xi).$$

Furthermore,

$$|\phi_T(-(\xi + \mathtt{i}\alpha))| \leq \left| \int_{\mathbb{R}} \mathrm{e}^{-\mathtt{i}(\xi+\mathtt{i}\alpha)x} q_T(x) \mathrm{d}x \right| \leq \int_{\mathbb{R}} \mathrm{e}^{\alpha x} q_T(x) \mathrm{d}x = \mathbb{E}(S_T^\alpha),$$

which is finite by assumption. Now,

$$\begin{aligned} H(T) = \mathbb{E}[h(X_T)] &= \int_{\mathbb{R}} \mathrm{e}^{\alpha x} h_\alpha(x) q_T(x) \mathrm{d}x = \int_{\mathbb{R}} \mathrm{e}^{\alpha x} \left(\frac{1}{\pi} \int_0^{+\infty} \mathrm{e}^{-\mathtt{i}x\xi} \hat{h}_\alpha(\xi) \mathrm{d}\xi \right) q_T(x) \mathrm{d}x \\ &= \frac{1}{\pi} \int_0^{+\infty} \left(\int_{\mathbb{R}} \mathrm{e}^{\alpha x} \mathrm{e}^{-\mathtt{i}x\xi} q_T(x) \mathrm{d}x \right) \hat{h}_\alpha(\xi) \mathrm{d}\xi \\ &= \frac{1}{\pi} \int_0^{+\infty} \phi_T(-(\xi + \mathtt{i}\alpha)) \hat{h}_\alpha(\xi) \mathrm{d}\xi, \end{aligned}$$

which concludes the proof. The second line follows by the Fourier inversion formula (since $\hat{h}_\alpha \in L^1(\mathbb{R})$ by assumption), and the third line by Fubini. □

Example. If h is the payoff of a European Call option, then

$$\hat{h}_\alpha^{\mathrm{Call}}(\xi) = \frac{\mathrm{e}^{(\mathtt{i}\xi+1-\alpha)k}}{(\mathtt{i}\xi - \alpha)(\mathtt{i}\xi - \alpha + 1)},$$

and $\hat{h}_\alpha^{\mathrm{Call}}(\xi) \equiv \hat{h}_\alpha^{\mathrm{Put}}(\xi)$, albeit with different restrictions on ξ.

We now investigate the case where the payoff function h is discontinuous.

Theorem 4.2.4. *Assume that h_α is integrable, that $\mathbb{E}(S_T^\alpha)$ is finite, and that the map $x \mapsto \mathbb{E}[h(X_T + x)]$ is continuous around $x = X_0 = 0$ and has bounded variation in a neighbourhood of X_0. Then*

$$H(T) = \frac{1}{\pi} \lim_{R \uparrow +\infty} \int_0^R \phi_T(-(\xi + \mathtt{i}\alpha)) \hat{h}_\alpha(\xi + \mathtt{i}\alpha) \mathrm{d}\xi.$$

The proof of the theorem is left as an exercise, and follows analogous lines to that of Theorem 4.2.3, together with Theorem 4.1.12.

Example. The following examples of payoffs are not continuous:

- digital option: $h(x) = \mathbb{1}_{\{\mathrm{e}^x \geq k\}}$ and $\hat{h}(\xi) = -\frac{\mathrm{e}^{(\mathtt{i}\xi-\alpha)k}}{\mathtt{i}\xi-\alpha}$, $\alpha > 0$;
- asset-or-nothing option: $h(x) = \mathrm{e}^x\mathbb{1}_{\{\mathrm{e}^x \geq k\}}$ and $\hat{h}(\xi) = -\frac{\mathrm{e}^{(1+\mathtt{i}\xi-\alpha)k}}{1+\mathtt{i}\xi-\alpha}$, $\alpha > 1$;
- double digital option ($k_1 < k_2$): $h(x) = \mathbb{1}_{\{k_1 \leq \mathrm{e}^x \leq k2\}}$ and $\hat{h}(\xi) = \frac{\mathrm{e}^{(\mathtt{i}\xi-\alpha)k_2} - \mathrm{e}^{(\mathtt{i}\xi-\alpha)k_1}}{\mathtt{i}\xi-\alpha}$, $\alpha \neq 0$;
- self-quanto option: $h(x) = \mathrm{e}^x(\mathrm{e}^x - \mathrm{e}^k)_+$ and $\hat{h}(\xi) = \frac{\mathrm{e}^{(2+\mathtt{i}\xi-\alpha)k}}{(1+\mathtt{i}\xi-\alpha)(2+\mathtt{i}\xi-\alpha)}$, $\alpha > 2$;
- power option: $h(x) = [(\mathrm{e}^x - \mathrm{e}^k)_+]^2$ and $\hat{h}(\xi) = \frac{2\mathrm{e}^{(2+\mathtt{i}\xi-\alpha)k}}{(\mathtt{i}\xi-\alpha)(1+\mathtt{i}\xi-\alpha)(2+\mathtt{i}\xi-\alpha)}$, $\alpha > 2$;

A note on bond pricing

In the context of bond pricing, one does not immediately need the characteristic function of the process itself, but it does come into play. Let $(r_t)_{t\geq 0}$ denote the instantaneous short rate process. Then the bond price with maturity T is given by $\mathbb{E}\left(\mathrm{e}^{-\int_0^T r_t \mathrm{d}t}\right)$, namely the characteristic function, evaluated at the point $\mathtt{i}$, of the integrated rate process $R_t := \int_0^t r_s \mathrm{d}s$. Consider the Vasicek model [54], introduced in 1977, under which the instantaneous sport interest rate follows an Ornstein-Uhlenbeck process, namely we consider the unique strong solution to the stochastic differential equation

$$\mathrm{d}r_t = \kappa(\theta - r_t)\mathrm{d}t + \xi \mathrm{d}W_t, \qquad r_0 > 0, \tag{4.2.4}$$

where W is a standard Brownian motion and κ, θ, ξ and r_0 strictly positive real numbers. It is immediate to see that, for any $t \geq 0$,

$$r_t = r_0\mathrm{e}^{-\kappa t} + \theta\left(1 - \mathrm{e}^{-\kappa t}\right) + \xi\int_0^t \mathrm{e}^{-\kappa(t-s)}\mathrm{d}W_s,$$

and

$$\mathbb{E}(r_t) = r_0\mathrm{e}^{-\kappa t} + \theta\left(1 - \mathrm{e}^{-\kappa t}\right) \qquad \text{and} \qquad \mathbb{V}(r_t) = \frac{\xi^2}{2\kappa}\left(1 - \mathrm{e}^{-2\kappa t}\right).$$

Straightforward computations show that, for any $t \geq 0$, the integrated rate R_t is Gaussian with mean and variance given by

$$\mathbb{E}(R_t) = \theta t + \frac{\theta - r_0}{\kappa}\left(\mathrm{e}^{-\kappa t} - 1\right) \qquad \text{and} \qquad \mathbb{V}(R_t) = \frac{\xi^2}{2\kappa^2}\left(\frac{4\mathrm{e}^{-\kappa t} - \mathrm{e}^{-2\kappa t} - 3}{2\kappa} + t\right),$$

so that $\mathbb{E}\left(\mathrm{e}^{\mathtt{i}uR_t}\right) = \exp\left(A(u,t) - B(u,t)r_0\right)$, where

$$\begin{cases} A(u,t) &= \left(\theta + \dfrac{\mathtt{i}\xi^2 u}{2\kappa^2}\right)(B(u,t) + \mathtt{i}u) - \dfrac{\xi^2}{4\kappa}B(u,t)^2, \\ B(u,t) &= \dfrac{\mathrm{e}^{-\kappa t} - 1}{\kappa}\mathtt{i}u. \end{cases}$$

4.3 Pricing via saddlepoint approximation

4.3.1 The Lugannani-Rice approximation

The Gaussian base

We consider here a continuous random variable X, taking values on the real line. We shall denote $M(u) := \mathbb{E}(\mathrm{e}^{uX})$ its moment generating function and $\Lambda(u) := \log M(u)$ its cumulant generating function. Obviously these two functions are only defined on some subset of the real line (including the origin), which is called the effective domain: $\mathcal{D}_X := \{u \in \mathbb{R} : M(u) < \infty\}$. The function Λ is differentiable on $\mathcal{D}_X$ and convex by Jensen's inequality. Therefore, for any $x \in \mathcal{D}'_X := \Lambda'(\mathcal{D}_X)$ (the image of $\mathcal{D}_X$ by Λ'), the equation $\Lambda'(s) = x$ has a unique solution in $\mathcal{D}_X$. We start with the following saddlepoint approximation theorem. We recall that $\mathcal{N}$ and n respectively denote the Gaussian cumulative distribution function and Gaussian density.

Theorem 4.3.1 (Lugannani and Rice [48]). *Let X be a continuous random variable on $\mathbb{R}$, then*

$$\mathbb{P}(X \geq x) \approx \begin{cases} 1 - \mathcal{N}(\hat{\omega}) + n(\hat{\omega}) \left(\dfrac{1}{\hat{u}} - \dfrac{1}{\hat{\omega}} \right), & \text{if } x \neq \mathbb{E}(X), \\ \dfrac{1}{2} - \dfrac{\Lambda'''(0)}{6\Lambda''(0)^{3/2}\sqrt{2\pi}}, & \text{if } x = \mathbb{E}(X), \end{cases}$$

where $\hat{\omega} := \operatorname{sgn}(\hat{s}(x))\sqrt{2[\hat{s}(x)x - \Lambda(\hat{s}(x))]}$, $\hat{u} := \hat{s}(x)\sqrt{\Lambda''(\hat{s}(x))}$, $\operatorname{sgn}(z) := 1\!\!1_{\{z>0\}} - 1\!\!1_{\{z<0\}}$, *and* $\hat{s}(x)$ *is the unique solution to* $\Lambda'(s) = x$ *in* $\mathcal{D}_X$.

Remark 4.3.2. Strictly speaking, Lugannani and Rice considered the mean $\overline{X}_n := n^{-1} \sum_{i=1}^n X_i$ of independent and identically distributed random variables $(X_i)_{i\in\mathbb{N}}$ under the condition that X_1 admits a density. They then proved that, as n tends to infinity, the expansion

$$\mathbb{P}(\overline{X}_n \geq x) = 1 - \mathcal{N}(\hat{\omega}) + n(\hat{\omega}) \left(\frac{1}{\hat{\omega}} - \frac{1}{\hat{u}} \right) + \mathcal{O}\left(n^{-3/2}\right)$$

holds. The approximation in Theorem 4.3.1 is thus obtained by considering the case $n = 1$. It is remarkable, as we shall see, that this approximation remains extremely accurate.

Example. Consider the standard Gaussian distribution, then $\hat{s} = \hat{w} = \hat{u} = x$, and the saddlepoint approximation is exact.

Example. The Gamma distribution with shape parameter $\alpha > 0$ and rate $\beta > 0$ has density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}, \qquad \text{for all } x > 0,$$

so that $M(u) = \left(1 - \frac{u}{\beta}\right)^{-\alpha}$, for any $u < \beta$, and therefore, for any $x > 0$,

$$\hat{s} = \beta - \frac{\alpha}{x}, \qquad \hat{w} = \operatorname{sgn}(\hat{s}(x))\sqrt{2\left(\beta x - \alpha\left[1 + \log\left(\frac{\beta x}{\alpha}\right)\right]\right)}, \qquad \hat{u} = \frac{\beta x - \alpha}{\sqrt{\alpha}}.$$

Exercise 44. Check the computations of the Gamma example above (Example 4.3.1) and plot the difference between the Gamma CDF and its saddlepoint approximation given in Theorem 4.3.1. Discuss the influence of the parameters α and β on the accuracy of this approximation.

Non-Gaussian bases

Wood, Booth and Butler [57] generalised the Lugannani-Rice approximation by allowing for a non-Gaussian base. Let Z be a given random variable, which we call the base, for which the cumulant generating function $\Lambda_Z(u) \equiv \log \mathbb{E}(\mathrm{e}^{uZ})$, the CDF F_Z and the density f_Z are known. We state below their results, which obviously reduces to the Lugannani-Rice approximation (Theorem 4.3.1) when Z is Gaussian:

Theorem 4.3.3. *Let X be a real continuous random variable. Then, for any $x \in \mathbb{R}$,*

$$\mathbb{P}(X > z) \approx 1 - F_Z(\hat{\xi}) + f_Z(\hat{\xi}) \left(\frac{1}{\hat{u}_{\hat{\xi}}} - \frac{1}{\omega_{\hat{\xi}}} \right),$$

where

$$\hat{u}_{\hat{\xi}} := \hat{s}\sqrt{\frac{\Lambda''(\hat{s})}{G''(\omega_{\hat{\xi}})}},$$

and where $\hat{s}$, $\hat{\xi}$ and $\omega_{\hat{\xi}}$ are the unique solutions to

$$\Lambda'(\hat{s}) = x, \qquad H(\hat{\xi}) = \Lambda(\hat{s}) - \hat{s}x, \qquad G'(\omega_{\hat{\xi}}) = \xi;$$

here Λ is the CGF of X and $H(\xi) \equiv G(\omega_\xi) - \xi\omega_\xi$.

4.3.2 Pricing with the Lugannani-Rice approximation

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a given filtered probability space. We consider here a European Put option $P(k,T)$ with strike e^k ($k \in \mathbb{R}$) and maturity $T > 0$, on a given stock price $(\mathrm{e}^{X_t})_{t\geq 0}$, assumed to be a strictly positive martingale under $\mathbb{P}$. Then

$$P(k,T) = \mathbb{E}^{\mathbb{P}}(\mathrm{e}^k - \mathrm{e}^{X_T})_+ = \mathbb{E}^{\mathbb{P}}\left(\left(\mathrm{e}^k - \mathrm{e}^{X_T}\right)\mathbb{1}_{\{X_T<k\}}\right) = \mathrm{e}^k\mathbb{P}(X_T < k) - \mathbb{E}^{\mathbb{P}}\left(\mathrm{e}^{X_T}\mathbb{1}_{\{X_T<k\}}\right).$$

Using the 'Share measure' $\mathbb{Q}$ defined via the Radon-Nikodym derivative by $\left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_{\mathcal{F}_T} = \frac{S_T}{S_0} = \exp(X_T - X_0)$, we have

$$\mathbb{E}^{\mathbb{P}}\left(\mathrm{e}^{X_T}\mathbb{1}_{\{X_T<k\}}\right) = \mathrm{e}^{X_0}\mathbb{E}^{\mathbb{P}}\left(\mathrm{e}^{X_T-X_0}\mathbb{1}_{\{X_T<k\}}\right) = \mathrm{e}^{X_0}\mathbb{E}^{\mathbb{P}}\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\mathbb{1}_{\{X_T<k\}}\right) = \mathrm{e}^{X_0}\mathbb{E}^{\mathbb{Q}}\left(\mathbb{1}_{\{X_T<k\}}\right),$$

and therefore

$$\mathbb{P}(k,T) = \mathrm{e}^k\mathbb{P}(X_T < k) - \mathrm{e}^{X_0}\mathbb{Q}(X_T < k). \tag{4.3.1}$$

Suppose now that the cumulant generating function of X_T, $\Lambda(u) := \log \mathbb{E}^{\mathbb{P}}\left(\mathrm{e}^{uX_T}\right)$, is known, obviously only on its effective domain $\mathcal{D}_X$. The first probability in (4.3.1) can be approximated directly using the Lugannani-Rice formula in Theorem 4.3.1. Regarding the second one, let us compute the cumulant generating function, $\Lambda_{\mathbb{Q}}$ of X_T under $\mathbb{Q}$:

$$\Lambda_{\mathbb{Q}}(u) := \log \mathbb{E}^{\mathbb{Q}}\left(\mathrm{e}^{uX_T}\right) = \log \mathbb{E}\left(\mathrm{e}^{uX_T}\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right) = \mathrm{e}^{X_0}\log \mathbb{E}\left(\mathrm{e}^{(u-1)X_T}\right)$$

Example. See the IPython notebook for an implementation.

4.4 Numerical integration and quadrature methods

As we mentioned above, if either the density of the characteristic function of the stock price process is available in closed-form, then pricing European vanilla options (Calls and Puts) boils down to a simple integration over (part of) the real line of some real-valued function. We shall now see how such an integration can be performed. We will first have a look at (standard) integration schemes, which can be seen essentially as refinements of the standard Riemann integration. We will then have a look at discrete and fast Fourier methods, which have proven to be extremely efficient algorithms for the pricing methodology developed in Section 4.2.

4.4.1 A primer on polynomial interpolation

Lagrange polynomials

Polynomial interpolation lies at the basis of function approximation. The basic question can be formulated as follows: given a continuous function f on an interval $[a, b] \subset \mathbb{R}$, can we find a smooth (simple) function that approximates f in some sense? The sense in which we consider this approximation is the sup norm, i.e.

$$\|f\|_\infty := \sup_{x\in[a,b]} |f(x)|.$$

The answer to the above question is given by Weierstrass theorem, who proves that the space of polynomials is dense in the space of continuous functions (on any interval of the real line):

Theorem 4.4.1 (Weierstrass)**.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function. For any $\varepsilon > 0$, there exists a polynomial P such that $\|f - P\|_\infty \leq \varepsilon$.*

Proof. There exist several proofs in the literature. We give here a constructive proof. For simplicity consider $[a, b] = [0, 1]$, and for any $n \in \mathbb{N}$, define the polynomial P_n by

$$P_n(x) := \sum_{k=0}^{n} \binom{n}{k} f(k/n)x^k(1-x)^k, \qquad \text{for any } x \in [0, 1].$$

Since f is continuous on the closed interval $[0, 1]$, it is bounded and the sequence $(P_n)_n$ converges uniformly to f in $[0, 1]$. □

Remark 4.4.2. Note that Weierstrass theorem ensures the existence of some polynomial close enough to f in the sup norm. It can be shown, however, that the sequence proposed in the proof has poor convergence properties.

This theorem motivates our use of polynomials as interpolating functions. Note that the Weierstrass theorem indicates the existence of a polynomial that approximates a given function in the sup norm, but does not provide any information concerning its construction. Theorem 4.4.5 below gives us such information, but we need some preliminary definitions.

Definition 4.4.3. Let $n \in \mathbb{N}$. Given a set of distinct points $\{x_0, \ldots, x_n\}$, the Lagrange polynomials are the $n+1$ polynomials satisfying

$$L_i^{(n)}(x_j) := \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \qquad \text{for each } 0 \leq i, j \leq n.$$

Let now $\{y_0, \ldots, y_n\}$ be a set of $n+1$ points, we further define the *interpolating polynomial* P_n as

$$P_n(x) := \sum_{i=0}^{n} y_i L_i^{(n)}(x).$$

The Lagrange polynomials can be further characterised by

$$L_i^{(n)}(x) := \prod_{k=0, k\neq i}^{n} \frac{x - x_k}{x_i - x_k}, \qquad \text{for each } i = 0, \ldots, n.$$

Remark 4.4.4. It is clear from the definition that $P_n(x_i) = y_i$ for any $i = 1, \ldots, n$, hence the name *interpolating polynomial.* Note further that P_n is of degree at most n.

The following proposition validates the use of these Lagrange polynomials.

Theorem 4.4.5. *Let $\{x_0, \ldots, x_n\}$ be $n+1$ distinct nodes, then for any $\{y_0, \ldots, y_n\}$ there exists a polynomial P of degree at most n such that $P(x_i) = y_i$, for any $i = 0, \ldots, n$.*

Proof. The existence follows immediately from the construction of the Lagrange polynomials above. Suppose now that there exist two such interpolating polynomials P and Q of degree at most n, and define $R \equiv P - Q$. R is a polynomial of degree at most n but has (at least) $n+1$ roots, so that it must be null everywhere, and the theorem follows. □

We could give another proof of existence of these interpolating polynomials, in a more constructive way. Consider a set of pairs $\{(x_0, y_0), \ldots, (x_n, y_n)\}$. We wish to construct the interpolating polynomials iteratively. Start with the initial constant polynomial $P_0(x) := y_0$, and define the family $(P_k)_{k\geq 1}$ recursively as

$$P_{k+1}(x) := P_k(x) + \gamma_k (x - x_0) \ldots (x - x_k), \qquad \text{for each } k \geq 1,$$

where γ_k is a constant chosen such that $y_{k+1} = P_{k+1}(x_{k+1})$, i.e.

$$\gamma_k = \frac{y_{k+1} - P_k(x_{k+1})}{(x - x_0) \dots (x - x_k)}.$$

This construction is called *Newton's algorithm.* It has the advantage that given a polynomial interpolating n points, one does not have to start the Lagrange algorithm from scratch in order to obtain a polynomial interpolating $n+1$ points. Note that we can rewrite the general term of the sequence of polynomials as

$$P_n(x) = y_0 + \sum_{k=0}^{n} \gamma_k \prod_{j=0}^{k-1} (x - x_j), \qquad \text{for any } n \geq 0,$$

where an empty product is by convention equal to one.

Interpolation error

It is then natural to wonder how good—in the sense of how close from the original function f—these interpolating polynomials are. The following theorem provides an answer to that question:

Theorem 4.4.6 (Interpolation Error Theorem)**.** *Let P_n be a polynomial of degree at most $n \geq 0$ interpolating the function f at the distinct nodes $x_0, \dots, x_n$ on the interval $[a, b]$ and assume that the $(n+1)$-th derivative $f^{(n+1)}$ is continuous on this interval, then for each $x \in [a, b]$, there exists $\xi \in [a, b]$ such that*

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i).$$

Proof. The theorem is trivial when x corresponds to one of the nodes $x_0, \dots, x_n$. Assume therefore that $x \neq x_i$ for any $i = 0, \dots, n$. Define

$$\begin{aligned} \psi(z) &:= \prod_{i=0}^{n} (z - x_i), \\ \gamma &:= \frac{f(x) - P_n(x)}{\psi(x)}, \\ \phi(z) &:= f(z) - P_n(z) - \gamma\psi(z). \end{aligned}$$

Note that

(i) γ is well defined;

(ii) the function ϕ has $n+2$ roots: $x_0, \dots, x_n$ and x;

(iii) the functions f, P_n and ψ are $C^{n+1}([a, b])$, and so is ϕ.

Rolle's theorem implies that ϕ' has $n+1$ roots, that ϕ'' has n roots ad so on, so that $\phi^{(n+1)}$ has exactly one root in $[a, b]$, which we denote ξ. Differentiating ϕ $(n+1)$ times, we see that $f^{(n+1)}(\xi) - \gamma\psi^{(n+1)}(\xi) = 0$ since P_n is a polynomial of degree at most n. We also note that $\psi^{(n+1)}(\xi) = (n+1)!$ and hence, replacing γ by its definition, the theorem follows. □

We usually do not have much control over the derivative $f^{(n+1)}$, so that the only way to reduce the error is to choose the nodes adequately. The following proposition outlines the interpolation error when using equally spaced nodes.

Proposition 4.4.7. *Consider $n+1$ equally spaced nodes $x_0, \dots, x_n$ of the interval $[a, b]$, i.e. $x_i := a + i\delta$, for $i = 0, \dots, n$, where $\delta := (b-a)/n$. Then*

$$\prod_{i=0}^{n} |x - x_i| \leq \frac{\delta^{n+1} n!}{4}, \qquad \text{for any } x \in [a, b].$$

Proof. If x is situated at one of the nodes, then the proposition is obvious since the left-hand side is null. Fix some $x \in [a, b]$; then there exists $j \in [0, n-1]$ such that $x \in (x_j, x_{j+1})$. Note that either $|x - x_j| \leq \delta/2$ or $|x - x_{j+1}| \leq \delta/2$. In either case, we have the inequality

$$|x - x_j| \, |x - x_{j+1}| \leq \frac{\delta^2}{4}.$$

Straightforward calculations then show that

$$\begin{aligned} |x - x_i| &\leq (j - i + 1)\delta, \quad \text{for } i < j, \\ |x - x_i| &\leq (i - j)\delta, \quad \text{for } i > j + 1. \end{aligned}$$

Therefore

$$\prod_{i=0}^{n} |x - x_i| \leq \frac{\delta^2}{4} \left((j+1)! \delta^j \right) \left((n-j)! \delta^{n-j-1} \right) \leq \frac{\delta^{n+1} n!}{4},$$

as claimed, where we have used the fact that $(j+1)!(n-j)! \leq n!$. □

The following example (stated as an exercise) should be seen as a warning when using polynomial interpolations.

Exercise 45. Consider the so-called Runge function defined by $f(x) := (1+25x^2)^{-1}$ on the interval $[-1, 1]$. Discuss the validity of the polynomial interpolation as the number of nodes increases.

Exercise 46. How many nodes are needed in order to interpolate the function $f(x) := \sin(x) + \cos(x)$ on the interval $[0, \pi]$ with a maximum error of 10^{-8}?

These results seem to indicate that equally spaced nodes may not be the most efficient choice. Note that from Theorem 4.4.6, we can write

$$\|f(x) - P_n(x)\|_\infty \leq \frac{\left\| f^{(n+1)} \right\|_\infty}{(n+1)!} \|\omega\|_\infty,$$

where $\omega(x) \equiv \prod_{i=0}^{n} (x - x_i)$. The only term depending on the nodes is ω, and therefore it sounds sensible to try and minimise $\|\omega\|_\infty$. Consider the Chebychev polynomials defined recursively on $[-1, 1]$ by

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_{k+1}(x) &= 2x T_k(x) - T_{k-1}(x), \qquad \text{for any } k \geq 1. \end{aligned}$$

One can show that this definition implies the characterisation $\cos(k\theta) = T_k(\cos\theta)$, for any $k \geq 0$, $\theta \in \mathbb{R}$ and that for any $k \geq 2$, T_k has exactly k roots:

$$x_i^{(k)} = \cos\left(\frac{i+1/2}{k}\pi\right), \qquad \text{for any } i = 0, \ldots, k-1.$$

It is then easy to show that (exercise) for the Chebychev polynomial T_k, we have

$$\prod_{i=0}^{k} \left|x - x_i^{(k)}\right| \leq 2^{-k}, \qquad \text{for any } x \in [-1,1],$$

which is clearly an improvement of Proposition 4.4.7. The motivation underlying the use of Chebychev polynomials is the following theorem:

Theorem 4.4.8. *For any fixed integer $n > 0$, then the minimisation problem*

$$\tau_n = \inf_{\deg(Q) \leq n-1} \left\{ \max_{x\in[-1,1]} |x^n + Q(x)| \right\}$$

has a unique solution equal to $\tau_n = 2^{1-n}$ which is attained at $Q^*(x) \equiv 2^{1-n}T_n(x) - x^n$, where T_n is the n-th Chebychev polynomial.

Orthogonal polynomials

We have just seen that equally spaced nodes are not optimal, and that greater accuracy can be achieved by choosing the roots of other polynomials such as Chebychev polynomials. We make this clearer and more rigorous here. Recall that Weierstrass theorem 4.4.1 considered the approximating error in the sup norm $\|\cdot\|_\infty$. We define here a new norm, more adapted to the current problem. Let us recall the following basic facts about vector spaces:

Definition 4.4.9. Let $\mathcal{H}$ be a (real) vector space and u and v two elements of H. An inner product $\langle u, v\rangle$ is a mapping from $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$ satisfying

- $\langle u+v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$, for $w \in \mathcal{H}$;
- $\langle \alpha u, v\rangle = \alpha \langle u, v\rangle$, for any $\alpha \in \mathbb{R}$;
- $\langle u, v\rangle = \langle v, u\rangle$;
- $\langle u, u\rangle \geq 0$;
- $\langle u, u\rangle = 0$ if and only if $u = 0$.

We shall be interested here in the weighted vector space $C([a,b])$ of continuous functions on the interval $[a,b]$, with inner product

$$\langle f, g\rangle_w := \int_a^b w(x)f(x)g(x)\mathrm{d}x, \qquad \text{for any } f, g \in C([a,b]),$$

where the weight function $w : [a, b] \to \mathbb{R}$ is continuous, non-negative, and does not vanish on any subinterval of $[a, b]$ of non-zero length. The corresponding norm $\|\cdot\|_{2,w}$ is the weighted L^2-norm. Let now $f \in \mathcal{H} = C([a, b])$ (and a given weight function w) and define $\mathcal{P}$ as a finite-dimensional subspace of $\mathcal{H}$ of say, dimension $m \geq 1$. This implies that there exists a basis of linearly independent vector $(\phi_1, \ldots, \phi_m) \in \mathcal{H}^m$ such that

$$\mathcal{P} = \left\{ \sum_{k=1}^{m} \alpha_k \phi_k, \quad (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m \right\}.$$

The problem we are considering can be written as

$$\text{Find } P \in \mathcal{P} \text{ that minimises } \|f - P\|_{2,w}. \tag{4.4.1}$$

The solution to this problem is contained in the following theorem, the proof of which is omitted.

Theorem 4.4.10. *The problem* (4.4.1) *has a unique solution* $P^* \in \mathcal{P}$, *which is the unique solution to* $\langle f - P, Q \rangle_w = 0$ *for all* $Q \in \mathcal{P}$.

Definition 4.4.11. We shall say that two polynomials P and Q are orthogonal in the Hilbert weighted space $\mathcal{H}$ if $\langle P, Q \rangle_w = 0$.

Remark 4.4.12. One could for instance consider $\mathcal{P}$ as the $(n+1)$-dimensional subspace generated by the basis $(1, x, \ldots, x^n)$. In this case, the function to minimise (from the minimisation problem (4.4.1)) reads

$$\Phi(\alpha_0, \ldots, \alpha_n) := \int_a^b w(x) \left(f(x) - \sum_{i=0}^{n} \alpha_i x^i \right)^2 \mathrm{d}x.$$

In order for $(\alpha_0, \ldots, \alpha_n)$ to be a minimum, we need the conditions $\partial_i \Phi(\alpha_0, \ldots, \alpha_n) = 0$, for $i = 0, \ldots, n$. This can be written as the following linear system

$$\sum_{j=0}^{n} \alpha_j \int_a^b w(x) x^{i+j} \mathrm{d}x = \int_a^b w(x) f(x) x^i \mathrm{d}x, \quad \text{for } i = 0, \ldots, n.$$

Consider the case where the weight function is constant and equal to one on $[a, b] = [0, 1]$. Then the linear system becomes

$$\sum_{j=0}^{n} \frac{\alpha_j}{i + j + 1} = \int_0^1 f(x) x^i \mathrm{d}x, \quad \text{for } i = 0, \ldots, n,$$

which can be written in matrix form as

$$\begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+2} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{1}{n+1} & \cdots & \cdots & \frac{1}{2n+1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \int_0^1 f(x) \mathrm{d}x \\ \vdots \\ \vdots \\ \int_0^1 x^n f(x) \mathrm{d}x \end{pmatrix}.$$

The matrix on the left-hand side is called the *Hilbert matrix*, and is invertible since its determinant is equal to

$$\det = \frac{\left(\prod_{k=1}^{n} k!\right)^4}{\prod_{k=1}^{2n+1} k!} \neq 0.$$

It can be shown that the solution to this problem is very sensitive to any small perturbation of the vector on the right-hand side of the equality, making the problem at hand very ill-conditioned. In fact, even though the Hilbert matrix is invertible, numerical instabilities do arise when performing the inversion. Note further that this basis of monomials $(1, x, \dots, x^n)$ is not orthogonal in the space $\mathcal{H}$, whatever the weight function w is. The problem here is that the functions $x \mapsto x^k$ becomes more and more similar as k increases.

In order to bypass this issue (Remark 4.4.12), we would like to use an orthogonal basis, which would then ensure that its elements do not become more and more alike. The following Gram-Schmidt orthogonalisation algorithm provides us with such a construction.

Theorem 4.4.13 (Gram-Schmidt orthogonalisation)**.** *For any given weight function w, there exists a unique sequence of orthogonal polynomials $(\psi_n)_{n\geq 0}$ in $\mathcal{H}$ with* $\mathrm{degree}(\psi_n) = n$ *such that $\|\psi_n\|_{2,w} = 1$ and the highest-order coefficient of ψ_n is strictly positive.*

Proof. The proof is constructive and determines the sequence recursively. Start with the initial (constant) polynomial $\psi_0 \equiv c$. The condition $\|\psi_0\|_{2,w} = 1$ implies $c = \left(\int_a^b w(x)\mathrm{d}x\right)^{-1/2}$. Then, construct ψ_1 from an auxiliary polynomial p_1 defined by

$$p_1(x) = x + \alpha_{1,0}\psi_0(x).$$

The orthogonality condition $\langle p_1, \psi_0\rangle_w = 0$ implies

$$\alpha_{1,0} = -\langle x, \psi_0\rangle_w = -\frac{\int_a^b x w(x)\mathrm{d}x}{\left(\int_a^b w(x)\mathrm{d}x\right)^{1/2}}.$$

Define now $\psi_1 := p_1/\|p_1\|_{2,w}$, and note that ψ_1 satisfies the conditions of the theorem. The general term of the sequence is then determined by

$$\psi_n := \frac{p_n}{\|p_n\|_{2,w}},$$

where $p_n(x) := x^n + \alpha_{n,n-1}\psi_{n-1}(x) + \ldots + \alpha_{n,0}\psi_0(x)$, and where the constants $\alpha_{n,n-1}, \dots, \alpha_{n,0}$ are chosen to ensure orthogonality, i.e. $\alpha_{n,k} = -\langle x^n, \psi_k\rangle_w$ for each $k = 0, \dots, n-1$. □

The most common families of orthogonal polynomials are given as follows:

Name	(a, b)	$w(x)$
Legendre	$(-1, 1)$	1
Chebychev	$(-1, 1)$	$\left(1 - x^2\right)^{-1/2}$
Laguerre	$(0, \infty)$	$\exp(-x)$
Hermite	$\mathbb{R}$	$\exp(-x^2)$

Remark 4.4.14. Note that each orthogonal basis is defined on a pre-specified interval (a, b). A simple mapping of the nodes allows one to use them on any desired interval.

The following theorem gives a recursion algorithm for the sequence of polynomials constructed from the Gram-Schmidt orthogonalisation.

Theorem 4.4.15. *Let $(\psi_n)_{n\geq 0}$ be a family of orthogonal polynomials with respect to the weight function w on some interval (a, b). We assume that the n-th element of the sequence has the form $\psi_n(x) = A_n x^n + B_n x^{n-1} + \ldots$, and define*

$$a_n := \frac{A_{n+1}}{A_n} \quad \textit{and} \quad \gamma_n := \langle \psi_n, \psi_n \rangle_w \,.$$

Then, for all $n \geq 1$, we have

$$\psi_{n+1}(x) = \left(a_n x + b_n\right)\psi_n(x) - c_n \psi_{n-1}(x),$$

where

$$b_n := a_n \left(\frac{B_{n+1}}{A_{n+1}} - \frac{B_n}{A_n}\right) \quad \textit{and} \quad c_n := \frac{A_{n-1}A_{n+1}}{A_n^2}\frac{\gamma_n}{\gamma_{n-1}}.$$

The proof of the theorem follows by a careful yet simple manipulation of the Gram-Schmidt construction, and we hence leave it as an exercise.

Example.

(i) (Legendre polynomials). Let $(a, b) = (-1, 1)$ and $w(x) \equiv 1$, then

$$P_n(x) = \frac{(-1)^n}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[\left(1 - x^2\right)^n\right] \quad \text{and} \quad \|P_n\|_{2,w}^2 = \frac{2}{2n+1}.$$

and we have the recursion

$$P_{n+1}(x) = \frac{2n+1}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x).$$

(ii) (Chebychev polynomials). Let $(a, b) = (-1, 1)$ and $w(x) \equiv \left(1 - x^2\right)^{-1/2}$, then

$$T_n(x) = \cos\left(\operatorname{acos}(x) n\right),$$

and

$$\langle T_n, T_m \rangle_w = \begin{cases} 0, & \text{if } n \neq m, \\ \pi, & \text{if } n = m = 0, \\ \pi/2, & \text{if } n = m \neq 0. \end{cases}$$

and we have the recursion

$$T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x).$$

(iii) (Laguerre polynomials). Let $[a,b) = [0,\infty)$ and $w(x) \equiv \mathrm{e}^{-x}$, then

$$L_n(x) = \frac{\mathrm{e}^x}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^n\mathrm{e}^{-x}\right) \qquad \text{and} \qquad \|L_n\|_{2,w}^2 = 1,$$

and we have the recursion

$$L_{n+1}(x) = \frac{2n+1-x}{n+1}L_n(x) - \frac{n}{n+1}L_{n-1}(x).$$

(iii) (Hermite polynomials). Let $(a,b) = \mathbb{R}$ and $w(x) \equiv \mathrm{e}^{-x^2}$, then

$$H_n(x) = (-1)^n\mathrm{e}^{x^2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(\mathrm{e}^{-x^2}\right),$$

and we have the recursion

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x).$$

The following property is fundamental for the next section:

Theorem 4.4.16. *For any $n \geq 1$, the orthogonal monic ($A_n = 1$) polynomial p_n defined by Theorem 4.4.15 has exactly n distinct roots in the interval (a,b) which are the eigenvalues of the symmetric tridiagonal matrix*

$$\begin{pmatrix} -b_1 & c_1 & 0 & 0 \\ c_1 & -b_2 & \ddots & 0 \\ 0 & \ddots & \ddots & c_{n-1} \\ 0 & 0 & c_{n-1} & -b_n \end{pmatrix}.$$

Proof. Suppose that p_n has fewer than n roots in (a,b), and denote them $\widetilde{x}_1,\ldots,\widetilde{x}_m$ with $m<n$. Define the polynomial $q(x) := (x-\widetilde{x}_1)\ldots(x-\widetilde{x}_m)$. Since the polynomial product p_nq does not change sign in the whole interval (a,b), we have $\langle p_n, q\rangle \neq 0$. By the Gram-Schmidt construction, we further know that p_n is orthogonal to any polynomial of degree strictly smaller than n, so in particular is orthogonal to and q, and the contradiction concludes the proof of the theorem. The characterisation of the roots as eigenvalues of the matrix is left as an exercise. □

Interpolation via splines

We have seen so far how to construct interpolating polynomials given a set of data points; one of the main drawbacks of this approach is that as the number of interpolating nodes increases, the interpolating polynomial becomes more and more oscillating. An alternative method has become increasingly popular since the 1960s in many areas of applied mathematics (computer graphics, numerical solutions of integrals,...). The idea is first to split the domain of the function into a number of subdomains and then to apply an interpolation method on each of them. To be more precise, we consider a sequence of points $(\overline{x}_0,\ldots,\overline{x}_n)$ subdividing the domain of the function under consideration, say f, and we define a piecewise polynomial $P : \mathbb{R} \to \mathbb{R}$ of order $r \geq 1$ as a polynomial of degree at most r on each subinterval $[\overline{x}_i, \overline{x}_{i+1}]$, for $i = 0, \ldots, n-1$. There are several ways of constructing these polynomials:

(i) one could first consider a fixed number of nodes on each interval $[\overline{x}_i, \overline{x}_{i+1}]$ and then construct an interpolating polynomial as above using Lagrange polynomials;

(ii) one could choose a weight function w and approximate the function f using orthogonal polynomials where the nodes within the interval $[\overline{x}_i, \overline{x}_{i+1}]$ correspond to the (scaled) roots of the orthogonal polynomial with respect to w;

(iii) if the derivatives of the function f are known (or easily computable), one could use *cubic splines* as explained below.

Items (i) and (ii) are rather obvious to construct from the previous results, and are called *composite rules*. Likewise, their errors are straightforward to compute. The idea of cubic splines is to find a sequence of cubic polynomials $\mathrm{P} := (P_0, \ldots, P_{n-1})$ satisfying the following conditions:

(a) $P_i(x_i) = f(x_i)$ and $P_i(x_{i+1}) = f(x_{i+1})$, for $i = 0, \ldots, n-1$;

(b) $P_i'(x_{i+1}) = P_{i+1}'(x_{i+1}) = 0$, for $i = 0, \ldots, n-2$;

(c) $P_i''(x_{i+1}) = P_{i+1}''(x_{i+1}) = 0$, for $i = 0, \ldots, n-2$;

Condition (a) ensures the continuity of P at each node, and Conditions (b) and (c) ensure the continuity of the first and second derivatives of P at each node. Since the cubic polynomial on each interval has four coefficients, there are exactly $4n$ coefficients to determine in this procedure. However, the sets of Conditions (a), (b) and (c) only lead to $2n + (n-1) + (n-1) = 4n - 2$ equations. Extra conditions are therefore needed in order to make the problem well posed, and this leads to different types of cubic splines:

- Natural cubic splines:
$$P_0''(x_0) = P_{n-1}''(x_n) = 0.$$

- Clamped boundary conditions:
$$P_0'(x_0) = \alpha_0 \quad \text{and} \quad P_{n-1}'(x_n) = \alpha_n,$$
where α_0 and α_n are usually set equal to the first derivative of the function f at x_0 and x_n.

- Not-a-knot conditions:
$$P_0'''(x_1) = P_1'''(x_1) \quad \text{and} \quad P_{n-2}'''(x_{n-1}) = P_{n-1}'''(x_{n-1}).$$
This means that P''' is continuous at the nodes x_1 and x_{n-1}.

4.4.2 Numerical integration via quadrature

We are now interested in finding a way to compute (one-dimensional) integrals as accurately as possible. With Riemann integration in mind, we look for an approximation of the form

$$\int_A f(x)\mathrm{d}x \approx \sum_{i=0}^n w_i f(x_i),$$

where $A \subset \mathbb{R}$ is the integration domain, $n+1$ is the number of points we wish to use, $(x_i)_{0\leq i\leq n}$ some points in A and w_i some weights. Such a representation is called a *quadrature rule* and the points $x_0, \ldots, x_n$ are the *quadrature nodes*.

Newton-Cotes formulae

Let us start with a few examples:

- *Rectangular rule.* We approximate the integral as the area of a rectangle of width $(b-a)$ and height $f(a)$ or $f(b)$:
$$\int_a^b f(x)\mathrm{d}x \approx (b-a)f(a) \quad \text{or} \quad \int_a^b f(x)\mathrm{d}x \approx (b-a)f(b).$$
- *Mid-point rule*:
$$\int_a^b f(x)\mathrm{d}x \approx (b-a)\,f\left(\frac{a+b}{2}\right).$$
- *Trapezoidal rule.* We approximate the integral as the area of a trapezoid with base $[a,b]$ and sidelines $f(a)$ and $f(b)$:
$$\int_a^b f(x)\mathrm{d}x \approx \frac{b-a}{2}\Big(f(a)+f(b)\Big).$$

Note that the first two rules follow by approximating the function f by a constant $f(a)$—or $f(b)$—and $f\left(\frac{a+b}{2}\right)$. The trapezoidal rule follows by integrating the polynomial of order one: $p_1(x) := f(a) + \frac{x-a}{b-a}\left(f(b)-f(a)\right)$. Let us now generalise this. Let $x_0, \ldots, x_n$ be $n+1$ distinct points in the interval $[a,b]$ and consider the Lagrange interpolating polynomial P_n defined in Definition 4.4.3 by

$$P_n(x) := \sum_{i=0}^n y_i L_i^{(n)}(x).$$

Definition 4.4.17. The Newton-Cotes quadrature formula is defined as $\int_a^b P_n(x)\mathrm{d}x$, where P_n is an interpolating polynomial of the function f with nodes $x_0, \ldots, x_n$.

From this definition, we can then write

$$\int_a^b f(x)\mathrm{d}x \approx \int_a^b P_n(x)(x)\mathrm{d}x = \sum_{i=0}^n w_i f(x_i),$$

where $w_i := \int_a^b L_i^{(n)}(x)\mathrm{d}x$ for any $i = 0, \ldots, n$.

Remark 4.4.18. The three rules introduced above can be recovered by taking $n = 0$ and $x_0 = a$ (rectangular), $n = 0$ and $x_0 = (a + b)/2$ (mid-point) and $n = 1$, $x_0 = a$, $x_n = b$ (trapezoidal).

Remark 4.4.19. A more subtle (and more widely used) rule is $n = 2$, $x_0 = a$, $x_1 = (a + b)/2$ and $x_2 = b$ and is called *Simpson's rule*, i.e. we use a quadratic interpolating polynomial on the interval $[a, b]$. The integration then reads

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &\approx \int_a^b \left(\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2) \right) \mathrm{d}x \\ &= \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right),\end{aligned}$$

where $h := (b - a)/2$.

Remark 4.4.20. Note that these integration rules can also be applied to solve numerically ordinary differential equations. Consider indeed an initial value problem for an ODE of the form

$$\dot{x}_t = f(t, x_t), \quad \text{for } t \in [a, b], \text{ with } x_a = x_0.$$

Integrating both sides of the equality between a and $t \in [a, b]$ gives

$$x_t = x_0 + \int_a^t f(s, x_s)\,\mathrm{d}s,$$

and we are hence left with an integration problem between a and t.

Newton-Cotes integration error

Let us now look at the error of this integration approximation. For simplicity we shall assume that the function f to integrate is smooth, i.e. has derivatives of all orders. From Theorem 4.4.6, we know that for a polynomial P of degree at most $n \geq 0$ interpolating the function f at the distinct nodes $x_0, \dots, x_n$, the error is worth

$$f(x) - P_n(x) = \varepsilon_n(x), \tag{4.4.2}$$

where

$$\varepsilon_n(x) := \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^n (x - x_i),$$

for some $\xi \in [a, b]$ (depending on x). Integrating (4.4.2), we obtain

$$\int_a^b f(x)\mathrm{d}x = \int_a^b P_n(x)\mathrm{d}x + \int_a^b \varepsilon_n(x)\mathrm{d}x = \sum_{i=0}^n w_i f(x_i) + E_n,$$

where $E_n := \int_a^b \varepsilon_n(x)\mathrm{d}x$ is called the *integration error*. It is then immediate to see that

$$|E_n| \leq \sup_{x\in[a,b]} \left| f^{(n+1)}(\xi) \right| \frac{1}{(n+1)!} \prod_{i=0}^n \int_a^b |x - x_i| \mathrm{d}x.$$

More precise results are available in the literature, but we omit them here for brevity. We however refer the interested reader to Isaacson & Keller [39] for more details. In the rules discussed above, we can compute the error explicitly as follows.

Proposition 4.4.21. *Let $h := (b-a)/2$. With the rectangular rule, we have*

$$E_0^R = 2h^2 f'(\xi), \quad \textit{for some } \xi \in [a,b].$$

With the trapezoidal rule, we have

$$E_1^T = \frac{2h^3}{3} f''(\xi), \quad \textit{for some } \xi \in [a,b].$$

With the mid-point rule, we have

$$E_1^M = \frac{h^3}{3} f''(\xi), \quad \textit{for some } \xi \in [a,b].$$

With the Simpson rule, we have

$$E_2^R = -\frac{h^5}{90} f'(\xi), \quad \textit{for some } \xi \in [a,b].$$

We leave the proof of this proposition as an exercise. This proof is based on the mean value theorem for integrals (Theorem 4.4.22 below) for the rectangular and the trapezoidal rules. In the other two cases, this theorem is not enough, and one needs to use a Taylor expansion (i) of order two around the mid-point of the interval in the mid-point rule and (ii) of order three around the mid-point of the interval in the Simpson rule. We recall the following theorem for completeness.

Theorem 4.4.22 (Mean value theorem)**.** *Let f and g be two continuous functions on the interval $[a,b]$. If $g(x) \geq 0$ for all $x \in [a,b]$, then there exists $\gamma \in [a,b]$ such that*

$$\int_a^b f(x)g(x)\mathrm{d}x = f(\gamma) \int_a^b g(x)\mathrm{d}x.$$

Remark 4.4.23. It is clear that the composite polynomial interpolation method presented above on page 124 carries over to integration method: we first split the integration domain into a number of subdomains, and then approximate each integrand by its interpolating polynomial. We shall not write out the details here since they are clearly straightforward.

We finish this part with the following definition and exercises concerning the order of accuracy of interpolation schemes.

Definition 4.4.24. A $(n+1)$-node quadrature formula

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{i=0}^n w_i f(x_i),$$

has a degree of accuracy $m \geq 0$ if it is exact for any polynomial p_k of degree smaller than m:

$$\int_a^b p_k(x)\mathrm{d}x = \sum_{i=0}^n w_i p_k(x_i),$$

and not exact for some polynomial of degree equal to $m+1$.

Example. For the trapezoidal rule, the order of accuracy is $m = 1$ whereas it is equal to $m = 3$ for the Simpson rule.

One can in particular show that the order of accuracy of any Newton-Cotes formula cannot exceed the total number of nodes. We may now wonder whether the Newton-Cotes approximation formula converges to the original integral as the degree of the polynomial approximation of the integrand tends to infinity. This shall also serve—as we will see—as an introduction to Gaussian quadrature methods in the next subsection. As an exercise, the reader should numerically check that the Newton-Cotes approximation to the integration of the Runge function does not converge when the number of nodes goes large:

$$\int_{-5}^{5} \frac{\mathrm{d}x}{1+x^2}.$$

The following theorem makes this precise:

Theorem 4.4.25. *Let $n \geq 1$ and*

$$I_n(f) := \sum_{i=0}^{n} w_{i,n} f(x_{i,n})$$

be a sequence of numerical integration formulae approximating the integral $I(f) := \int_a^b f(x)\mathrm{d}x$. We write here $w_{i,n}$ and $x_{i,n}$ to highlight the dependence on the number of nodes $n+1$. Let then $\mathcal{F}$ be a dense family in the space $C([a,b])$ of continuous functions on the interval $[a,b]$. Then $I_n(f)$ converges to $I(f)$ for all $f \in C([a,b])$ if and only if the following two conditions are satisfied:

(i) *$I_n(f)$ converges to $I(f)$ for all $f \in \mathcal{F}$;*

(ii) $\sup_{n\geq 1} \sum_{i=0}^{n} |w_{i,n}| < \infty$.

Exercise 47. Using the theorem, check why the Newton-Cotes formulae do not converge for the Runge function.

Gaussian quadratures

As we mentioned above, the order of accuracy of a Newton-Cotes formula can never exceed the total number of interpolation / integration nodes. Gaussian integration finds the optimal choice of such nodes in order to maximise the order of accuracy of the scheme. Let us start with the following example: we wish to approximate the integral $\int_{-1}^{1} f(x)\mathrm{d}x$ by $w_1 f(x_1) + w_2 f(x_2)$, where the two constants w_1 and w_2 and the two nodes x_1 and x_2 are such that the order of accuracy is maximised. These four unknowns lead to a system of four equations via a polynomial of degree three (the monomials 1, x, x^2 and x^3):

$$\begin{aligned}
\int_{-1}^{1} 1\mathrm{d}x &= 2 = w_1 + w_2,\\
\int_{-1}^{1} x\mathrm{d}x &= 0 = w_1 x_1 + w_2 x_2,\\
\int_{-1}^{1} x^2\mathrm{d}x &= 2/3 = w_1 x_1^2 + w_2 x_2^2,\\
\int_{-1}^{1} x^3\mathrm{d}x &= 0 = w_1 x_1^3 + w_2 x_2^3,
\end{aligned}$$

which gives the unique solution $(w_1, w_2, x_1, x_2) = \left(1, 1, -\sqrt{3}/3, \sqrt{3}/3\right)$. Note that this integration rule gives an order of accuracy of 3 with only two nodes. As a comparison, with two nodes the trapezoidal rule has an order of accuracy of one and the Simpson's rule requires three nodes to attain an order of accuracy of three. This example serves as an introduction to the following definition of Gaussian quadrature:

Definition 4.4.26. A Gaussian quadrature consists in choosing n integration / interpolation nodes that maximise the order of accuracy to a value of $2n - 1$.

In the general case, let us consider n nodes labelled $x_1, \ldots, x_n$ and n corresponding weights $w_1, \ldots, w_2$. We wish again to obtain a numerical scheme for the integral $\int_{-1}^{1} f(x)\mathrm{d}x$. We consider all the monomials of degree smaller than $2n - 1$ (since there are $2n$ parameters to determine), which leads to the following system of equations:

$$\sum_{i=1}^{n} w_i x_i^k = \begin{cases} 0 & \text{if } k \text{ is odd},\\ \dfrac{2}{k+1} & \text{if } k \text{ is even}.\end{cases}$$

Note however than the non-linearity of the problem makes it difficult to solve for general $n \in \mathbb{N}$. We however have the following theorem:

Theorem 4.4.27. *Let $(\phi_n)_{n\geq 0}$ be a family of orthogonal polynomials on the interval $[a, b]$ with respect to a given weight function w, of the form*

$$\phi_n(x) = \sum_{k=0}^{n} A_k x^k.$$

Denote by $x_1, \ldots, x_n \in [a, b]$ the zeros of ϕ_n, let $a_n := A_{n+1}/A_n$ and $\gamma_n := \|\phi_n\|_{2,w}^2 = \int_a^b w(x)\phi_n^2(x)\mathrm{d}x$. For each $n \geq 1$, there is a unique integration approximation formula

$$I(f) := \int_a^b f(x)\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i) =: I_n(f)$$

with order of accuracy equal to $2n - 1$. If the function f is $2n$ times differentiable on $[a, b]$, then there exists $\xi \in [a, b]$ such that the integration error E_n reads

$$E_n := I(f) - I_n(f) = \frac{\gamma_n}{(2n)!A_n^2} f^{(2n)}(\xi).$$

Furthermore, the optimal nodes are given by the roots of ϕ_n and the weights by

$$w_i = \frac{-a_n\gamma_n}{\phi_n'(x_i)\phi_{n+1}(x_i)}, \quad \textit{for all } i = 1, \ldots, n.$$

Example. According to the table of orthogonal polynomials on page 121, the Legendre polynomials have weight function constant equal to one. In particular, the first of the series read

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{3x^2}{2} - \frac{1}{2}, \quad P_3(x) = \frac{5x^3}{2} - \frac{3x}{2}, \quad P_4(x) = \frac{35x^4}{8} - \frac{30x^2}{8} + \frac{3}{8}.$$

The Gauss-Legendre quadrature of $\int_{-1}^1 f(x)\mathrm{d}x$ therefore reads $\sum_{i=1}^n w_i f(x_i)$, where $x_1, \ldots, x_n$ are the n real roots of the Legendre polynomial of degree n, and the weights are given by

$$w_i = \frac{-2}{(n+1)P_n'(x_i)P_{n+1}(x_i)}, \quad \text{for } i = 1, \ldots, n,$$

and the integration approximation error reads

$$E_n = \frac{2^{2n+1}(n!)^4}{(2n+1)((2n)!)^2}\frac{f^{(2n)}(\xi)}{(2n)!},$$

for some $\xi \in (-1, 1)$. Note that a simple rescaling allows us to consider the integral over some interval $[a, b]$ as follows:

$$\int_a^b f(x)\mathrm{d}x = \frac{b-a}{2}\int_{-1}^1 f\left(\frac{a+b+(b-a)x}{2}\right)\mathrm{d}x.$$

Remark 4.4.28. Using the integration error formula, we can actually prove—using Stirling's approximation—that the convergence of the Gauss-Legendre quadrature is exponentially fast. This is to be compared to the trapezoidal and the Simpson rules which converge at a speed of n^{-2} and n^{-4} respectively.

Exercise 48. Compute the Gauss-Legendre quadrature to compute the integral $\int_0^\pi \mathrm{e}^x \cos(x)\mathrm{d}x$ and compare with a standard trapezoidal or Simpson rule. The true value of the integral is equal to $-(1 + \mathrm{e}^\pi)/2$.

Many other orthogonal polynomials exist in the literature, and we shall not present them here. One last alternative has been proposed and used in the financial literature, namely the *Gauss-Lobatto* quadrature. Contrary to what it may seem, it is not based on some possible Lobatto polynomials, but on the Legendre polynomials. The idea is that, when performing a Gauss-Legendre quadrature for the integral $\int_a^b f(x)\mathrm{d}x$, the two endpoints a and b are actually not taken into account by Theorem 4.4.16. The Gauss-Lobatto deforms the Gauss-Legendre quadrature in order to incorporate these two endpoints. It takes as weight function $w \equiv 1$, and the general Gauss-Lobatto quadrature formula reads (for clarity we normalised the integration domain to $[-1, 1]$)

$$\int_{-1}^1 f(x)\mathrm{d}x = \frac{2}{n(n-1)}(f(-1) + f(1)) + \sum_{i=2}^{n-1} w_i f(x_i),$$

where the abscissas $x_2, \dots, x_{n-1}$ are the roots of the polynomial P'_{n-1}, where P_{n-1} is a Legendre polynomial. The weights are computed explicitly as

$$w_i = \frac{2}{n\,(n-1)\,P^2_{n-1}(x_i)}, \quad \text{for } i = 2, \dots, n-1.$$

We refer the interested reader to [27] for more details and a precise implementation in Matlab.

Remark 4.4.29. Tables of roots of orthogonal polynomials (and hence of integration nodes) are available in [1].

Adaptive quadrature

A last refinement of integration by quadrature is adaptive quadrature. This is a straightforward yet powerful extension, and we shall hence present it very briefly. Suppose one wishes to evaluate the integral $I(f) := \int_a^b f(x)\mathrm{d}x$. A Gaussian quadrature (or Newton-Cotes) gives an estimate $I_n(f)$. Let us now specify a tolerance $\varepsilon > 0$ and split the interval (a, b) in two. Perform then a quadrature on each subinterval, which gives two values $I_{n,1}(f)$ and $I_{n,2}(f)$. If $|I_n(f) - (I_{n,1}(f) + I_{n,2}(f))| > \varepsilon$, then we keep splitting the subintervals. This methodology allows for a more refined grid where the function is less smooth or oscillates more rapidly. We shall see below some practical examples.

Numerical integration example

Quadratures are straightforward in MATLAB. Suppose we are interested in computing the integral $\int_0^{\pi/2} f(x)\mathrm{d}x$, where $f : x \in [0, \pi/2] \mapsto \mathrm{e}^{2x}\cos(x)$. Three commands are available to compute such an integral: **trapz** uses the trapezoidal rule, **quad** implements an adaptive Simpson rule while **quadl** evaluates the integral according to an adaptive Gauss-Lobatto quadrature, as proposed in [27]. As an example, consider the following MATLAB code:

Adaptive Simpson integration

$f = @(x)\exp(2.*x).*\cos(x)$ *(Note that we use the vector formulation)*

$[I, n] =$ quad(f,0,π/2,10^{-6}, 'trace on')

The output I gives the value of the integral, n is the number of function evaluations needed to obtain this value with a tolerance of 10^{-6}, and the argument *trace on* outputs the following table:

Function evaluations	x_i	$x_{i+1} - x_i$	I
9	0.0000000000	0.426596866	0.6489482541
11	0.0000000000	0.213298433	0.2637880598
13	0.2132984332	0.213298433	0.3851604435
15	0.4265968664	0.717602594	2.3779211477
17	0.4265968664	0.358801297	0.9919640373
19	0.4265968664	0.179400648	0.4385709242
21	0.6059975149	0.179400648	0.5533932720
23	0.7853981634	0.358801297	1.3859845427
25	0.7853981634	0.179400648	0.6602023206
27	0.9647988119	0.179400648	0.7257824978
29	1.1441994604	0.426596866	1.2012396448
31	1.1441994604	0.213298433	0.8045126483
33	1.1441994604	0.106649217	0.4251261784
35	1.2508486770	0.106649217	0.3793864797
37	1.3574978936	0.213298433	0.3967283285
39	1.3574978936	0.106649217	0.2825300514
41	1.4641471102	0.106649217	0.1141982878

The x_i represent the abscissa nodes at which the integrand is computed and the last column, I is the value of the integral on each subinterval (x_{i-1}, x_i).

4.4.3 Fast Fourier transform methods

The FFT algorithm

The Fast Fourier Transform is an algorithm to compute efficiently a vector $(\mathbf{F}_1, \ldots, \mathbf{F}_n)$ (for some $n \in \mathbb{N}$) given as the discrete Fourier transform of some vector $(f_1, \ldots, f_n)$:

$$\mathbf{F}_k := \sum_{j=1}^{n} \exp\left(-\frac{2\mathrm{i}\pi}{n}(j-1)(k-1)\right) f_j, \quad \text{for } k = 1, \ldots, n, \tag{4.4.3}$$

More precisely, consider the vectors $\mathbf{f} = (f_1, \ldots, f_n) \in \mathbb{R}^n$, and $\mathbf{F} = (F_1, \ldots, F_n) \in \mathbb{C}^n$. If $\mathbf{F}$ denotes the discrete Fourier transform of the vector $\mathbf{f}$, then we can write $\mathbf{F} = \mathbf{W}_n \mathbf{f}$, where the complex matrix $\mathbf{W}_n \in \mathcal{M}_n(\mathbb{C})$ is given by

$$\mathbf{W}_n = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & \omega_n & \omega_n^2 & \ddots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \ddots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \ldots & \ldots & \omega_n^{(n-1)(n-1)} \end{pmatrix},$$

with $\omega_n := \exp\left(-\frac{2\mathrm{i}\pi}{n}\right)$. The computation of the vector $\mathbf{F}$ requires an order $\mathcal{O}(n^2)$ operations. The FFT algorithm, in particular the one proposed by Colley and Tukey [12], reduces the number of computations to $\mathcal{O}(n\log_2(n))$. Assume that there exists $L \in \mathbb{N}$ such that $n = 2^L$, and define the sequences $(x_j)_{1\leq j\leq n/2}$ and $(y_j)_{1\leq j\leq n/2}$ by

$$x_j := f_{2j-1} \qquad \text{and} \qquad x_j := f_{2j}, \qquad \text{for } j = 1, \dots, n/2.$$

We can therefore rewrite, for any $k = 1, \dots, n$:

$$\begin{aligned} F_k = (\mathbf{W}_n\mathbf{f})_k &= \sum_{j=1}^{n} f_j \omega_n^{(j-1)(k-1)} \\ &= \sum_{j=1}^{n/2} \left[f_{2j-1}\omega_n^{(2j-2)(k-1)} + f_{2j}\omega_n^{(2j-1)(k-1)}\right] \\ &= \sum_{j=1}^{n/2} \left[x_j\omega_{n/2}^{(j-1)(k-1)} + y_j\omega_{n/2}^{(j-1)(k-1)}\omega_n^{k-1}\right] \\ &=: X_k + Y_k\omega_n^{k-1}. \end{aligned}$$

where we used the identity $\omega_n^{2(j-1)(k-1)} = \omega_{n/2}^{(j-1)(k-1)}$. We therefore obtain the so-called ‘butterfly relations’, for $k = 1, \dots, n/2$:

$$\left\{\begin{aligned} F_k &= X_k + Y_k\omega_n^{k-1}, \\ F_{k+n/2} &= X_{k+n/2} + Y_{k+n/2}\omega_n^{k-1+n/2}. \end{aligned}\right. \tag{\textit{butterfly relations}}$$

Note further that $X_{k+n/2} = X_k$, that $Y_{k+n/2} = Y_k$ and that $\omega_n^{n/2} = -1$, so that the butterfly relations read

$$\left\{\begin{aligned} F_k &= X_k + Y_k\omega_n^{k-1}, \\ F_{k+n/2} &= X_k - Y_k\omega_n^{k-1}. \end{aligned}\right.$$

This splitting therefore reduces the computation to two discrete Fourier Transforms of size $n/2$. We can iterate this procedure L times until we obtain sequences of length one. After $L = \log_2(n)$ steps we have n butterfly relations to evaluate; which require one multiplication and two additions. Therefore the total computational cost therefore is of order $\mathcal{O}(n\log_2(n))$.

Remark 4.4.30. It is possible to extend the algorithm to the case where n is not a power of 2, but this is outside the scope of these notes.

Application to option pricing

In view of the pricing formulae (4.2.3) or (4.2.1), we can use the above quadrature methods to evaluate the integrals, and hence the call option price. We can also use Fast Fourier transform (FFT) methods, which are often more tailored for numerical integration problems involving Fourier transforms. As mentioned above, the FFT is an efficient way to compute sums of the form

$$\Phi(k) := \sum_{j=1}^{n} \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)(k-1)} f(\xi_j), \quad \text{for } k = 1, \dots, n, \tag{4.4.4}$$

where n is an integer, usually of the form 2^p ($p \in \mathbb{N}$) and $(\xi_j)_{1\le j\le n}$ are nodes. Standard algorithms (the *discrete Fourier transform*) requires a total of n^2 multiplications to compute all the terms $\Phi(1), \dots, \Phi(n)$. The FFT algorithm actually reduces this quantity to $\mathcal{O}\left(n\log(n)\right)$. We shall not detail here the FFT procedure and refer the interested reader to [52] for different FFT algorithms used in practice.

Let us consider the FFT algorithm applied to the Carr-Madan formula (4.2.3). Using the trapezoidal rule for the integral in (4.2.3), and setting $\xi_j := \eta\,(j-1)$ for some $\eta > 0$, we obtain

$$C_T(k) \approx \frac{\mathrm{e}^{-\alpha k}}{\pi}\eta\left(2\mathrm{e}^{-\mathtt{i}k\xi_1}\psi(\xi_1) + 2\mathrm{e}^{-\mathtt{i}k\xi_{n+1}}\psi(\xi_{n+1}) + \sum_{j=2}^{n}\mathrm{e}^{-\mathtt{i}k\xi_j}\psi(\xi_j)\right).$$

Note that the integration domain $[0,\infty)$ has been truncated to $[0, n\eta]$. Since the FFT algorithm returns a vector of n values $(k_1, \dots, k_n)$, we have to select these abscissas first. We define a regular grid on the k-axis with step $\lambda > 0$ so that

$$k_u := -b + \lambda\,(u-1)\,, \quad \text{for } u = 1, \dots, n.$$

In order to have a symmetric grid (around the origin), we choose $b = \lambda n/2$. The final term $\mathrm{e}^{-\mathtt{i}k\xi_{n+1}}\psi(\xi_{n+1})$ is exponentially smaller than the others, so we may discard it, which yields the approximation

$$\begin{aligned} C_T(k_u) &\approx \frac{\mathrm{e}^{-\alpha k_u}}{\pi}\sum_{j=1}^{n}\exp\Big\{-\mathtt{i}\xi_j\left(-b+\lambda(u-1)\right)\Big\}\psi(\xi_j)\eta_j \\ &\approx \frac{\mathrm{e}^{-\alpha k_u}}{\pi}\sum_{j=1}^{n}\exp\Big\{-\mathtt{i}\lambda\eta(j-1)(u-1)\Big\}\mathrm{e}^{\mathtt{i}b\xi_j}\psi(\xi_j)\eta_j, \end{aligned}$$

where $\eta_j := \eta(1+\delta_{j-1})$. In order to apply the FFT algorithm, in view of (4.4.4), we need to set $\lambda\eta = 2\pi/n$. Note that this condition imposes some constraint on the methodology. A small value for η creates a fine grid for the discretisation of the inverse Fourier transform integral. However this also implies that the grid in the k-space becomes less dense, and hence one may not be able to obtain a precise accuracy for some strikes not on this grid. In order to take this into account, one can add a Simpson's rule to finally lead to the approximation

$$C_T(k_u) \approx \frac{\mathrm{e}^{-\alpha k_u}}{\pi}\sum_{j=1}^{n}\exp\left(-\frac{2\mathtt{i}\pi}{n}\,(j-1)\,(u-1)\right)\mathrm{e}^{\mathtt{i}b\xi_j}\psi\left(\xi_j\right)\frac{\eta}{3}\left(3+(-1)^j-\delta_{j-1}\right).$$

Example. See the implementation in the IPython notebook.

4.4.4 Fractional FFT methods

In the Fast Fourier transform approach (Section 4.4.3), the constraint $\lambda\eta = 2\pi/n$ is imposed on the discretisation parameters in order for the method to work. For a fixed computational cost (the dimension n), the FFT imposes a tradeoff between the accuracy of the integration (the truncation of the integration domain is $[0, n\eta])$ and the spacing of the strikes, $\lambda = 2\pi/(n\eta)$. For a very accurate integration, strikes might not be spaced densely enough to match observed strikes, and the required interpolation between two strikes shall create additional error. The fractional FFT, introduced in 1991 by Bailey and Swarztrauber [4], is a step further in order to bypass this issue, and computes sums of the form

$$\Phi(k) := \sum_{j=1}^{n} \mathrm{e}^{-2\mathtt{i}\pi\gamma(j-1)(k-1)} f(\xi_j), \quad \text{for } k = 1, \ldots, n, \tag{4.4.5}$$

where n is an integer and $(\xi_j)_{1\leq j\leq n}$ are the nodes. In the standard FFT method (4.4.4), $\gamma = 1/n$. Using the identity $2(j-1)(k-1) = (j-1)^2 + (k-1)^2 - (k-j)^2$, we can write, for any $k = 1, \ldots, n$,

$$\begin{aligned}\Phi(k) &= \sum_{j=1}^{n} \mathrm{e}^{-\mathtt{i}\gamma\pi\left[(j-1)^2+(k-1)^2-(k-j)^2\right]} f(\xi_j) \\ &= \mathrm{e}^{-\pi\gamma(k-1)^2} \sum_{j=1}^{n} \mathrm{e}^{-\mathtt{i}\pi\gamma\left[(j-1)^2-(k-j)^2\right]} f(\xi_j) \\ &= \mathrm{e}^{-\pi\gamma(k-1)^2} \sum_{j=1}^{n} y_j z_{k-j},\end{aligned} \tag{4.4.6}$$

where $y_j := \mathrm{e}^{-\mathtt{i}\pi\gamma(j-1)^2} f(\xi_j)$ and $z_j := \mathrm{e}^{-\mathtt{i}\pi\gamma j^2}$. Recall now that, for a given vector $\mathrm{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, the discrete Fourier transform is a vector $\hat{\mathrm{x}} = (\hat{x}_k)_{k=1,\ldots,n}$ satisfying

$$\mathcal{F}(\mathrm{x})_k := \hat{x}_k := \sum_{j=1}^{n} \exp\left(-\frac{2\mathtt{i}\pi}{n}(j-1)(k-1)\right) x_j,$$

and the inverse Fourier transform reads, for any $j = 1, \ldots, n$,

$$\mathcal{F}^{-1}(\hat{\mathrm{x}})_j := x_j = \sum_{k=1}^{n} \exp\left(\frac{2\mathtt{i}\pi}{n}(k-1)(j-1)\right) \hat{x}_k.$$

For two vectors x and w in $\mathbb{R}^n$, the convolution is defined as the following operation:

$$(\mathrm{x} * \mathrm{w})_l := \sum_{j=1}^{n} x_j w_{j-l} = \sum_{j=1}^{n} w_j x_{j-l} = (\mathrm{w} * \mathrm{x})_l, \qquad \text{for any } l = 1, \ldots, n.$$

Recall now the convolution duality, the proof of which follows by simple manipulations:

Theorem 4.4.31. *For any two vectors* x *and* w *in* $\mathbb{R}^n$*, the identity* $\mathcal{F}(\mathrm{x} * \mathrm{w}) = (\hat{\mathrm{x}} \odot \hat{\mathrm{w}})$ *holds.*

Remark 4.4.32. The symbol $\odot$ denotes the component-by-component multiplication, so that the expression in the theorem reads, component-wise: $\mathcal{F}(\mathrm{x} * \mathrm{w})_l = \hat{\mathrm{x}}_l \hat{\mathrm{w}}_l$, for each $k = 1, \ldots, n$.

Therefore, the expression (4.4.6) can be rewritten, for any $k = 1, \ldots, n$, as

$$\Phi(k) = \mathrm{e}^{-\pi\gamma(k-1)^2} \left(\mathrm{y} * \mathrm{z}\right)_k = \mathrm{e}^{-\pi\gamma(k-1)^2} \mathcal{F}^{-1}\Big(\hat{\mathrm{y}} \odot \hat{\mathrm{z}}\Big)_k = \mathrm{e}^{-\pi\gamma(k-1)^2} \mathcal{F}^{-1}\Big(\mathcal{F}(\mathrm{y})_k \mathcal{F}(\mathrm{z})_k\Big).$$

Standard Fourier transform procedures (such as the FFT above) requires some circular property of the input vectors, which is not quite the case here since $z_{k-j} = z_{j-k}$. The idea of the fractional

FFT is to fictitiously extend the vectors y and z in $\mathbb{R}^n$ to $\overline{\mathrm{y}}$ and $\overline{\mathrm{z}}$ in $\mathbb{R}^{2n}$ in the following way:

$$\begin{aligned} \overline{y}_j &= y_j, && 1 \leq j \leq n,\\ \overline{y}_j &= 0, && n < j \leq 2n,\\ \overline{z}_j &= z_j, && 1 \leq j \leq n,\\ \overline{z}_j &= z_{2n-(j-1)}, && n < j \leq 2n, \end{aligned}$$

so that, for any $k = 1, \ldots, 2n$,

$$\Phi(k) = \mathrm{e}^{-\pi\gamma(k-1)^2}\mathcal{F}^{-1}\Big(\mathcal{F}(\overline{\mathrm{y}}) \odot \mathcal{F}(\overline{\mathrm{z}})\Big)_k.$$

Example. See the IPython notebook.

4.4.5 Sine / Cosine methods

Description of the method

One of the main drawbacks of the FFT method presented above is that it does not seem to be able to handle path-dependent options. We now present a method, still based on the knowledge of the characteristic function of the underlying process, due to Fang and Osterlee [25] (see also the second author's webpage for several related papers), which, not only improves the computational pricing time, but allows for path-dependent options.

For a function $f := [0, \pi] \to \mathbb{R}$, its Fourier cosine series expansion reads

$$f(\theta) = \frac{1}{2}\alpha_0 + \sum_{n\geq 1}\alpha_n \cos(n\theta) =: \overline{\sum_{n\geq 0}}\alpha_n \cos(n\theta),$$

where the Fourier coefficients are given by

$$\alpha_n = \frac{2}{\pi}\int_0^\pi f(\theta)\cos(n\theta)\mathrm{d}\theta.$$

By scaling, it is straightforward to extend this definition to any closed interval $[a, b]$; with the mapping $\theta = (x-a)\pi/(b-a)$, or $x = a + (b-a)\theta/\pi$, we can write

$$f(x) = \overline{\sum_{n\geq 0}}\alpha_n \cos\left(\frac{x-a}{b-a}n\pi\right), \tag{4.4.7}$$

with

$$\alpha_n = \frac{2}{b-a}\int_a^b f(x)\cos\left(\frac{x-a}{b-a}n\pi\right)\mathrm{d}x.$$

Consider a one-dimensional random variable with density f and characteristic function ϕ:

$$\phi(\xi) := \int_{\mathbb{R}} \mathrm{e}^{\mathtt{i}\xi x} f(x)\mathrm{d}x.$$

Evaluation the function ϕ at the point $n\pi/(b-a)$, for some $n \geq 0$, and truncating the integral to the compact interval $[a, b]$, we can write

$$\widetilde{\phi}\left(\frac{n\pi}{b-a}\right) := \int_a^b \exp\left(\frac{\mathtt{i}n\pi x}{b-a}\right) f(x)\mathrm{d}x,$$

and therefore

$$\widetilde{\phi}\left(\frac{n\pi}{b-a}\right)\exp\left(-\frac{\mathtt{i}n\pi a}{b-a}\right) = \int_a^b \exp\left\{\mathtt{i}n\pi\left(\frac{x-a}{b-a}\right)\right\} f(x)\mathrm{d}x$$
$$= \int_a^b \left\{\cos\left(n\pi\left(\frac{x-a}{b-a}\right)\right) + \mathtt{i}\sin\left(n\pi\left(\frac{x-a}{b-a}\right)\right)\right\} f(x)\mathrm{d}x.$$

Taking the real part on both sides and assuming that the truncation error is negligible, by identification with the Fourier coefficients above, we have

$$\alpha_n = \frac{2}{b-a}\Re\left\{\widetilde{\phi}\left(\frac{n\pi}{b-a}\right)\exp\left(-\frac{\mathtt{i}n\pi a}{b-a}\right)\right\} \approx \frac{2}{b-a}\Re\left\{\phi\left(\frac{n\pi}{b-a}\right)\exp\left(-\frac{\mathtt{i}n\pi a}{b-a}\right)\right\}. \tag{4.4.8}$$

Combining this and (4.4.7), we obtain the cosine series expansion for the density f:

$$f(x) \approx \overline{\sum}_{n\geq 0}\alpha_n \cos\left(\frac{x-a}{b-a}n\pi\right) \approx \overline{\sum}_{n=0}^{N}\alpha_n \cos\left(\frac{x-a}{b-a}n\pi\right). \tag{4.4.9}$$

Exercise 49. Study the numerical efficiency (as a function of N) of the approximation (4.4.9) for the Gaussian density $f(x) \equiv (2\pi)^{-1-2}\exp\left(-\frac{1}{2}x^2\right)$, with $[a,b] = [-10,10]$.

Application to option pricing

Let us now consider a path-dependent option with payoff $h(\cdot)$ at maturity T, and denote by $u(x,t)$ its value at time $t \in [0,T]$, where $x \in \mathbb{R}$ denotes the initial value of the log-stock price (or a rescaled version of it), which we assume has a transition density f between t and T, so that

$$v(x,t) = \int_{\mathbb{R}} h(z)f(z|x)\mathrm{d}z.$$

Assume that the density is supported on a compact interval $[a,b]$ (or that we can neglect the tail parts $\mathbb{R}\setminus[a,b]$), then (4.4.9) yields

$$v(x,t) = \int_a^b h(z)\overline{\sum}_{n\geq 0}\alpha_n\cos\left(\frac{z-a}{b-a}n\theta\right)\mathrm{d}z,$$

where the coefficients $(\alpha_n)_{n\geq 0}$ are approximated via (4.4.8). Truncating the infinite sum at some level N, we finally obtain the following approximation for the option price:

$$v(x,t) \approx \overline{\sum}_{n=0}^{N}\alpha_n\int_a^b h(z)\cos\left(\frac{z-a}{b-a}n\pi\right)\mathrm{d}z =: \overline{\sum}_{n=0}^{N}\alpha_n V_n.$$

In the case of a Call option price with strike K and maturity T, with $h(z) \equiv K(\mathrm{e}^z-1)_+$,

$$V_n^{\mathrm{Call}} = \int_a^b h(z)\cos\left(\frac{z-a}{b-a}n\pi\right)\mathrm{d}z = K\int_a^b(\mathrm{e}^z-1)_+\cos\left(\frac{z-a}{b-a}n\pi\right)\mathrm{d}z = \frac{2K}{b-a}\Big\{\chi_n(0,b)-\varphi_n(0,b)\Big\},$$

and for a Put option with $h(z) \equiv K(1-\mathrm{e}^z)_+$,

$$V_n^{\mathrm{Put}} = \int_a^b h(z)\cos\left(\frac{z-a}{b-a}n\pi\right)\mathrm{d}z = K\int_a^b(1-\mathrm{e}^z)_+\cos\left(\frac{z-a}{b-a}n\pi\right)\mathrm{d}z = \frac{2K}{b-a}\Big\{\varphi_n(a,0)-\chi_n(a,0)\Big\},$$

where

$$\begin{aligned}\chi_n(c,d) &:= \int_c^d \mathrm{e}^y \cos\left(\frac{y-a}{b-a}n\pi\right)\mathrm{d}y\\ &= \frac{\cos\left(\frac{d-a}{b-a}n\pi\right)\mathrm{e}^d - \cos\left(\frac{c-a}{b-a}n\pi\right)\mathrm{e}^c + \frac{n\pi}{b-a}\sin\left(\frac{d-a}{b-a}n\pi\right)\mathrm{e}^d - \frac{n\pi}{b-a}\sin\left(\frac{c-a}{b-a}n\pi\right)\mathrm{e}^c}{1+\left(\frac{n\pi}{b-a}\right)^2}\end{aligned}$$

and

$$\varphi_n(c,d) = \begin{cases} \left[\sin\left(\dfrac{d-a}{b-a}n\pi\right) - \sin\left(\dfrac{c-a}{b-a}n\pi\right)\right]\dfrac{b-a}{n\pi}, & \text{if } n \neq 0,\\ d-c, & \text{otherwise.}\end{cases}$$

Exercise 50. Consider a digital Call option with payoff $h(z) \equiv \mathbb{1}_{\{\mathrm{e}^z \geq K\}}$. Show that

$$V_n = \frac{2K}{b-a}\varphi_n(0,b).$$

Remark 4.4.33. It is not clear, a priori, how to choose the truncation domain $[a,b]$. In [25], the authors propose the following:

$$[a,b] = \left[\mathrm{c}_1 - L\sqrt{\mathrm{c}_2 + \sqrt{\mathrm{c}_4}}, \mathrm{c}_1 + L\sqrt{\mathrm{c}_2 + \sqrt{\mathrm{c}_4}}\right],$$

with $L = 10$, and where $\mathrm{c}_i := \partial_u^i \log \mathbb{E}\left(\mathrm{e}^{uX}\right)$ is the ith cumulant of X. Higher-order truncation intervals are also possible, involving higher cumulants. These are necessary when dealing, for example, with short-maturity options, where the accuracy of this method—and of the FFT method as well—is low.

Exercise 51. Implement the Cosine method to compute a European Call option in the Black-Scholes model, and analyse the error.

Chapter 5

Model calibration

In this chapter, we shall endeavour to introduce numerical methods with a view towards everyday practical issues. As a motivating example, let us introduce the concept of *implied volatility.* Recall that, in the Black-Scholes model, the stock price is assumed to have the following dynamics:

$$S_t = S_0 \exp\left(\left(r - \frac{\sigma^2}{2}\right) t + \sigma W_t\right),$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion, $r \geq 0$ is the risk-free interest rate, and $\sigma > 0$ is the instantaneous volatility. As already proved, the Black-Scholes price (at time zero) of a European call option $C^{\mathrm{BS}}(S_0, K, T, \sigma)$ written on the underlying S, with strike $K > 0$ and maturity $T > 0$ has the following closed-form expression:

$$C_{\mathrm{BS}}(S_0, K, T, \sigma) = S_0 \mathcal{N}(d_+) - K\mathrm{e}^{-rT}\mathcal{N}(d_-), \tag{5.0.1}$$

where

$$d_\pm := \frac{\log(S_0/K) + \left(r \pm \sigma^2/2\right) T}{\sigma\sqrt{T}}.$$

This leads to the following concept of implied volatility.

Definition 5.0.34. Given an underlying S, a strike $K > 0$, a maturity $T > 0$ and an observed European call option price $C^{\mathrm{obs}}(S_0, K, T)$, the implied volatility is the unique $\sigma > 0$ such that the following equality holds:

$$C^{\mathrm{BS}}(S_0, K, T, \sigma) = C^{\mathrm{obs}}(S_0, K, T).$$

The fact that the implied volatility is defined uniquely for each strike and maturity follows directly from the fact that the map $\sigma \mapsto C^{\mathrm{BS}}(S_0, K, T, \sigma)$ is strictly increasing on $\mathbb{R}_+ \setminus \{0\}$ (since the Vega—the derivative of the call price with respect to the volatility— is strictly positive, see Exercise 5 on page 29). The implied volatility has many useful properties and is one of the most important concepts in mathematical finance. In fact, most vanilla options are actually quoted in terms of the implied volatility rather than in terms of the option value. The implied volatility is therefore a function of both the strike and the maturity, and we shall hence use the notation $\sigma_T(K)$. The map $(K, T) \mapsto \sigma_T(K)$ is called the implied volatility surface, and for each $T > 0$, the function $K \mapsto \sigma_T(K)$ is an implied volatility slice. Let now $C^{\mathrm{M}}\left(S_0, K, T, \overrightarrow{\theta}\right)$ be the call option price with the same characteristics (underlying stock price, strike, maturity) in some model (other than Black-Scholes) depending on a vector of parameters $\overrightarrow{\theta}$. The corresponding implied volatility is then the volatility $\sigma > 0$ plugged in the Black-Scholes formula (5.0.1) so that the Black-Scholes price and the model price are equal. In this chapter, we shall be interested in the following questions:

(i) Given a set of observed option prices, how does one recover the implied volatility?

(ii) Given a set of observed option prices—or implied volatilities—how does on calibrate the vector of parameters $\overrightarrow{\theta}$?

The first question is clearly related to root-finding. The second question is more subtle and is an optimisation problem. Let us temporarily leave the option pricing framework and consider a portfolio $(\Pi_t)_{t \geq 0}$ constructed as

$$\Pi_t := \sum_{i=1}^{n} w_i \pi_t^i, \quad \text{for all } t \geq 0,$$

where π_t^i represents the value at time t of some contingent claim (European option, American option, barrier option, single stock price,...) and $w_i \in (0, 1)$ represents its weight in the whole portfolio. At some future time $T > 0$, the value Π_T of the portfolio is random and one is interested in determining the optimal weights $\mathrm{w} = (w_1, \ldots, w_n)$ in order to maximise its expectation. The optimisation problem can therefore be written as

$$\sup_{\mathrm{w}} \mathbb{E}\left(\Pi_T\right), \quad \text{subject to} \begin{cases} w_i \in (0,1) \text{ for } i = 1, \ldots, n, \\ \sum_{i=1}^{n} w_i = 1. \end{cases}$$

One could (and does) add further constraints, such as diversification, i.e. each weight is bounded by some constant in $(0, 1)$. The famous Markowitz efficient frontier problem is to maximise such an expectation while minimising the variance of the portfolio at maturity T.

5.1 Solving non-linear equations

In this section we shall consider a function $f : [a, b] \to \mathbb{R}$, where $[a, b]$ is some interval of the real line, and we are interested in solving the equation $f(x) = 0$ for $x \in [a, b]$. Let us first recall—without proof—the following elementary theorem (Intermediate value theorem) from calculus:

Theorem 5.1.1. *If the function f is continuous and $f(a)f(b) \leq 0$ then the equation $f(x) = 0$ admits at least one solution in $[a, b]$.*

All the methods presented below rely on the construction of a sequence converging to the solution of the equation. In order to study the speed of convergence of this sequence, the following definition will be fundamental:

Definition 5.1.2. A sequence $(x_n)_{n\geq 0}$ is said to converge to x^* with order $p \geq 1$ if there exists $\gamma > 0$ such that

$$|x_{n+1} - x^*| \leq \gamma |x_n - x^*|^p, \quad \text{for any } n \geq 0.$$

The convergence is said to be linear if $p = 1$ and quadratic if $p = 2$. In the case $p = 1$, we further require that $\gamma \in (0, 1)$.

Remark 5.1.3. Note that when $p = 1$, we can iterate the definition to obtain $|x_n - x^*| \leq \gamma^n |x_0 - x^*|$. In this case, we may alternatively say that the convergence is linear with rate γ.

5.1.1 Bisection method

The bisection method constructs a sequence of couples $(x_n, y_n)_{n\geq 0}$ in $[a, b]$ defined recursively by $(x_0, y_0) := (a, b)$ and

$$(x_{n+1}, y_{n+1}) := \begin{cases} \left(\dfrac{x_n + y_n}{2}, y_n\right), & \text{if } f\left(\dfrac{x_n + y_n}{2}\right) f(y_n) < 0, \\ \left(x_n, \dfrac{x_n + y_n}{2}\right), & \text{if } f\left(\dfrac{x_n + y_n}{2}\right) f(y_n) > 0, \\ (x_n, y_n), & \text{if } f(x_n)f(y_n) = 0, \end{cases}$$

for all $n \geq 0$. The algorithm clearly stops in the third case, where an exact solution is found (either x_n or y_n). As usual, we want to make sure that the algorithm does converge to some limiting value as n tends to infinity. It is straightforward to see that the root x^* satisfies $\left|\frac{x_n+y_n}{2} - x^*\right| \leq 2^{-n}(b - a)$, so that the algorithm converges linearly with a rate equal to $1/2$. The major advantages of the bisection method are (i) that it converges provided that the function f is continuous and that $f(a)f(b) \leq 0$ and (ii) that we have an estimate of the error. However, the algorithm does not take into account the precision of the computer, for instance if the function does not vary much in the vicinity of x^*, the evaluation of $f(x_n)$ might lead to a wrong sign when x_n is close to x^*. Furthermore, the algorithm does not work for roots of even multiplicity, for instance when $f(x) \equiv x^2$. Finally, the convergence is rather slow—as opposed to the methods below—and it may not always be easy to find quantities a and b satisfying $f(a)f(b) \leq 0$.

5.1.2 Newton-Raphson method

As in the bisection method above, the idea here is to construct a sequence $(x_n)_{n\geq 0}$ converging to the root of the equation $f(x) = 0$ in the interval $[a, b]$. Start with some initial guess $x_0 \in [a, b]$. A Taylor series expansion of the function f around this point gives

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \mathcal{O}\left((x - x_0)^2\right).$$

The idea of Newton-Raphson algorithm is to locally replace the problem $f(x) = 0$ by the linear problem $f(x_0) + (x - x_0)\, f'(x_0) = 0$, which gives the solution $x_1 := x_0 - f(x_0)/f'(x_0)$. We now iterate this procedure as

$$x_{n+1} := x_n - f(x_n)/f'(x_n), \qquad \text{for } n \geq 0.$$

The quality of this algorithm is stated in the following theorem:

Theorem 5.1.4. *Assume that the functions f, f' and f'' are continuous in some neighbourhood of x^* and that $f'(x^*) \neq 0$. If x_0 is sufficiently close to x^*, the algorithm converges to x^* and*

$$\lim_{n\to\infty} \frac{x^* - x_{n+1}}{(x^* - x_n)^2} = -\frac{f''(x^*)}{2f'(x^*)},$$

and hence the algorithm has an order of convergence equal to two.

Proof. By continuity of the map f', we can choose a small neighbourhood $I := [x^* - \varepsilon, x^* + \varepsilon]$ around x^* for some $\varepsilon > 0$, so that the quantity

$$M := \frac{\max_{x\in I} |f''(x)|}{2 \min_{x\in I} |f'(x)|}$$

is well defined. Now, a Taylor expansion up to second order of f around x_n (for some $n \geq 1$) gives

$$f(x) = f(x_n) + (x - x_n)\, f'(x_n) + \frac{f''(\xi)}{2} (x - x_n)^2,$$

for some ξ between x and x_n. Therefore with $x = x^*$, we obtain

$$x^* = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{(x^* - x_n)^2}{2} \frac{f''(\xi)}{f'(x_n)},$$

which can be written as

$$x^* - x_{n+1} = -\frac{(x^* - x_n)^2}{2} \frac{f''(\xi)}{f'(x_n)}, \quad \text{for } n \geq 0. \tag{5.1.1}$$

Therefore $|x^* - x_1| = M\, |x^* - x_0|^2$ (and clearly $M\, |x^* - x_1| = (M\, |x^* - x_0|)^2$). Take now x_0 such that $|x^* - x_0| \leq \varepsilon$ and $M\, |x^* - x_0| < 1$. This implies that $M\, |x^* - x_1| < 1$ and $M\, |x^* - x_1| < M\, |x^* - x_0|$ so that $|x^* - x_1| \leq \varepsilon$. The same inequalities then hold for $|x^* - x_n|$ for any $n \geq 0$. Equation (5.1.1) implies by induction that

$$|x^* - x_n| \leq M^{-1} (M\, |x^* - x_0|)^{2n},$$

and therefore, since $M\, |x^* - x_0| < 1$, the sequence $(x_n)_{n\geq 0}$ converges to x^* as n tends to infinity. Now, since the variable ξ in (5.1.1) lies between x_n and x^* we obtain the limit stated in the theorem by continuity of the functions f' and f''. □

Remark 5.1.5. The proof of the theorem sheds a light on how close the initial value x_0 should be from the root x^*. The Newton-Raphson algorithm converges much more quickly than the bisection method, and it does not require the initial value to be in some predefined interval. However, the convergence is local, and the algorithm requires the computation of the first derivative, which may be computationally intensive. Note finally that if the function f is rather flat in the vicinity of x^* then the algorithm may be computationally awkward when dividing by $f'(x_n)$ at some step $n \geq 0$.

5.1.3 The secant method

The secant method builds upon the Newton-Raphson algorithm, and uses an approximation of the first derivative of the function f. Consider two initial estimates x_0 and x_1 of the root of the equation $f(x) = 0$. Approximate the graph of f by the secant line determined by the two points $(x_0, f(x_0))$ and $(x_1, f(x_1))$, and denote by x_2 the intersection of this line with the horizontal axis. Matching the slopes gives

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_1)}{x_1 - x_2},$$

so that

$$x_2 = x_1 - f(x_1)\frac{x_1 - x_0}{f(x_1) - f(x_0)}.$$

We can iterate this procedure and construct the sequence (x_n) by

$$x_{n+1} := x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \quad \text{for } n \geq 0.$$

The following theorem—we leave its proof as an exercise similar to that of Theorem 5.1.4—provides the accuracy of this algorithm.

Theorem 5.1.6. *Assume that the functions f, f' and f'' are continuous in a neighbourhood of x^* and that $f'(x^*) \neq 0$. If the initial guesses x_0 and x_1 are sufficiently close to x^*, then the algorithm converges to x^* with an order of convergence equal to $\left(1 + \sqrt{5}\right)/2$.*

5.1.4 The fixed-point algorithm

This last method encompasses other algorithms, in particular Newton-Raphson. This method is in fact a reformulation of the root-finding problem and is sometimes more convenient to use. We consider here an interval $[a, b] \subset \mathbb{R}$ and a continuous function ϕ on this interval. We wish to solve the fixed-point problem $\phi(x) = x$ on $[a, b]$. The iteration we consider here is

$$x_{n+1} = \phi(x_n), \quad \text{for } n \geq 0, \text{ where we start with an initial guess } x_0 \in [a, b]. \tag{5.1.2}$$

Note that taking $\phi(x) \equiv x - f(x)/f'(x)$ reduces this problem to a root finding issue in the Newton-Raphson framework. We prove a series of results concerning the quality of this algorithm. We shall assume from now on that $\phi\left([a, b]\right) \subset [a, b]$.

Lemma 5.1.7. *With the assumptions above, the equation $x = \phi(x)$ has at least one solution in the interval $[a, b]$.*

The proof simply follows by applying the intermediate value theorem to the function f defined by $f(x) := x - \phi(x)$ on $[a, b]$.

Lemma 5.1.8. *Assume that there exists* $\gamma \in (0, 1)$ *such that*

$$|\phi(x) - \phi(y)| \leq \gamma |x - y|, \quad \textit{for all } x, y \textit{ in } [a, b],$$

then the equation $\phi(x) = x$ *has a unique solution* x^* *in* $[a, b]$ *and the algorithm* (5.1.2) *converges to* x^* *for any initial guess* x_0 *with*

$$|x^* - x_n| \leq \frac{\gamma^n}{1 - \gamma} |x_1 - x_0|, \quad \textit{for any } n \geq 1.$$

Proof. Existence follows from Lemma 5.1.7. Assume that there exists another solution $y^* \in [a, b]$. Then $|\phi(x^*) - \phi(y^*)| = |x^* - y^*|$, which contradicts the existence of $\gamma \in (0, 1)$. For any $n \geq 1$, we can now write

$$|x^* - x_n| = |\phi(x^*) - \phi(x_{n-1})| \leq \gamma |x^* - x_{n-1}| \leq \ldots \leq \gamma^n |x^* - x_0|. \tag{5.1.3}$$

Since $\gamma \in (0, 1)$, this proves the convergence for any $x_0 \in [a, b]$. The triangle inequality then implies

$$|x^* - x_0| \leq |x^* - x_1| + |x_1 - x_0| \leq \gamma |x^* - x_0| + |x_1 - x_0|,$$

so that $|x^* - x_0| \leq (1 - \gamma)^{-1} |x_1 - x_0|$, and the lemma follows from (5.1.3). □

The following theorem is somehow a reformulation of the above lemmas, and we leave its straightforward proof as an exercise.

Theorem 5.1.9. *Assume that* $\phi \in C^1([a, b])$ *and that*

$$\gamma := \max_{x \in [a,b]} |\phi'(x)| < 1.$$

Then the equation $\phi(x) = x$ *has a unique solution* $x^* \in [a, b]$, *the sequence* (x_n) *defined in* (5.1.2) *converges to* x^* *for any* x_0 *and* $|x^* - x_n| \leq \frac{\gamma^n}{1-\gamma} |x_1 - x_0|$. *Furthermore*

$$\lim_{n \to \infty} \frac{x^* - x_{n+1}}{x^* - x_n} = \phi'(x^*).$$

5.2 Optimisation

5.2.1 Unconstrained optimisation

We are interested in this section in solving the following problem:

$$\min_{\mathrm{x} \in \mathbb{R}^n} f(\mathrm{x}),$$

subject to some constraints on the variables, where f is a map from $\mathbb{R}^n$ to $\mathbb{R}$, called the objective, or the cost function, which we shall assume to be smooth for simplicity. We first review some fundamental concepts of optimisation, before going into some details on some particular methods.

Definition 5.2.1. A point $\mathrm{z} \in \mathbb{R}^n$ is called a local minimiser if the inequality $f(\mathrm{z}) \leq f(\mathrm{x})$ holds locally around z, i.e. if there exists some $\varepsilon > 0$ such that $f(\mathrm{z}) \leq f(\mathrm{x})$ holds for all x in a ball with centre z and radius ε.

The definition of a global minimiser follows straightforwardly. When the function f is simple enough, it may be possible to compute the minimiser (if it exists) directly, i.e. to compute all the values $f(\mathrm{x})$ for x close enough to z. However in the general case the definition is impracticable and we shall see some more useful results to determine whether a point is a (local) minimiser or not.

Theorem 5.2.2 (First-order conditions)**.** *If* z *is a local minimiser and* f *is of class* C^1 *in a neighbourhood of* z*, then* $\nabla f(\mathrm{z}) = 0$.

Proof. Assume that $\nabla f(\mathrm{z}) \neq 0$, and define the vector $\mathrm{p} := -\nabla f(\mathrm{z})$. This implies the inequality $\mathrm{p}^T \nabla f(\mathrm{z}) = -\|\nabla f(\mathrm{z})\|^2 < 0$, and hence there exists $\overline{\alpha} > 0$ such that $\mathrm{p}^T \nabla f(\mathrm{z} + \alpha \mathrm{p}) < 0$ for all $\alpha \in [0, \overline{\alpha}]$. Let now fix some $\alpha \in [0, \overline{\alpha}]$. A Taylor expansion around the point z gives $f(\mathrm{z} + \alpha\mathrm{p}) = f(\mathrm{z}) + \alpha \mathrm{p}^T \nabla f(\mathrm{z} + \alpha_0 \mathrm{p})$ for some $\alpha_0 \in (0, \alpha)$. Therefore $f(\mathrm{z} + \alpha \mathrm{p}) < f(\mathrm{z})$, i.e. the function f decreases in the direction of the vector p, and the theorem follows by contradiction. □

Remark 5.2.3. The statement $\nabla f(\mathrm{z}) = 0$ does not necessarily imply that the point z is a minimiser. It however defines a *stationary point*. For instance, consider the function $f : (x_1, x_2) \mapsto x_1 x_2$ and the point $\mathrm{z} = (0, 0)$. It is clear that $\nabla f(\mathrm{z}) = 0$ but z is not a minimiser of the function.

Example. Recall the polynomial interpolation problem: we observe a function f at some points $x_1, \ldots, x_n$, and we want to find a polynomial P of the form $P(x) = \sum_{i=0}^{d-1} a_i x^i$, for some $\mathrm{a} := (a_0, \ldots, a_{d-1}) \in \mathbb{R}^d$, with $d < n$. We can write this as the following minimisation problem

$$\min_{\mathrm{a} \in \mathbb{R}^d} \sum_{k=1}^n (f(x_k) - P(x_k))^2 = \min_{\mathrm{a} \in \mathbb{R}^d} \phi(\mathrm{a}),$$

where we can write the function ϕ as $\phi(\mathrm{a}) := \mathrm{a}^T \mathrm{Q} \mathrm{a} - 2\mathrm{b}^T \mathrm{a} + c$, where $\mathrm{Q} = (q_{ij})$, $\mathrm{b} = (b_i)$ and

$$q_{ij} := \sum_{k=1}^n x_k^{i+j}, \quad b_j := \sum_{k=1}^n f(x_k) x_k^j, \quad c := \sum_{k=1}^n f(x_k)^2.$$

The first-order conditions read $\mathrm{Qa} = \mathrm{b}$.

The following theorem gives a necessary condition for the minimum:

Theorem 5.2.4 (Second-order necessary conditions)**.** *If* z *is a local minimiser of the function* f *and the Hessian matrix* $\nabla^2 f$ *exists and is continuous in a neighbourhood of* z*, then* $\nabla f(\mathrm{z}) = 0$ *and the matrix* $\nabla^2 f(\mathrm{z})$ *is positive semi-definite, i.e.* $\mathrm{p}^T \nabla^2 f(\mathrm{z}) \mathrm{p} \geq 0$ *for any* $\mathrm{p} \in \mathbb{R}^n$.

Proof. Assume that the Hessian is not positive semi-definite. Then there exists a vector $\mathrm{p} \in \mathbb{R}^n$ such that $\mathrm{p}^T \nabla^2 f(\mathrm{z}) \mathrm{p} < 0$. By continuity of the Hessian, we have $\mathrm{p}^T \nabla^2 f(\mathrm{z} + \alpha \mathrm{p}) \mathrm{p} < 0$ for sufficiently small $\alpha > 0$. A Taylor series expansion then gives, for some $\alpha_0 \in [0, \alpha]$,

$$f(\mathrm{z} + \alpha \mathrm{p}) = f(\mathrm{z}) + \alpha \mathrm{p}^T \nabla f(\mathrm{z}) + \frac{1}{2} \alpha \mathrm{p}^T \nabla^2 f(\mathrm{z} + \alpha_0 \mathrm{p}) \mathrm{p} < f(\mathrm{z}).$$

Again this implies that the function decreases (locally) in the direction of the vector p, which is a contradiction and the theorem follows. □

We now state the following sufficient conditions ensuring that z is a local minimiser, and we leave the proof as an exercise.

Theorem 5.2.5 (Second-order sufficient conditions)**.** *Assume that the Hessian matrix is continuous in a neighbourhood of* z *and that* $\nabla f(\mathrm{z}) = 0$*. Then the point* z *is a (strict) local minimiser if the Hessian matrix is positive definite.*

In the context of convex optimisation, many problems are simplified. Let us first recall that a function is convex in some domain $K \subset \mathbb{R}^n$ if

$$f\left(\alpha \mathrm{x} + (1-\alpha)\,\mathrm{y}\right) \leq \alpha f(\mathrm{x}) + (1-\alpha)\, f(\mathrm{y}),$$

for all $\alpha \in [0,1]$ and all $(\mathrm{x},\mathrm{y}) \in K \times K$. The following theorem is fundamental and clarifies how things get simpler in a convex setting.

Theorem 5.2.6. *For a convex function, any local minimiser is also global. Furthermore if the function is differentiable, then any stationary point is a global minimiser.*

Proof. Suppose that the point z is a local but not a global minimiser. There exists then z_0 such that $f(\mathrm{z}) > f(\mathrm{z}_0)$, and we consider the segment between z and z_0 (excluding z):

$$\mathrm{L} := \left\{\alpha \mathrm{z}_0 + (1-\alpha)\,\mathrm{z}, \quad \alpha \in (0,1]\right\}.$$

Note that the convexity of f implies that $f(\mathrm{x}) < f(\mathrm{z})$ for any $\mathrm{x} \in \mathrm{L}$. Fix a point $\mathrm{x} \in \mathrm{L}$, then

$$\begin{aligned}
\nabla f(\mathrm{z}) \cdot (\mathrm{x}-\mathrm{z}) &= \left.\frac{\mathrm{d}}{\mathrm{d}\alpha} f\left(\mathrm{z} + \alpha\left(\mathrm{x}-\mathrm{z}\right)\right)\right|_{\alpha=0} \\
&= \lim_{\alpha \to 0} \frac{f\left(\mathrm{z}+\alpha\left(\mathrm{x}-\mathrm{z}\right)\right) - f(\mathrm{z})}{\alpha} \\
&\leq \lim_{\alpha \to 0} \frac{\alpha f(\mathrm{x}) + (1-\alpha)\, f(\mathrm{z}) - f(\mathrm{z})}{\alpha}, \quad \text{by convexity} \\
&= f(\mathrm{x}) - f(\mathrm{z}) < 0,
\end{aligned}$$

and hence z cannot be a stationary point and the theorem follows by contradiction. □

Exercise 52. Consider the parametric function $f_\alpha(\mathrm{x}) := \frac{1}{2}\mathrm{x}^T \mathrm{H}_\alpha \mathrm{x} - \mathrm{x}^T \mathrm{b}$, where

$$\mathrm{H}_\alpha := \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix} \quad \text{and} \quad \mathrm{b} := \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and $\alpha \in \mathbb{R}$. Discuss the existence of local and global minimisers.

Exercise 53. Consider the function $f(x) \equiv \cos(\pi x)$. Implement and discuss the application of Newton's algorithm to find a minimiser of f, depending on the starting point of the algorithm.

5.2.2 Line search methods

The idea of line search method is to construct a sequence $(\mathrm{x}_k)_{k\geq 0}$ converging to the minimiser x^* according to the rule

$$\mathrm{x}_{k+1} = \mathrm{x}_k + \alpha_k \mathrm{p}_k, \quad \text{for all } k \geq 0,$$

where we start with an initial guess x_0 and where α_k is a positive scalar called the *step length* and p_k is a vector indicating the direction of the search. Since we want to solve a minimisation problem, the vector p_k will be a descent vector, i.e. such that the function decreases in its direction. This procedure now requires the computation of α_k and the direction vector p_k. A sensible choice for α_k is to find the minimiser of the one-dimensional function ϕ defined by $\phi(\alpha) := f(\mathrm{x}_k + \alpha \mathrm{p}_k)$. However, it is usually too computer-intensive to determine such a minimum. A lighter requirement is to find α such that, locally, the function f decreases, i.e. find α such that $f(\mathrm{x}_k + \alpha \mathrm{p}_k) < f(\mathrm{x}_k)$. This approach does not however converge to the minimiser x^*. A popular line search method is to provide *sufficient* decrease to the function f via

$$\phi(\alpha) = f(\mathrm{x}_k + \alpha \mathrm{p}_k) < f(\mathrm{x}_k) + \eta_1 \alpha \nabla f(\mathrm{x})^T \mathrm{p}_k, \tag{5.2.1}$$

for some $\eta_1 \in (0,1)$. Note that this is tantamount to finding α such that the graph of the function ϕ lies below the line given by the right-hand side of (5.2.1). This inequality is called the *Armijo condition* and imposes a decrease of the function proportional to the step length and the directional derivative. It can be shown that this condition can be satisfied for very small values of the parameter α, which clearly implies a very slow convergence (if any). One can further impose a *curvature condition* as

$$\phi'(\alpha) = \nabla f(\mathrm{x}_k + \alpha \mathrm{p}_k)^T \mathrm{p}_k \geq \eta_2 \nabla f(\mathrm{x}_k)^T \mathrm{p}_k, \tag{5.2.2}$$

for some $\eta_2 \in (\eta_1, 1)$. We leave the geometrical interpretation of this result to the reader. Conditions (5.2.1) and (5.2.2) are called the *Wolfe conditions* and are fundamental in determining an optimal algorithm for the minimisation problem. The following lemma ensures the existence of step lengths.

Lemma 5.2.7. *Suppose that the function f is continuously differentiable and assume that it is bounded along any ray $\{\mathrm{x}_k + \alpha \mathrm{p}_k, \alpha > 0\}$. If $0 < \eta_1 < \eta_2 < 1$, then there exist intervals of step lengths satisfying the Wolfe conditions* (5.2.1) *and* (5.2.2).

Proof. By definition and assumption, the function ϕ is bounded below for any $\alpha > 0$. For any $k \geq 0$, define the linear functional

$$\mathrm{L}_k(\alpha) := f(\mathrm{x}_k) + \alpha \eta_1 \mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k), \quad \text{for any } \alpha > 0.$$

It is clear that L_k is a decreasing function, and hence its graph has (at least) an intersection with the graph of ϕ. If we denote $\alpha_1 > 0$ the smallest of these, it satisfies

$$f(\mathrm{x}_k + \alpha_1 \mathrm{p}_k) = f(\mathrm{x}_k) + \alpha_1 \eta_1 \mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k).$$

The mean value theorem then implies the existence of some $\alpha_2 \in (0, \alpha_1)$ such that

$$f(\mathrm{x}_k + \alpha_1 \mathrm{p}_k) - f(\mathrm{x}_k) = \alpha_1 \mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k + \alpha_2 \mathrm{p}_k),$$

and straightforward algebra shows that the point α_2 satisfies the Wolfe conditions. □

The importance of the Wolfe conditions is revealed in the following theorem.

Theorem 5.2.8 (Zoutendijk Theorem (see [59]))**.** *We consider the sequence $(\mathrm{x}_k)_{k\geq 0}$ defined as above, where α_k satisfies the two Wolfe conditions* (5.2.1) *and* (5.2.2)*, and p_k is a descent direction. Assume that there exists an open set $K \subset \mathbb{R}^n$ such that $f \in C^1(K)$ and $\{\mathrm{x} : f(\mathrm{x}) \leq f(\mathrm{x}_0)\} \subset K$. If there exists some constant $\gamma > 0$ such that*

$$\|\nabla f(\mathrm{x}) - \nabla f(\mathrm{y})\| \leq \gamma \|\mathrm{x} - \mathrm{y}\|, \quad \textit{for all } \mathrm{x}, \mathrm{y} \textit{ in } K,$$

then

$$\sum_{k \geq 0} \left(\frac{\mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k)}{\|\nabla f(\mathrm{p}_k)\|} \right)^2 < \infty.$$

Remark 5.2.9. Note that if we denote θ_k the angle between the vector p_k and the direction of steepest descent $-\nabla f(\mathrm{x}_k)$, then

$$\cos(\theta_k) = -\frac{\mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k)}{\|\nabla f(\mathrm{x}_k)\| \, \|\nabla f(\mathrm{p}_k)\|}, \tag{5.2.3}$$

so that the last statement of the theorem reads $\sum_{k\geq 0} \cos^2(\theta_k) \|\nabla f(\mathrm{x}_k)\|^2 < \infty$, which implies that we must have $\lim_{k\to\infty} \left(\cos^2(\theta_k) \|\nabla f(\mathrm{x}_k)\|^2\right) = 0$. In particular, if p_k is the steepest descent direction $-\nabla f(\mathrm{x}_k)$, then $\theta_k = 0$, and the theorem implies that $\lim_{k\to\infty} \|\nabla f(\mathrm{x}_k)\|^2 = 0$, so that we have convergence of the algorithm.

Proof of Theorem 5.2.8. From the second Wolfe condition, we have

$$\mathrm{p}_k^{\mathrm{T}} \left(\nabla f(\mathrm{x}_{k+1}) - \nabla f(\mathrm{x}_k)\right) \geq (\eta_2 - 1) \, \mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k).$$

Since the gradient of the function f is Lipschitz continuous, Cauchy-Schwarz inequality gives

$$\mathrm{p}_k^{\mathrm{T}} \left(\nabla f(\mathrm{x}_{k+1}) - \nabla f(\mathrm{x}_k)\right) \leq \alpha_k \gamma \|\mathrm{p}_k\|^2,$$

which can be rewritten as

$$\alpha_k \geq \frac{\eta_2 - 1}{\gamma} \frac{\mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k)}{\|\mathrm{p}_k\|^2}.$$

Combining this with the first Wolfe condition leads to

$$f(\mathrm{x}_{k+1}) \leq f(\mathrm{x}_k) - \eta_1 \frac{\eta_2 - 1}{\gamma} \left(\frac{\mathrm{p}_k^{\mathrm{T}} \nabla f(\mathrm{x}_k)}{\|\mathrm{p}_k\|} \right)^2,$$

and hence

$$f(\mathrm{x}_{k+1}) \leq f(\mathrm{x}_k) - \eta \cos^2 \left(\theta_k \left\| \nabla f(\mathrm{x}_k) \right\|^2 \right),$$

where the angle θ_k is defined in (5.2.3) and where $\eta := \eta_1(1 - \eta_2)/\gamma$. We can now iterate this inequality to obtain

$$f(\mathrm{x}_{k+1}) \leq f(\mathrm{x}_0) - \eta \sum_{i=0}^{k} \cos^2 \left(\theta_i \left\| \nabla f(\mathrm{x}_i) \right\|^2 \right).$$

Since the function f is bounded below the theorem follows by letting k tend to infinity. □

Let us look at a simplified example of the form

$$f(\mathrm{x}) := \frac{1}{2} \mathrm{x}^T \mathrm{Q} \mathrm{x} - \mathrm{b}^T \mathrm{x}, \tag{5.2.4}$$

where Q is a symmetric positive definite matrix and b a given vector. The gradient of the function reads $\nabla f(\mathrm{x}) = \mathrm{Qx} - \mathrm{b}$, and the unique minimiser is the solution to the matrix equation $\mathrm{Qx} = \mathrm{b}$. If we wish to compute the optimal α in order to set up a line search algorithm, we need to minimise the map $\alpha \mapsto f(\mathrm{x} + \alpha \mathrm{p}_k)$ along a direction vector p_k. It is clear that the *steepest descent* vector at the point x_k is given by $\mathrm{p}_k = -\nabla f(\mathrm{x}_k)$. Since

$$f\left(\mathrm{x}_k - \alpha \nabla f(\mathrm{x}_k)\right) = \frac{1}{2} \left(\mathrm{x}_k - \alpha \nabla f(\mathrm{x}_k)\right)^T \mathrm{Q} \left(\mathrm{x}_k - \alpha \nabla f(\mathrm{x}_k)\right) - \mathrm{b}^T \left(\mathrm{x}_k - \alpha \nabla f(\mathrm{x}_k)\right),$$

the minimiser α^* is clearly equal to

$$\alpha^* = \frac{\nabla f(\mathrm{x}_k)^T \nabla f(\mathrm{x}_k)}{\nabla f(\mathrm{x}_k)^T \mathrm{Q} \nabla f(\mathrm{x}_k)}.$$

Note that this leads to a fully explicit expression for x_{k+1} as a function of x_k. Using the fact that the minimiser satisfies $\mathrm{Qx}^* = \mathrm{b}$, it is easy to show that

$$\frac{1}{2} \left\| \mathrm{x} - \mathrm{x}^* \right\|_{\mathrm{Q}}^2 = f(\mathrm{x}) - f(\mathrm{x}^*),$$

where the $\| \cdot \|_{\mathrm{Q}}$-norm is defined as $\|\mathrm{x}\|_{\mathrm{Q}}^2 := \mathrm{x}^T \mathrm{Q} \mathrm{x}$. This norm is a useful tool in order to quantify the rate of convergence of the algorithm and the following theorem provides a precise convergence result. We refer the interested reader to [47] for full details.

Theorem 5.2.10. *In the strongly convex problem* (5.2.4)*, the error norm in the steepest descent algorithm satisfies*

$$\left\| \mathrm{x}_{k+1} - \mathrm{x}^* \right\|_{\mathrm{Q}}^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right) \left\| \mathrm{x}_k - \mathrm{x}^* \right\|_{\mathrm{Q}}^2,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the matrix Q.

Note that when the matrix Q is proportional to the identity matrix, then there is a unique eigenvalue (with multiplicity n). In that case, the contour plot of the function consists of circles and not ellipsis, and the steepest descent direction always points directly to the minimiser. In this case, the convergence is further attained after only one iteration.

5.2.3 Minimisation via the Newton method

From the first and second order conditions proved in the previous section, the minimisation problem $\min_{\mathrm{x}\in\mathbb{R}^n} f(\mathrm{x})$ is tantamount to solving the non-linear equation $\nabla f(\mathrm{x}) = 0$, where the function f maps $\mathbb{R}^n$ to $\mathbb{R}$. This is indeed true as soon as the function f is convex. In the same spirit as the Newton algorithm above, consider a Taylor expansion around a point x—close to the minimiser z—in the direction of a vector $\mathrm{p} \in \mathbb{R}^n$:

$$f(\mathrm{z}) = f(\mathrm{x}+\mathrm{p}) = f(\mathrm{x}) + \mathrm{p}^T\nabla f(\mathrm{x}) + \mathcal{O}\left(\|\mathrm{p}\|^2\right),$$

which, together with the condition $\nabla f(\mathrm{z}) = 0$, implies that

$$\nabla f(\mathrm{x}) + \mathrm{p}^T\nabla^2 f(\mathrm{x}) + \mathcal{O}\left(\|\mathrm{p}\|^2\right) = \nabla f(\mathrm{z}) = 0.$$

This leads to $\mathrm{p} \approx -\left(\nabla^2 f(\mathrm{x})\right)^{-1}\nabla f(\mathrm{x})$. Pick now an initial guess $\mathrm{x}_0 \in \mathbb{R}^n$ and define the sequence $(\mathrm{x}_k)_{k\geq 0}$ recursively by

$$\mathrm{x}_{k+1} = \mathrm{x}_k - \left(\nabla^2 f(\mathrm{x}_k)\right)^{-1}\nabla f\left(\mathrm{x}_k\right). \tag{5.2.5}$$

Note that this is nothing else than a line search method, where the direction vector is given by $\mathrm{p} = -\left(\nabla^2 f(\mathrm{x}_k)\right)^{-1}\nabla f\left(\mathrm{x}_k\right)$, and hence is indeed of steepest descent. Note further the analogy with the Newton-Raphson developed on the real line in Section 5.1.2. This matrix equation is highly unstable and it may be wiser to solve $\nabla f(\mathrm{x}_k)\left(\mathrm{x}_{k+1} - \mathrm{x}_k\right) = -f\left(\mathrm{x}_k\right)$. We can now state the main theorem, which is an extension of Theorem 5.1.4.

Theorem 5.2.11. *Assume that f is twice continuously differentiable and that there exists $\mathrm{z} \in \mathbb{R}^n$ such that $\nabla f(\mathrm{z}) = 0$ and $\nabla^2 f(\mathrm{z})$ is positive definite, then the iteration (5.2.5) converges to z if f is convex. Alternatively, if the function f is only locally convex in a neighbourhood of z and the initial guess x_0 lies in this neighbourhood, then the algorithm converges. Furthermore, the converges is quadratic and the sequence $\left(|\nabla f(\mathrm{x}_k)|\right)_{k\geq 0}$ converges to zero quadratically.*

The proof follows similar steps as in the one-dimensional case and we omit it. We now give an example of the Newton method in two dimensions applied to a problem arising in economic theory.

Example (Cournot equilibrium)**.** The Cournot equilibrium is a classic problem in economic theory. Consider two companies producing the same product in quantities q_1 and q_2. The total cost of producing this product is $C_i(q_i) := \frac{1}{2}\gamma_i q_i^2$ for $i = 1, 2$, where γ_1 and γ_2 are two strictly positive constants. The price per unit of product is $P(q_1 + q_2) = (q_1 + q_2)^{-1/\alpha}$ for some $\alpha > 0$. The Cournot equilibrium corresponds to the state where both profits are maximised, where the profit π_i is defined by

$$\pi_i(q_1, q_2) = P(q_1 + q_2) - C_i(q_i), \quad \text{for } i = 1, 2.$$

The first-order conditions read $\partial_{q_1}\pi_1(q_1, q_2) = \partial_{q_2}\pi_2(q_1, q_2) = 0$.

5.2.4 Constrained optimisation

Consider a function f from $\mathbb{R}^n$ to $\mathbb{R}$ and the optimisation problem

$$\min_{\mathrm{x}\in\mathbb{R}^n} f(\mathrm{x}) \quad \text{subject to} \quad \mathrm{g}(\mathrm{x}) = \mathrm{b}, \tag{5.2.6}$$

where $\mathrm{g}(\mathrm{x}) := (g_1(\mathrm{x}), \ldots, g_m(\mathrm{x}))$ and $\mathrm{b} \in \mathbb{R}^m$. Consider the Lagrange function

$$\mathrm{L}(\mathrm{x}, \lambda) := f(\mathrm{x}) + \lambda^\top (\mathrm{g}(\mathrm{x}) - \mathrm{b}),$$

where $\lambda := (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ are called the *Lagrange multipliers.* The theory of Lagrange multipliers states that the optimisation problem (5.2.6) is tantamount to finding the stationary points of the functional L on the extended state space $\mathbb{R}^n \times \mathbb{R}^m$, i.e. to solving

$$\nabla \mathrm{f}(\mathrm{x}, \lambda) = \left(\frac{\partial \mathrm{L}}{\partial x_1}, \ldots, \frac{\partial \mathrm{L}}{\partial x_n}, \frac{\partial \mathrm{L}}{\partial \lambda_1}, \ldots, \frac{\partial \mathrm{L}}{\partial \lambda_m}\right)^\top = \begin{pmatrix} \nabla f(\mathrm{x}) + \mathrm{Dg}(\mathrm{x})^\top \lambda \\ \mathrm{g}(\mathrm{x}) - \mathrm{b} \end{pmatrix} = 0,$$

where Dg is the Jacobian matrix $(\partial_{x_i} g_j)_{1\leq i\leq n}^{1\leq j\leq m}$. The Hessian of the Lagrange function reads

$$\nabla^2 \mathrm{L}(\mathrm{x}, \lambda) = \begin{pmatrix} \nabla^2 f + \sum_{i=1}^m \lambda_i \nabla^2 g_i & (\mathrm{Dg})^\top \\ \mathrm{Dg} & 0 \end{pmatrix} (\mathrm{x}, \lambda).$$

We can then solve the minimisation problem by solving the equation $\nabla \mathrm{L}(\mathrm{x}, \lambda) = 0$ The whole theory of constrained optimisation is a wide topic and time constraints do not allow us such a detour. We hence leave it as the root-finding problem of the gradient of the Lagrange function on the extended state space $\mathbb{R}^n \times \mathbb{R}^m$.

5.2.5 Constrained optimisation

We now consider the following constrained optimisation problem:

$$\min_{\mathrm{x}\in\mathbb{R}^n} f(\mathrm{x}), \quad \text{subject to } \mathrm{h}(\mathrm{x}) = 0,$$

where $\mathrm{h} = (h_1, \ldots, h_m)^\top$, with each map $h_i : \mathbb{R} \to (0, \infty)$ assumed to be continuously differentiable. For any $i = 1, \ldots, m$, the constraint $h_i(\mathrm{x}) = 0$ defines a hypersurface $\mathcal{S} \subset \mathbb{R}^n$. Consider now a (smooth) curve $(x(t))_{0\leq t\leq 1}$ lying on $\mathcal{S}$, a point $x^* \in \mathcal{S}$, such that there exists $t^* \in [0, 1]$ for which $x(t^*) = x^*$. The vector $\dot{x}(t^*) := \frac{\mathrm{d}x(t)}{\mathrm{d}t}|_{t=t^*}$ is called the tangent vector of the curve $x(\cdot)$ at the point (t^*, x^*). The tangent space $\mathcal{T}$ at the point $x^* \in \mathcal{S}$ is then defined as the subspace of $\mathbb{R}^n$ spanned by all tangent vectors $\dot{x}(t^*)$.

Definition 5.2.12. A point $\mathrm{x} \in \mathbb{R}^n$ satisfying the constraints $h(\mathrm{x}) = 0$ is called regular if the vectors $\nabla h_1(\mathrm{x}), \ldots, \nabla h_m(\mathrm{x})$ are linearly independent (the matrix $\nabla \mathrm{h}$ has full rank).

Note that, at a point $\mathrm{x} \in \mathbb{R}^n$, $\nabla h(\mathrm{x}) = (\nabla h_1(\mathrm{x}), \ldots \nabla h_m(\mathrm{x}))^\top$ is a matrix in $\mathcal{M}_{mn}(\mathbb{R})$. We can then state the following result:

Theorem 5.2.13. *Let* x *be a regular point on the hypersurface* $\{\mathrm{x} : h(\mathrm{x}) = 0\}$. *The tangent space* $\mathcal{T}$ *is then the nullspace of the matrix* ∇h*:* $\mathcal{T} = \{\mathrm{y} \in \mathbb{R}^n : \nabla h(\mathrm{x})\mathrm{y} = 0\}$.

Example. Let $n = 2$, $m = 1$, and consider the function $h(\mathrm{x}) \equiv x_1^2$. Therefore $\mathcal{S} := \{\mathrm{x} = (x_1, x_2) \in \mathbb{R}^2 : h(\mathrm{x}) = 0\} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0\}$. The gradient reads $\nabla h(\mathrm{x}) = (2x_1, 0)$ and is null everywhere on the curve $\mathcal{S}$, so that the set of regular points is empty, and the tangent space at $\mathrm{x} \in \mathcal{S}$ reads $\mathcal{T} = \{(y_1, y_2) \in \mathbb{R}^n : \nabla(\mathrm{x})\mathrm{y} = 2x_1y_1 = 0\} = \mathbb{R}^2$.

We can now state the first-order and second-order necessary, and the sufficient, conditions for a minimum of f to exist:

Theorem 5.2.14. *Let* $\mathrm{x}^* \in \mathbb{R}^n$ *be a local minimum of the function* f *satisfying the constraints* $h(\mathrm{x}) = 0$, *and assume that* x^* *is a regular point. Then the following hold:*

- *there exists* $\lambda \in \mathbb{R}^m$ *such that* $\nabla f(\mathrm{x}^*) + \lambda^\top \nabla h(\mathrm{x}^*) = 0$*; furthermore, the matrix*

$$\mathrm{H}(\mathrm{x}^*) := \nabla^2 f(\mathrm{x}^*) + \lambda^\top \nabla^2 h(\mathrm{x}^*)$$

  is positive semidefinite on the tangent space $\mathcal{T}(\mathrm{x}^*) = \{\mathrm{y} \in \mathbb{R}^n : \nabla h(\mathrm{x}^*)\mathrm{y} = 0\}$.

Theorem 5.2.15. *Assume that there exists* $\mathrm{x}^* \in \mathbb{R}^n$ *and a vector* $\lambda \in \mathbb{R}^m$ *such that*

$$h(\mathrm{x}^*) = 0 = \nabla f(\mathrm{x}^*) + \lambda^\top \nabla h(\mathrm{x}^*).$$

If furthermore there exists a positive definite matrix $H(\mathrm{x}^*)$ *on the tangent space* $\mathcal{T}(\mathrm{x}^*)$, *then* x^* *is a strict local minimum of* f *satisfying the constraints.*

Chapter 6

Linear programming and duality

6.1 Separation theorems

We consider here a subset $\mathcal{S}$ of $\mathbb{R}^n$ $(n \geq 1)$, and define the distance to this set by

$$d_{\mathcal{S}}(\mathrm{z}) := \inf \{\|s - \mathrm{z}\|, s \in \mathcal{S}\},$$

where $\|\cdot\|$ denotes the usual Euclidian distance norm. We recall that $\mathcal{S}$ is said to be convex if, for any $x, y \in \mathcal{S}$, the line $\{\lambda x + (1-\lambda)y, \lambda \in [0,1]\}$ lies in $\mathcal{S}$. A point $s_0 \in \mathcal{S}$ is called the nearest point of $\mathcal{S}$ to z if $\|s_0 - \mathrm{z}\| = d_{\mathcal{S}}(\mathrm{z})$.

Lemma 6.1.1. *Let $C \subset \mathbb{R}^n$ be a non empty closed convex set. Then, for any $\mathrm{z} \in \mathbb{R}^n$, there exists a unique nearest point to $\mathrm{z} \in C$.*

Proof. Recall that the set $\mathcal{S}$ is closed if it contains the limit points of every convergent sequence in $\mathcal{S}$. This essentially implies the existence of a nearest point. Assume now that both x and y are nearest points to $z \in \mathcal{S}$, and define $\delta := d_{\mathcal{S}}(z) = \|z - x\| = \|z - y\|$, i.e. x and y both lie on the boundary of the closed ball $B_\delta(z) := \{y \in \mathbb{R}^n : \|y - z\| \leq \delta\}$. The point $(x+y)/2$ also lies in $\mathcal{S}$ by convexity and in the interior of $B_\delta(z)$, which is a contradiction. □

For any given non-zero vector a and a real number α, we define the hyperplane $\mathcal{H}_{\mathrm{a},\alpha} \subset \mathbb{R}^n$ by

$$\mathcal{H} := \left\{\mathrm{x} \in \mathbb{R}^n : \mathrm{a}^\top \mathrm{x} = \alpha\right\},$$

and a is called the normal vector of $\mathcal{H}_{a,\alpha}$. For instance, in $\mathbb{R}^2$, any hyperplane is a line, and in $\mathbb{R}^3$, any hyperplane is a plane. We shall further define the half-spaces

$$\mathcal{H}^+_{\mathrm{a},\alpha} := \left\{\mathrm{x} \in \mathbb{R}^n : \mathrm{a}^\top \mathrm{x} \leq \alpha\right\} \qquad \text{and} \qquad \mathcal{H}^-_{\mathrm{a},\alpha} := \left\{\mathrm{x} \in \mathbb{R}^n : \mathrm{a}^\top \mathrm{x} \geq \alpha\right\}.$$

Theorem 6.1.2. *Let $\mathcal{C} \subset \mathbb{R}^n$ be a non-empty closed convex set and $z \in \mathbb{R}^n \setminus \mathcal{C}$. Then z and $\mathcal{C}$ can be strongly separated, i.e., there exists a hyperplane $\mathcal{H}_{a,\alpha}$ such that $z \in \mathcal{H}^+_{\mathrm{a},\alpha}$ and $\mathcal{C} \subset \mathcal{H}^-_{\mathrm{a},\alpha}$.*

Proof. Let p be the nearest point to z in $\mathcal{C}$. For any $\mathrm{x} \in \mathcal{C}$, the open segment $\{(1-\lambda)\mathrm{p} + \lambda\mathrm{x}, \lambda \in (0,1)\}$ also belongs to $\mathcal{C}$ by convexity, and

$$\|\mathrm{p} - \mathrm{z} + \lambda(\mathrm{x} - \mathrm{p})\| = \|(1-\lambda)\mathrm{p} + \lambda\mathrm{x} - \mathrm{z}\| \geq \|\mathrm{p} - \mathrm{z}\|.$$

Squaring both sides, we obtain, applying the triangle inequality:

$$\|\mathrm{p} - \mathrm{z}\|^2 + 2\lambda(\mathrm{p} - \mathrm{z})^\top(\mathrm{x} - \mathrm{p}) + \lambda^2\|\mathrm{x} - \mathrm{p}\|^2 \geq \|\mathrm{p} - \mathrm{z}\|^2,$$

so that, as λ tends to zero, $(\mathrm{z} - \mathrm{p})^\top(\mathrm{x} - \mathrm{p}) \leq 0$. Define now the vector $\mathrm{a} := \mathrm{z} - \mathrm{p}$ and the real number $\alpha := \mathrm{a}^\top\mathrm{p}$, and consider the hyperplane

$$\mathcal{H}_{\mathrm{a},\alpha} := \{\mathrm{x} \in \mathbb{R}^n : \mathrm{a}^\top\mathrm{x} = \alpha\} = \{\mathrm{x} \in \mathbb{R}^n : (\mathrm{z} - \mathrm{p})^\top\mathrm{x} = \alpha\}.$$

Then clearly the set $\mathcal{C}$ belongs to the half space $\mathcal{H}^-_{\mathrm{a},\alpha}$, while z does not. Let now $\mathcal{H}^*$ be the hyperplane parallel to $\mathcal{H}_{\mathrm{a},\alpha}$ (i.e. with the same normal vector) and containing the point $\frac{1}{2}(\mathrm{z} + \mathrm{p})$. Then $\mathcal{H}^*$ strongly separates z and $\mathcal{C}$, and the theorem follows. □

We now state and prove the following fundamental lemma. For a vector $\mathrm{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we shall write $\mathrm{x} \geq \mathcal{O}_n$ whenever $x_i \geq 0$ for all $i = 1, \ldots, n$; the vector $\mathcal{O}_n$ on its own shall denote the $\mathbb{R}^n$-vector with null entries.

Theorem 6.1.3 (Farkas' lemma)**.** *Let* $A \in \mathcal{M}_{m,n}(\mathbb{R})$ *and* $\mathrm{b} \in \mathbb{R}^m$*. There exists a vector* $\mathrm{x} \geq \mathcal{O}_n$ *satisfying* $A\mathrm{x} = \mathrm{b}$ *if and only if for each* $\mathrm{y} \in \mathbb{R}^m$ *with* $\mathrm{y}^\top A \geq \mathcal{O}_n$*, it also holds that* $\mathrm{y}^\top\mathrm{b} \geq 0$.

Proof. Let $a_1, \ldots, a_n$ denote the column vectors of the matrix A, and define the (closed) convex cone[1] generated by these vectors:

$$\mathcal{C} := \left\{\sum_{i=1}^n \lambda_j a_j : \lambda \geq \mathcal{O}_n\right\} \subset \mathbb{R}_m.$$

Clearly, the equation $A\mathrm{x} = \mathrm{b}$ admits a non-negative solution if and only if $\mathrm{b} \in \mathcal{C}$. Assume now that x is such a solution satisfying $\mathrm{x} \geq \mathcal{O}_n$. If a vector $\mathrm{y} \in \mathbb{R}^m$ is such that $\mathrm{y}^\top A \geq \mathcal{O}_n$, then $\mathrm{y}^\top\mathrm{b} = \mathrm{y}^\top(A\mathrm{x}) = (\mathrm{y}^\top A)\mathrm{x} \geq 0$. Conversely, if the equation $A\mathrm{x} = \mathrm{b}$ does not admit any non-negative solution, then $\mathrm{b} \notin \mathcal{C}$, so that Theorem 6.1.2 implies that the vector b and the set $\mathcal{C}$ can be strongly separated, i.e. there exists $\mathrm{y} \neq \mathcal{O}_m$ and $\gamma \in \mathbb{R}$ such that

$$\mathrm{y}^\top\mathrm{x} \geq \gamma, \quad \text{for each } \mathrm{x} \in \mathcal{C}, \qquad \text{and} \qquad \mathrm{y}^\top\mathrm{b} < \gamma. \tag{6.1.1}$$

Since the vector null clearly belongs to $\mathcal{C}$, necessarily $\gamma \leq 0$. Now, $\mathrm{y}^\top\mathrm{x} \geq 0$ for each $\mathrm{x} \in \mathcal{C}$: indeed, if there exists $\mathrm{x} \in \mathcal{C}$ such that $\mathrm{y}^\top\mathrm{x} < 0$, then there exists $\lambda\mathrm{x} \in \mathcal{C}$ such that $\mathrm{y}^\top(\lambda\mathrm{x}) < \gamma$ for some $\lambda > 0$, which is obviously a contradiction. Finally, for any $i = 1, \ldots, n$, $\mathrm{y}^\top a_i \geq 0$ (because $a_i \in \mathcal{C}$), so that clearly $\mathrm{y}^\top A \geq \mathcal{O}_n$. Since $\mathrm{y}^\top\mathrm{b} < 0$ by (6.1.1), the lemma follows. □

[1] Recall that $\mathcal{C}$ is called a convex cone if the set $\{\lambda_1 x_1 + \lambda_2 x_2 : x_1, x_2 \in \mathcal{C}, \lambda_1, \lambda_2 \geq 0\}$ belongs to $\mathcal{C}$.

Remark 6.1.4. The geometric interpretation of Farkas' lemma[2] is that the following two statements are equivalent:

(i) the vector b belongs to the cone $\mathcal{C}$;

(ii) it is not possible to find a hyperplane $\mathcal{H}_{\cdot,0}$ that separates b and $\mathcal{C}$,

or equivalently

(i) there exists $\mathrm{x} \geq \mathcal{O}_n$ such that $A\mathrm{x} = \mathrm{b}$;

(ii) there exists a vector $\mathrm{y} \in \mathbb{R}^m$ such that $\mathrm{y}^\top A \geq 0$ and $\mathrm{y}\top\mathrm{b} < 0$.

6.2 Linear Programming Duality

A linear problem (which we shall call by convention the Primal Problem) has the following form:

$$\text{(Primal Problem)} \qquad \begin{aligned} &\sup \mathrm{c}^\top \mathrm{x} \\ &\text{subject to } A\mathrm{x} \leq \mathrm{b}, \end{aligned} \tag{6.2.1}$$

where $\mathrm{c} \in \mathbb{R}^n$, $\mathrm{b} \in \mathbb{R}^m$, $\mathrm{x} \in \mathbb{R}^n$ and $A \in \mathcal{M}_{m,n}(\mathbb{R})$. The problem is called *feasible* if there exists a vector $\mathrm{x} \in \mathbb{R}^n$ such that $A\mathrm{x} \leq \mathrm{b}$, namely if the constraint is satisfied. A feasible solution x_0 is further said to be optimal if $\mathrm{c}^\top \mathrm{x}_0 = \sup\{\mathrm{c}^\top \mathrm{x} : A\mathrm{x} \leq \mathrm{b}\}$. We allow the value $\pm\infty$ for the problem, namely $-\infty$ whenever the problem is not feasible, and $+\infty$ if the problem is unbounded, e.g. if there exists a sequence of feasible solutions $(\mathrm{x}_k)_{k\in\mathbb{N}}$ such that $\mathrm{c}^\top \mathrm{x}$ diverges to infinity as k tends to infinity. To each (primal) problem, one can associate a dual version, as follows:

$$\text{(Dual Problem)} \qquad \begin{aligned} &\inf \mathrm{b}^\top \mathrm{y} \\ &\text{subject to } A^\top \mathrm{y} = \mathrm{c} \text{ and } \mathrm{y} \geq 0_m. \end{aligned} \tag{6.2.2}$$

Theorem 6.2.1 (Duality Theorem)**.**

(i) If the primal problem (6.2.2) *has an optimal solution, then the dual solution has one as well, and there is no duality gap (both problems have the same value);*

(ii) If one of the problems is unbounded, then the other is not feasible; when at least one problem is feasible, then there is no duality gap.

Proof. □

Remark 6.2.2. The case b = 0 is of particular interest, since, in that case, the objective function in the dual problem is zero. Optimal solutions and feasible solutions are therefore equivalent.

[2] Gyula Farkas (March 28, 1847 December 27, 1930) was a Hungarian mathematician and physicist.

6.3 Application to the fundamental theorem of asset pricing

In the late seventies (or the twentieth century), no-arbitrage arguments became a central study in the understanding of financial market theory. The seminal papers by Harrison, Pliska and Kreps [33, 34, 35] were devoted to finite probability spaces (for example the Cox-Ross-Rubinstein's binomial model), and a full theory for general probability spaces was uncovered by Delbaen and Schachermayer [17, 18]. We are interested here in a simple finite-dimensional framework, in which linear programming plays a central role. Let us consider a market consisting of n assets with m possible scenarios (finite probability space), and denote $\Pi := (\pi_{i,j})_{i=1,\dots,m;j=1,\dots,n}$ the payoff matrix. A vector $\mathrm{h} \in \mathbb{R}^n$ will denote a trading strategy, and $\mathrm{x} \in \mathbb{R}^m$ a payoff, i.e. the outcome of a trading strategy.

Definition 6.3.1. A risk-neutral probability is a vector $\mathrm{y} \geq \mathcal{O}_m$ such that $\displaystyle\sum_{i=1}^m y_i = 1$ and $\mathrm{y}\Pi = \mathcal{O}_n$.

The linear programming problem reads as follows:

$$\text{(Primal Problem)} \qquad \begin{array}{l} \displaystyle\max \sum_{i=1}^m x_i \\ \text{subject to } \mathrm{x} = \Pi\mathrm{h} \text{ and } \mathrm{x} \geq \mathcal{O}_m. \end{array} \tag{6.3.1}$$

Definition 6.3.2. An arbitrage exists if and only if the optimal value of the linear programming problem (6.3.1) is strictly positive.

Theorem 6.3.3. *There is no arbitrage if and only if there exists a risk-neutral probability measure.*

Proof. Let I_n and e_n denote respectively the identity matrix and the unit vector (in dimension n), and rewrite the linear programming problem in the following form:

$$\text{(Primal Problem)} \qquad \max\left\{ (\mathcal{O}_n \ \mathrm{e}_m) \begin{pmatrix} \mathrm{h} \\ \mathrm{x} \end{pmatrix} : \begin{pmatrix} \Pi & -\mathrm{I} \\ -\Pi & \mathrm{I} \\ \mathcal{O}_{m,n} & -\mathrm{I} \end{pmatrix} \begin{pmatrix} \mathrm{h} \\ \mathrm{x} \end{pmatrix} \leq 0 \right\}.$$

The corresponding dual problem can therefore be written as

$$\text{(Dual Problem)} \qquad \min\left\{ \begin{pmatrix} \mathcal{O}_m \\ \mathcal{O}_m \\ \mathcal{O}_m \end{pmatrix}^\top \begin{pmatrix} \mathrm{y}^1 \\ \mathrm{y}^2 \\ \mathrm{y}^3 \end{pmatrix} : \begin{pmatrix} \Pi^\top & -\Pi^\top & \mathcal{O}_m \\ -\mathrm{I}_m & \mathrm{I}_m & -\mathrm{I}_m \end{pmatrix} \begin{pmatrix} \mathrm{y}^1 \\ \mathrm{y}^2 \\ \mathrm{y}^3 \end{pmatrix} = \begin{pmatrix} \mathcal{O}_n \\ \mathrm{e}_n \end{pmatrix}^\top, \mathrm{y}^1, \mathrm{y}^2, \mathrm{y}^3 \geq \mathcal{O}_m \right\}.$$

The objective function of the dual is equal to zero. Define $\mathrm{y} := \mathrm{y}^2 - \mathrm{y}^1$ and $\mathrm{z} := \mathrm{y}^3$, then the dual problem simplifies to

$$\text{(Dual Problem)} \qquad \min\left\{0 : \Pi^\top \mathrm{y} = \mathcal{O}_n, \mathrm{y} = \mathrm{z} + \mathrm{e}_m, \mathrm{z} \geq \mathcal{O}_m\right\}.$$

Clearly, there is a feasible solution if and only if there exists[3] $\mathrm{y} \in \mathrm{Ker}(\Pi^\top) = \mathrm{Col}(\Pi)^\perp$ such that $\mathrm{y} \geq \mathrm{e}_m$. This is then equivalent to the existence of a vector $\mathrm{y} \in \mathrm{Ker}(\Pi^\top)$ such that $\mathrm{y} > \mathcal{O}_m$ and (by scaling) $\sum_{i=1}^m y_i = 1$, and the theorem follows. □

[3] Recall that the column space Col(Π) is the set of all possible linear combinations of its column vectors

The fundamental theorem of asset pricing ensures the existence of a risk-neutral probability in a no-arbitrage framework. This in turn implies that (European) financial derivatives prices can be computed by expectation of their final payoffs. We are now interested in the following problem:

$$\text{(Primal Problem)'} \qquad \begin{aligned} &\max \varepsilon \\ &\text{subject to } \mathrm{x} = \Pi\, \mathrm{h} \text{ and } \mathrm{x} \geq \varepsilon \mathrm{e}, \end{aligned} \tag{6.3.2}$$

where inequalities are considered component-wise. Here, the zero vector is a trivial feasible solution.

Definition 6.3.4. A dominant strategy exists if the optimal value of (6.3.2) is strictly positive.

We shall say that a pricing measure is linear if some of its components can be equal to zero.

Theorem 6.3.5. *There is no dominant trading strategy if and only if there exists a linear pricing measure.*

Proof. The Primal problem (6.3.2) can be rewritten as

$$\text{(Primal Problem)'} \qquad \max \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}^\top \begin{pmatrix} \mathrm{h} \\ \mathrm{x} \\ \varepsilon \end{pmatrix} : \begin{pmatrix} \Pi & -\mathrm{I} & 0 \\ -\Pi & \mathrm{I} & 0 \\ \mathcal{O}_{m,n} & -\mathrm{I} & \mathrm{e} \end{pmatrix} \begin{pmatrix} \mathrm{h} \\ \mathrm{x} \end{pmatrix} \leq 0 \right\},$$

and the corresponding dual:

$$\text{(Dual Problem)'} \qquad \min \left\{ 0 : \Pi^\top (\mathrm{y}^1 - \mathrm{y}^2) = 0, -\mathrm{y}^1 + \mathrm{y}^2 - \mathrm{y}^3 = 0, \mathrm{e}^\top \mathrm{y}^3 = 1, \mathrm{y}^1, \mathrm{y}^2, \mathrm{y}^3 \geq 0 \right\}.$$

Let $\mathrm{y} := \mathrm{y}^2 - \mathrm{y}^1$ and $z := \mathrm{y}^3$, so that the dual reads

$$\text{(Dual Problem)'} \qquad \min \left\{ 0 : \Pi^\top \mathrm{z} = 0, \sum_i z_i = 1, \mathrm{z} \geq \mathcal{O}_m \right\},$$

and clearly a solution to this dual problem is a linear pricing measure. □

6.4 Application to arbitrage detection

Assume that n derivatives written on some underlying stock price S are given, all with the same maturity, but with piecewise linear payoffs $\psi_i(\cdot)$ with a single breakpoint K_i. The standard examples are Calls, Puts, Straddles, and assume that the strikes are ordered: $0 < K_1 \leq \cdots < K_n$. We shall introduce a fictitious strike $K_0 = 0$ to simplify some of the computations below. Consider also a portfolio $\mathrm{x} = (x_1, \ldots, x_n)^\top$, where each component represents the quantity of each derivative, so that the payoff of the portfolio at maturity is

$$\psi^{\mathrm{x}}(\cdot) = \sum_{i=1}^n x_i \psi_i(\cdot) = \mathrm{x}^\top \psi(\cdot).$$

At valuation time, say inception of the contract for example, the price of the portfolio reads $\Pi^{\mathrm{x}} = \sum x_i \Pi_i = \mathrm{x}^\top \Pi$, where each Π_i denotes the present value of derivative i.

Definition 6.4.1.

- The portfolio x is called an arbitrage opportunity of type A if

$$\Pi^{\mathrm{x}} = 0 \text{ and there exists } s \geq 0 \text{ such that } \psi^{\mathrm{x}}(s) > 0.$$

- The portfolio x is called an arbitrage opportunity of type B if

$$\Pi^{\mathrm{x}} < 0 \text{ and there exists } s \geq 0 \text{ such that } \psi^{\mathrm{x}}(s) \geq 0.$$

Note that since each derivative payoff is piecewise linear with a single breakpoint, the payoff of the whole portfolio is non-negative if and only if if it so at each breakpoint, and increasing after the last one. Mathematically, we can state this in the following way:

Proposition 6.4.2. *The function ψ^{x} is non-negative if and only if*

$$\psi^{\mathrm{x}}(K_i) \geq 0 \textit{ for all } i = 0, \ldots, n, \qquad \textit{and} \qquad \psi^{\mathrm{x}}(1 + K_n) \geq \psi^{\mathrm{x}}(K_n).$$

We can therefore formulate the primal problem as follows:

$$\text{(Primal Problem)} \qquad \begin{array}{l} \inf \mathrm{x}^\top \Pi \\ \text{subject to } \begin{cases} \mathrm{x}^\top \psi(K_j) \geq 0, \text{ for } j = 0, \ldots, n, \\ \mathrm{x}^\top \Big(\psi(1 + K_n) - \psi(K_n) \Big) \geq 0. \end{cases} \end{array} \tag{6.4.1}$$

or

$$\text{(Primal Problem)} \qquad \begin{array}{l} \inf_{\mathrm{x}} \mathrm{c}^\top \mathrm{x}, \\ \text{subject to } \mathrm{L}\mathrm{x} \geq 0, \end{array}$$

which admits, as a dual problem:

$$\text{(Dual Problem)} \qquad \begin{array}{l} \sup_{\mathrm{y}} 0, \\ \text{subject to } \mathrm{L}^\top \mathrm{y} = \mathrm{c} \quad \text{and} \quad \mathrm{y} \geq 0 \end{array}$$

The following theorem is the main result of this section:

Theorem 6.4.3. *There is no arbitrage if and only if the dual problem admits a strictly positive solution.*

Proof. The theorem follows directly from the following two claims:

(i) The dual problem is feasible if and only if there is no Type-B arbitrage;

(ii) Assume that there is no Type-B arbitrage; then there is no Type-A arbitrage if and only if the dual problem admits a strictly positive solution.

Claim (i) is simple to prove: if there is no Type-B arbitrage, then clearly the primal problem is bounded (since its optimal value is null), and therefore, by duality (Theorem 6.2.1), the dual problem is feasible. Conversely, assuming that the dual is feasible, its optimal value is therefore null, so is that of the primal by duality, and Type-B arbitrage cannot occur. Claim (ii) is slightly more subtle to prove. Assume absence of Type-B arbitrage, and consider the 'only if' part of the claim. Let h be a portfolio of derivatives, with payoff h at maturity and initial cost $\mathrm{c}^\top \mathrm{h} = \mathrm{y}^\top \mathrm{Lh}$, where y solves the dual problem. If $\mathrm{y} > 0$, then if $\mathrm{Lh} \geq 0$ (but not equal to zero), then $\mathrm{y}^\top \mathrm{Lh} > 0$, which rules out Type-A arbitrage. Consider now the 'if' part of the claim, and suppose that any y solving the dual problem has at least one null entry, say $y_i = 0$. Denote by $\mathrm{e} = (e_1, \ldots, e_m)$ the vector such that $e_j = 0$ whenever $j \neq i$ and $e_i = 1$. Therefore, the primal problem

$$\sup_{\mathrm{y}} \mathrm{e}^\top \mathrm{y}, \text{ subject to } \mathrm{L}^\top \mathrm{y} = c, \mathrm{y} \geq 0$$

has an optimal value equal to zero, and so does its dual

$$\inf_{\mathrm{x}} \mathrm{c}^\top \mathrm{x}, \text{ subject to } \mathrm{Lx} \geq \mathrm{e},$$

which precisely gives the minimum cost for a portfolio with payoff greater than e, and therefore we have constructed an Type-A arbitrage portfolio, and the claim follows by contraposition. □

Application to Calls and Puts

We now apply the framework developed above to the case of European Call options. For a fixed maturity, we shall show that no arbitrage (in the sense of Definition 6.4.1) is equivalent to the convexity of the Call option prices. Assume that we can observe the Call options $\psi_i(K_j) = (K_j - K_i)_+$. The primal problem therefore reads

$$\text{(Primal Problem)} \qquad \inf_{\mathrm{x}} \mathrm{c}^\top \mathrm{x}, \quad \text{such that} \quad \mathrm{Lx} \geq 0,$$

where $\mathrm{c} = (c(K_1), \ldots, c(K_N))^\top$, and the matrix L reads

$$\mathrm{L} = \begin{pmatrix} K_1 - K_1 & 0 & 0 & \cdots & 0 \\ K_3 - K_1 & K_3 - K_2 & 0\cdots & 0 & \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & \cdots & \cdots & 1 \end{pmatrix} \in \mathcal{M}_{NN}(\mathbb{R}).$$

Theorem 6.4.4. *There is no arbitrage if and only if the (discrete) map $K \mapsto c(K)$ is decreasing, convex and strictly positive.*

Proof. From Theorem 6.4.3, there is no arbitrage if and only if the dual problem admits a strictly positive solution. The dual constraints read

$$\begin{cases} \sum_{j=1}^{N-i} (K_{i+j} - K_i)\, y_{i+j-1} + y_N \quad c(K_i), & i = 1, \ldots, N-1, \\ \qquad\qquad\qquad\qquad y_N \quad = c(K_N). & \end{cases} \tag{6.4.2}$$

Subtracting the $(i+1)$th equation from the ith one and dividing by $k_{i+1} - K_i > 0$, we obtain

$$y_i + y_{i+1} + \cdots + y_{N-1} = \frac{c(K_i) - c(K_{i+1})}{K_{i+1} - K_i}, \quad \text{for all } i = 1, \dots, N-2,$$

which yields, by recursion,

$$y_i + \frac{c(K_{i+1}) - c(K_{i+2})}{K_{i+2} - K_i} = \frac{c(K_i) - c(K_{i+1})}{K_{i+1} - K_i}, \quad \text{for all } i = 2, \dots, N-2. \tag{6.4.3}$$

We first prove the necessity part of the claim. Assume that there is no arbitrage opportunity (the dual admits a strictly positive solution). Since $y_N > 0$, from (6.4.2), $c(K_N) > 0$ and $c(K_{N-1}) > c(K_N)$. Using (6.4.3), we obtain, for any $i = 2, \dots, N-2$, that

$$\frac{c(K_{i+1}) - c(K_{i+2})}{K_{i+2} - K_i} = \frac{c(K_i) - c(K_{i+1})}{K_{i+1} - K_i},$$

which proves convexity. To prove sufficiency, we apply the exact same recursion, assuming that convexity and the strict decreasing property holds which, from the equations above, yields that $y_i > 0$, for all $i = 1, \dots, N-2$, so that the dual problem admits a strictly positive solution, and arbitrage opportunities can hence not arise. □

Appendix A

Useful tools in probability theory and PDE

A.1 Essentials of probability theory

We provide here a brief overview of standard results in probability theory and convergence of random variables needed in these lecture notes. The reader is invited to consult [56] for instance for a more thorough treatment of the subject.

A.1.1 PDF, CDF and characteristic functions

In the following, $(\Omega, \mathcal{F}, \mathbb{P})$ shall denote a probability space and X a random variable defined on it. We define the cumulative distribution function $F : \mathbb{R} \to [0, 1]$ of S by

$$F(x) := \mathbb{P}\left(X \leq x\right), \qquad \text{for all } x \in \mathbb{R}.$$

The function F is increasing and right-continuous and satisfies the identities $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. If the function F is absolutely continuous, then the random variable X has a probability density function $f : \mathbb{R} \to \mathbb{R}_+$ defined by $f(x) = F'(x)$, for all real number x. Note that this in particular implies the equality $F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u$. Recall that a function $F : \mathcal{D} \subset \mathbb{R} \to \mathbb{R}$ is said to be *absolutely continuous* if for any $\varepsilon > 0$, there exists $\delta > 0$ such that the implication

$$\sum_n |b_n - a_n| < \delta \qquad \Longrightarrow \qquad \sum_n |F(b_n) - F(a_n)| < \delta$$

holds for any sequence of pairwise disjoint intervals $(a_n, b_n) \subset \mathcal{D}$. Define now the characteristic function $\phi : \mathbb{R} \to \mathbb{C}$ of the random variable X by

$$\phi(u) := \mathbb{E}\left(\mathrm{e}^{\mathrm{i}uX}\right).$$

Note that it is well defined for all real number u and that we always have $|\phi(u)| \leq 1$. Extending it to the complex plane ($u \in \mathbb{C}$) is more subtle and shall be dealt with in Chapter 4, along with some properties of characteristic functions.

A.1.2 Gaussian distribution

A random variable X is said to have a Gaussian distribution (or Normal distribution) with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, and we write $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ if and only if its density reads

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^2\right), \qquad \text{for all } x \in \mathbb{R}.$$

For such a random variable, the following identities are obvious:

$$\mathbb{E}\left(\mathrm{e}^{\mathtt{i}uX}\right) = \exp\left(\mathtt{i}\mu u - \frac{1}{2}u^2\sigma^2\right), \qquad \text{and} \qquad \mathbb{E}\left(\mathrm{e}^{uX}\right) = \exp\left(\mu u + \frac{1}{2}u^2\sigma^2\right),$$

for all $u \in \mathbb{R}$. The first quantity is the characteristic function whereas the second one is the Laplace transform or the random variable. If $X \in \mathcal{N}\left(\mu, \sigma^2\right)$, then the random variable $Y := \exp(X)$ is said to be lognormal and

$$\mathbb{E}(Y) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \qquad \text{and} \qquad \mathbb{E}\left(Y^2\right) = \exp\left(2\mu + 2\sigma^2\right).$$

A.1.3 Convergence of random variables

We recall here the different types of convergence for family of random variables $(X_n)_{n\geq 1}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We shall denote $F_n : \mathbb{R} \to [0,1]$ the corresponding cumulative distribution functions and $f_n : \mathbb{R} \to \mathbb{R}_+$ their densities whenever they exist. We start with a definition of convergence for functions, which we shall use repeatedly.

Definition A.1.1. Let $(h_n)_{n\geq 1}$ be a family of functions from $\mathbb{R}$ to $\mathbb{R}$. We say that the family converges pointwise to a function $h : \mathbb{R} \to \mathbb{R}$ if and only if the equality $\lim_{n\to\infty} h_n(x) = h(x)$ holds for all real number x.

Convergence in distribution

This is the weakest form of convergence, and is the one appearing in the central limit theorem.

Definition A.1.2. The family $(X_n)_{n\geq 1}$ converges in distribution—or weakly or in law—to a random variable X if and only if the family $(F_n)_{n\geq 1}$ converges pointwise to a function $F : \mathbb{R} \to [0,1]$, i.e. the equality

$$\lim_{n\to\infty} F_n(x) = F(x),$$

holds for all real number x where F is continuous. Furthermore, the function F is the CDF of the random variable X.

Example. Consider the family $(X_n)_{n\geq 1}$ such that each X_n is uniformly distributed on the interval $\left[0, n^{-1}\right]$. We then have $F_n(x) = nx\mathbb{1}_{\{x\in[0,1/n]\}} + \mathbb{1}_{\{x\geq 1/n\}}$. It is clear that the family of random variable converges weakly to the degenerate random variable $X = 0$. However, for any $n \geq 1$, we have $F_n(0) = 0$ and $F(0) = 1$. The function F is not continuous at 1, but the definition still holds.

Example. Weak convergence does not imply convergence of the densities, even when they exist. Consider the family such that $f_n(x) = \Big(1 - \cos(2\pi n x)\Big)\mathbb{1}_{\{x\in(0,1)\}}$.

Even though convergence in law is a weak form of convergence, it has a number of fundamental consequences for applications. We list them here without proof and refer the interested reader to [6] for details

Corollary A.1.3. *Assume that the family $(X_n)_{n\geq 1}$ converges weakly to the random variable X. Then the following statements hold*

1. $\lim_{n\to\infty} \mathbb{E}(h(X_n)) = \mathbb{E}(h(X))$ *for all bounded and continuous function h.*
2. $\lim_{n\to\infty} \mathbb{E}(h(X_n)) = \mathbb{E}(h(X))$ *for all Lipschitz function h.*
3. $\lim \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$ *for all continuity sets A of X.*
4. *(Continous mapping theorem). The sequence $(h(X_n))_{n\geq 1}$ converges in law to $h(X)$ for every continuous function h.*

The following theorem shall be of fundamental importance in many applications, and we therefore state it separately.

Theorem A.1.4 (Lévy's continuity theorem)**.** *The family $(X_n)_{n\geq 1}$ converges weakly to the random variable X if and only if the sequence of characteristic functions ϕ_n converges pointwise to the characteristic function ϕ of X and ϕ is is continuous at the origin.*

Convergence in probability

Definition A.1.5. The family $(X_n)_{n\geq 1}$ converges in probability to the random variable X if, for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$

Remark A.1.6. The continuous mapping theorem still holds under this form of convergence.

Almost sure convergence

This form of convergence is the strongest form of convergence and can be seen as an analogue for random variables of the pointwise convergence for functions.

Definition A.1.7. The family $(X_n)_{n\geq 1}$ converges almost surely to the random variable X if

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1.$$

Convergence in mean

Definition A.1.8. Let $r \in \mathbb{N}^*$. The family $(X_n)_{n\geq 1}$ converges in the L^r norm to the random variable X if the r-th absolute moments of X_n and X exist for all $n \geq 1$ and if

$$\lim_{n\to\infty} \mathbb{E}\left(|X_n - X|^r\right) = 0.$$

Properties

- Almost sure convergence implies convergence in probability.
- Convergence in probability implies weak convergence.
- Convergence in the L^r norm implies convergence in probability.
- For any $r \geq s \geq 1$, convergence in the L^r norm implies convergence in the L^s norm.

A.1.4 Central limit theorem and Berry-Esséen inequality

Let $(X_i)_{i=1\ldots,n}$ form a sequence of independent and identically distributed random variables with finite mean μ and finite variance $\sigma^2 > 0$, and define the sequences of random variables $(\overline{X}_n)_{n\geq 1}$ and $(Z_n)_{n\geq 1}$ by

$$\overline{X}_n := \sum_{i=1}^{n} X_i \qquad \text{and} \qquad Z_n := \frac{\overline{X}_n - n\mu}{\sigma\sqrt{n}}, \qquad \text{for each } n \geq 1. \tag{A.1.1}$$

Recall now the central limit theorem:

Theorem A.1.9 (Central limit theorem)**.** *The family $(Z_n)_{n\geq 1}$ converges in distribution to a Gaussian distribution with zero mean and unit variance. In particular for any $a < b$, we have $\lim_{n\to\infty} \mathbb{P}(Z_n \in [a,b]) = \mathcal{N}(b) - \mathcal{N}(a)$.*

The central limit theorem provides information about the limiting behaviour of the probabilities, but does not tell anything aboug the rate of convergence or the error made when approximating the Gaussian distribution by the distribution of Z_n for $n \geq 1$ fixed. The following theorem, proved by Berry [5] and Esséen [24] gives such estimates

Theorem A.1.10. *Assume that $\mathbb{E}\left(|X|^3\right) < \infty$. Then there exists a strictly positive universal (i.e. independent of n) constant C such that*

$$\sup_x \left|\mathbb{P}(Z_n \leq x) - \mathcal{N}(x)\right| \leq \frac{C\rho}{\sqrt{n}},$$

where $\rho := \mathbb{E}\left(\frac{|X_1 - \mu|^3}{\sigma^3}\right)$.

A.2 Useful tools in linear algebra

Let $n \in \mathbb{N}$ and consider a matrix $A = (a_{ij})_{1\le i,j\le n} \in \mathcal{M}_n(\mathbb{R})$.

Definition A.2.1. The matrix A is said to be *positive definite* (respectively *positive semi-definite*) if $\mathrm{x}^t \mathrm{A}\mathrm{x} > 0$ (resp ≥ 0) for all non null vector $\mathrm{x} \in \mathbb{R}^n$.

For a matrix $\mathrm{A} \in \mathcal{M}_n(\mathbb{R})$, we define its *principal minors* as

$$\Delta_1 := a_{11}, \qquad \Delta_2 := \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad \ldots, \qquad \Delta_n := \det(\mathrm{A}).$$

Proposition A.2.2. *The following statements are equivalent:*

(i) A *is positive definite;*

(ii) all the eigenvalues of A *are positive;*

(iii) all leading principal minors of A *are positive.*

Exercise 54. Let $S \subset \mathbb{R}^n$ be a convex and open set. Let f be a continuously differentiable function on S. Recall that the function f is convex in S if for any two points x and y in S, the following inequality holds:

$$f\left(\alpha \mathrm{x} + (1-\alpha)\mathrm{y}\right) \leq \alpha f(\mathrm{x}) + (1-\alpha) f(\mathrm{y}), \quad \text{for any } \alpha \in [0,1].$$

Show the following:

(i) f is convex if and only if $f(\mathrm{y}) \geq f(x) + \nabla f(\mathrm{x})^T \cdot (\mathrm{y} - \mathrm{x})$ for all $(\mathrm{x},\mathrm{y}) \in S \times S$;

(ii) if f is twice continuously differentiable on S, then f is convex if and only if the matrix $\nabla^2 f(\mathrm{x})$ is positive semi-definite for all $\mathrm{x} \in S$.

Definition A.2.3. The *spectral radius* ρ of a matrix $\mathrm{A} \in \mathcal{M}_n(\mathbb{R})$ is defined by $\rho(\mathrm{A}) := \max_{1\le i\le n} \lambda_i$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

A.3 Useful tools in analysis

Theorem A.3.1 (Fubini theorem on $\mathbb{R}$ or $\mathbb{C}$)**.** *Let A and B be two subsets of $\mathbb{C}$ and $f : A \times B \to \mathbb{C}$ a function. If $\int_{A\times B} |f(x,y)|\,\mathrm{d}(x,y) < \infty$ then*

$$\int_A \int_B f(x,y)\mathrm{d}x\mathrm{d}y = \int_B \int_A f(x,y)\mathrm{d}y\mathrm{d}x = \int_{A\times B} f(x,y)\mathrm{d}(x,y).$$

Bibliography

[1] M. Abramowitz and I. Stegun. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover Publications, 1972.

[2] L.B.G. Andersen, P. Jäckel and C. Kahl. Simulation of Square-Root Processes. Encyclopedia of Quantitative Finance, 2010.

[3] K.E. Atkinson. An introduction to numerical analysis, Second Edition. Wiley, 1989.

[4] D. H. Bailey and P. N. Swarztrauber. The fractional Fourier transform and applications. *SIAM Review*,33: 389-404, 1991.

[5] A.C. Berry. The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Transactions of the American Mathematical Society*, 49 (1): 122-136, 1941.

[6] P. Billingsley. Convergence of probability measures (2nd ed.). John Wiley & Sons, 1999.

[7] F. Black and M. Scholes. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81 (3): 637-654, 1973.

[8] P. Boyle. Option Valuation Using a Three-Jump Process. *International Options Journal* 3, 7-12, 1986.

[9] P Carr and D. Madan. Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2 (4): 61-73, 1999.

[10] P Carr and D. Madan. Saddlepoint Methods for Option Pricing. *Journal of Computational Finance*, 13 (1): 4961, 2009.

[11] A.L. Cauchy. Cours d'analyse de l'Ecole Royale Polytechnique. Imprimerie royale, 1821. Reissued by Cambridge University Press, 2009.

[12] J.W. Cooley and J.W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19: 297-301, 1965.

[13] R. Courant, K. Friedrichs and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100 (1): 32-74, 1928.

[14] J.C. Cox, J.E. Ingersoll and S.A. Ross. A Theory of the Term Structure of Interest Rates. *Econometrica*, 53: 385-407.

[15] J.C. Cox, S.A. Ross and M. Rubinstein. Option Pricing: A Simplified Approach. *Journal of Financial Economics*, 7: 229-263, 1979.

[16] J.Crank and P. Nicolson. A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of Heat Conduction Type. *Proceedings of the Cambridge Philosophical Society* 43: 50-67, 1947.

[17] F. Delbaen and W. Schachermayer. A general version of the fundamental theorem of asset pricing. *Math Ann*, 300(1): 463-520, 1994.

[18] F. Delbaen and W. Schachermayer. The Mathematics of arbitrage. Springer, 2008.

[19] J. Douglas and H.H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the AMS*, 82:421-439, 1956.

[20] D. Duffie, D. Filipovic, and W. Schachermayer. Affine processes and applications in finance. *The Annals of Applied Probability*, 13(3): 984-1053, 2003.

[21] D. Duffie, J. Pan, and K. Singleton. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6): 1343-1376, 2000.

[22] D.J. Duffy. Finite Difference Methods in Financial Engineering. Wiley, Chichester, 2006.

[23] D.J. Duffy. A Critique of the Crank-Nicolson scheme strenghts and weaknesses for financial instrument pricing. *Wilmott Magazine*, July-August, 2004

[24] C.G. Esséen. A moment inequality with an application to the central limit theorem. *Skand. Aktuarietidskr.*, 39: 160170, 1956.

[25] F. Fang and K. Osterlee. A novel pricing method for European options based on Fourier-cosine series expansions. *SIAM Journal of Scientific Computing*, 31: 826-848, 2008.

[26] W. Feller. Two Singular Diusion Problems. *Annals of Mathematics* 54 (1), 1951.

[27] W. Gander and W. Gautschi. Adaptive Quadrature Revisited. BIT, 40: 84-101, 2000.

[28] I. M.Gelfand. Normierte Ringe. *Mat. Sbornik*, 9: 3-24, 1941.

[29] J. Gil-Pelaez. Note on the inversion theorem. *Biometrika*, 38 (3-4): 481-482, 1951.

[30] P. Glasserman. Monte Carlo methods in financial engineering, Springer-Verlag, 2003.

[31] L. Grafakos. Classical Fourier Analysis. Springer, 249 (2nd edition), 2008.

[32] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13, 49-52, 1902.

[33] J.M. Harrison and D. M. Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20: 381-408, 1979.

[34] J.M. Harrison and S. R. Pliska. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications*, 11: 215-260, 1981.

[35] J.M. Harrison and S. R. Pliska. A stochastic calculus model of continuous trading: Complete markets. *Stochastic Processes and their Applications*, 15: 313-316, 1983.

[36] S.L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6: 237-343, 1993.

[37] A. Hadjidimos. Successive Overrelaxation (SOR) and related methods. *Journal of Computational and Applied Mathematics*, 123: 177-199, 2000.

[38] J.C. Hull and A. White. Numerical procedures for implementing term structure models I: Single-factor models. *Journal of Derivatives*, 2 (1): 7-16, 1994.

[39] E. Isaacson and H. Keller. Analysis of numerical methods. Wiley, New-York, 1966.

[40] S. Karlin and H. Taylor. A Second course in Stochastic processes. Academic Press, 1981.

[41] B. Kamrad and P. Ritchken. Multinomial Approximating Models for Options with k State Variables. *Management Science*, 37 (12): 1640-1652, 1991.

[42] M. Keller-Ressel. Moment Explosions and Long-Term Behavior of Affine Stochastic Volatility Models. *Mathematical Finance*, 21(1): 73-98, 2011.

[43] P. Kloeden and E. Platen. Numerical solutions of stochastic differential equations. Springer-Verlag, New-York, 1999.

[44] D.E. Knuth. The Art of computer programming, volume II: Seminumerical algorithms. Third Edition, Addison Wesley Longman, Reading, Mass, 1998.

[45] H. J. Kushner and P. Dupuis. Numerical Methods for Stochastic Control Problems in Continuous Time. Springer, 2001.

[46] R. W. Lee. Option Pricing by Transform Methods: Extensions, Unification, and Error Control. *Journal of Computational Finance*, 7 (3): 51-86, 2004.

[47] D.G. Luenberger, Y. Ye. Linear and nonlinear programming. Springer, 2010.

[48] R. Lugannani and S.O. Rice. Saddlepoint approximations for the distribution of the sum of independent random variables. *Advances in Applied Probability*, 12: 475-490, 1980.

[49] G. Marsaglia. Random numbers fall mainly in the planes. *Proceedings of the National Academy of Sciences*, 61:25-28, 1968.

[50] R. Merton. The Theory of Rational Option Pricing. *Bell Journal of Economics and Management Science*, 4(1): 141-183, 1973.

[51] D.W. Peaceman and H.H. Rachford. The numerical solution of parabolic and elliptic differential equations. *Journal of SIAM*, 3: 28-41, 1955.

[52] B. Flannery, W.H. Press, S. Teukolsky and W. Vetterling. Numerical Recipes. Cambridge University Press, Third Edition, 2007.

[53] J.C. Strikwerda. Finite difference schemes and partial differential equations, Second Edition. Chapman and Hall, Pacific Grove, 1989.

[54] O. Vasicek. An equilibrium characterization of the term structure *Journal of Financial Economics*, 5: 177-188, 1977.

[55] K. Weierstrass. Zur Theorie der Potenzreihen. *Werke*, 1:67-74, 1894.

[56] D. Williams. Probability with martingales. Cambridge University Press, 1991.

[57] A.T.A Wood, J.G. Booth and R.W. Butler. Saddlepoint approximations with nonnormal limit distributions. *Journal of the American Statistical Association*, 88: 680-686, 1993.

[58] Zeliade Systems. Heston 2010. www.zeliade.com/whitepapers/zwp-0004.pdf, 2011.

[59] G. Zoutendijk. Methods of feasible directions. Elsevier, Amsterdam, 1960

[60] K. R. Zvan, P.A. Forsyth and K. Vetzal. Robust Numerical Methods for PDE Models of Asian Options. *Journal of Computational Finance*, 1:39-78, 1998.