Chapter 2 Generating Random Numbers with Specified Distributions

Simulation and valuation of finance instruments require numbers with specified distributions. For example, in Section 1.6 we have used numbers Z drawn from a standard normal distribution, $Z \sim \mathcal{N}(0,1)$. If possible the numbers should be random. But the generation of "random numbers" by digital computers, after all, is done in a deterministic and entirely predictable way. If this point is to be stressed, one uses the term *pseudo-random*[1].

Computer-generated random numbers mimic the properties of true random numbers as much as possible. This is discussed for uniformly distributed numbers in Section 2.1. Suitable transformations generate normally distributed numbers (Sections 2.2, 2.3). Section 2.3 includes the vector case, where normally distributed numbers are calculated with prescribed correlation.

Another approach is to dispense with randomness and to generate *quasi-random numbers*, which aim at avoiding one disadvantage of random numbers, namely, the potential lack of equidistributedness. The resulting *low-discrepancy* numbers will be discussed in Section 2.5. These numbers are used for the deterministic Monte Carlo integration (Section 2.4).

Definition 2.1 (sample from a distribution)

A sequence of numbers is called a *sample from* F if the numbers are independent realizations of a random variable with distribution function F.

If F is the uniform distribution over the interval $[0,1)$ or $[0,1]$, then we call the samples from F *uniform deviates (variates)*, notation $\sim \mathcal{U}[0,1]$. If F is the standard normal distribution then we call the samples from F *standard normal deviates (variates)*; as notation we use $\sim \mathcal{N}(0,1)$. The basis of the random-number generation is to draw uniform deviates.

[1] Since in our context the predictable origin is clear we omit the modifier "pseudo," and hereafter use the term "random number." Similarly we talk about randomness of these numbers when we mean apparent randomness.

2.1 Uniform Deviates

A standard approach to calculate uniform deviates is provided by linear congruential generators.

2.1.1 Linear Congruential Generators

Choose integers M, a, b, with $a, b < M$, $a \neq 0$. For $N_0 \in \mathbb{N}$ a sequence of integers N_i is defined by

Algorithm 2.2 (linear congruential generator)

$$\boxed{\begin{array}{l} \text{Choose } N_0. \\ \text{For } i = 1, 2, \ldots \text{ calculate} \\ N_i = (aN_{i-1} + b) \bmod M \end{array}} \qquad (2.1)$$

The *modulo* congruence $N = Y \bmod M$ between two numbers N and Y is an equivalence relation [Gen98]. The number N_0 is called the *seed*. Numbers $U_i \in [0, 1)$ are defined by

$$U_i = N_i/M\,, \qquad (2.2)$$

and will be taken as uniform deviates. Whether the numbers U_i are suitable will depend on the choice of M, a, b and will be discussed next.

Properties 2.3 (periodicity)

(a) $N_i \in \{0, 1, ..., M - 1\}$

(b) The N_i are periodic with period $\leq M$.
(Because there are not $M + 1$ different N_i. So two in $\{N_0, ..., N_M\}$ must be equal, $N_i = N_{i+p}$ with $p \leq M$.)

Obviously, some peculiarities must be excluded. For example, $N = 0$ must be ruled out in case $b = 0$, because otherwise $N_i = 0$ would repeat. In case $a = 1$ the generator settles down to $N_n = (N_0 + nb) \bmod M$. This sequence is too easily predictable. Various other properties and requirements are discussed in the literature, in particular in [Knu95]. In case the period is M, the numbers U_i are distributed "evenly" when exactly M numbers are needed. Then each grid point on a mesh on [0,1] with mesh size $\frac{1}{M}$ is occupied once.

After these observations we start searching for good choices of M, a, b. There are numerous possible choices with bad properties. For serious computations we recommend to rely on suggestions of the literature. [PrTVF92] presents a table of "quick and dirty" generators, for example, $M = 244944$, $a = 1597, b = 51749$. Criteria are needed to decide which of the many possible generators are recommendable.

2.1.2 Quality of Generators

What are good random numbers? A practical answer is the requirement that the numbers should meet “all” aims, or rather pass as many tests as possible. The requirements on good number generators can roughly be divided into three groups.

The first requirement is that of a large period. In view of Property 2.3 the number M must be as large as possible, because a small set of numbers makes the outcome easier to predict —a contrast to randomness. This leads to select M close to the largest integer machine number. But a period p close to M is only achieved if a and b are chosen properly. Criteria for relations among M, p, a, b have been derived by number-theoretic arguments. This is outlined in [Rip87], [Knu95]. For 32-bit computers, a common choice has been $M = 2^{31} - 1$, $a = 16807$, $b = 0$.

A second group of requirements are the *statistical tests* that check whether the numbers are distributed as intended. The simplest of such tests evaluates the sample mean $\hat{\mu}$ and the sample variance $\hat{s}^2$ (B1.11) of the calculated random variates, and compares to the desired values of μ and σ^2. (Recall $\mu = 1/2$ and $\sigma^2 = 1/12$ for the uniform distribution.) Another simple test is to check correlations. For example, it would not be desirable if small numbers are likely to be followed by small numbers.

A slightly more involved test checks how well the probability distribution is approximated. This works for general distributions ($\longrightarrow$ Exercise 2.14). Here we briefly summarize an approach for uniform deviates. Calculate j samples from a random number generator, and investigate how the samples distribute on the unit interval. To this end, divide the unit interval into subintervals of equal length ΔU, and denote by j_k the number of samples that fall into the kth subinterval

$$k\Delta U \leq U < (k+1)\Delta U \,.$$

Then j_k/j should be close the desired probability, which for this setup is ΔU. Hence a plot of the quotients

$$\frac{j_k}{j\Delta U} \quad \text{for all } k$$

against $k\Delta U$ should be a good approximation of 1, the density of the uniform distribution. This procedure is just the simplest test; for more ambitious tests, consult [Knu95].

The third group of tests is to check how well the random numbers distribute in higher-dimensional spaces. This issue of the *lattice structure* is discussed next. We derive a priori analytical results on *where* the random numbers produced by Algorithm 2.2 are distributed.

2.1.3 Random Vectors and Lattice Structure

Random numbers N_i can be arranged in m-tuples $(N_i, N_{i+1}, ..., N_{i+m-1})$ for $i \geq 1$. Then the tuples or the corresponding points $(U_i, ..., U_{i+m-1}) \in [0,1)^m$ are analyzed with respect to correlation and distribution. The sequences defined by the generator of Algorithm 2.2 lie on $(m-1)$-dimensional hyperplanes. This statement is trivial since it holds for the M parallel planes through $U = i/M$, $i = 0, ..., M-1$. But if all points fall on only a small number of parallel hyperplanes (with large empty gaps in between), then the generator would be impractical in many applications. Next we analyze the generator whether such unfavorable planes exist, restricting ourselves to the case $m = 2$.

For $m = 2$ the hyperplanes are straight lines, and are defined by $z_0 N_{i-1} + z_1 N_i = \lambda$, with parameters z_0, z_1, λ. The modulus operation can be written

$$\begin{aligned} N_i &= (aN_{i-1} + b) \bmod M \\ &= aN_{i-1} + b - kM \quad \text{ for } \; kM \leq aN_{i-1} + b < (k+1)M \,, \end{aligned}$$

k an integer, $k = k(i)$. A side calculation for arbitrary z_0, z_1 shows

$$\begin{aligned} z_0 N_{i-1} + z_1 N_i &= z_0 N_{i-1} + z_1 (aN_{i-1} + b - kM) \\ &= N_{i-1}(z_0 + az_1) + z_1 b - z_1 kM \\ &= M \cdot \underbrace{\{N_{i-1} \frac{z_0 + az_1}{M} - z_1 k\}}_{=:c} + z_1 b \,. \end{aligned}$$

We divide by M and obtain the equation of a straight line in the (U_{i-1}, U_i)-plane, namely,

$$z_0 U_{i-1} + z_1 U_i = c + z_1 b M^{-1} \,, \tag{2.3}$$

one line for each parameter c. The points calculated by Algorithm 2.2 lie on these straight lines. To eliminate the seed we take $i > 1$. For each tuple (z_0, z_1), the equation (2.3) defines a family of parallel straight lines, one for each number out of the finite set of c's. The question is whether there exists a tuple (z_0, z_1) such that only few of the straight lines cut the square $[0,1)^2$? In this case wide areas of the square would be free of random points, which violates the requirement of a "uniform" distribution of the points. The minimum number of parallel straight lines (hyperplanes) cutting the square, or equivalently the maximum distance between them serve as measures of the equidistributedness. We now analyze the number of straight lines, searching for the worst case.

For integers (z_0, z_1) satisfying

$$z_0 + az_1 = 0 \bmod M \tag{2.4}$$

the parameter c is integer. By solving (2.3) for $c = z_0 U_{i-1} + z_1 U_i - z_1 b M^{-1}$ and applying $0 \leq U < 1$ we obtain the maximal interval I_c such that for each integer $c \in I_c$ its straight line cuts or touches the square $[0,1)^2$. We count how many such c's exist, and have the information we need. For some constellations of a, M, z_0 and z_1 it may be possible that the points (U_{i-1}, U_i) lie on very few of these straight lines!

Example 2.4 $N_i = 2N_{i-1} \bmod 11$ (that is, $a = 2,\ b = 0,\ M = 11$)

We choose $z_0 = -2,\ z_1 = 1$, which is one tuple satisfying (2.4), and investigate the family (2.3) of straight lines

$$-2U_{i-1} + U_i = c$$

in the (U_{i-1}, U_i)-plane. For $U_i \in [0,1)$ we have $-2 < c < 1$. In view of (2.4) c is integer and so only the two integers $c = -1$ and $c = 0$ remain. The two corresponding straight lines cut the interior of $[0,1)^2$. As Figure 2.1 illustrates, the points generated by the algorithm form a lattice. All points on the lattice lie on these two straight lines. The figure lets us discover also other parallel straight lines such that all points are caught (for other tuples z_0, z_1). The practical question is: What is the largest gap? ($\longrightarrow$ Exercise 2.1)

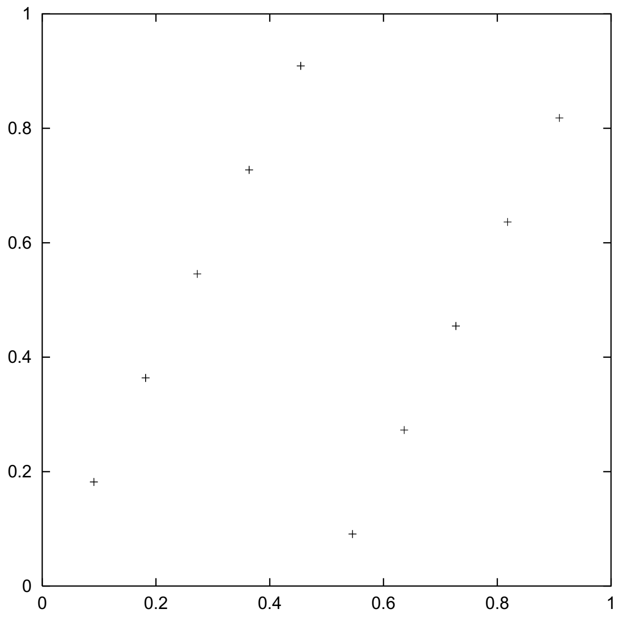

Fig. 2.1. The points (U_{i-1}, U_i) of Example 2.4

Example 2.5 $N_i = (1229 N_{i-1} + 1) \bmod 2048$

The requirement of equation (2.4)

$$\frac{z_0 + 1229 z_1}{2048} \quad \text{integer}$$

is satisfied by $z_0 = -1,\ z_1 = 5$, because

$$-1 + 1229 \cdot 5 = 6144 = 3 \cdot 2048 \,.$$

For c from (2.3) and $U_i \in [0, 1)$ we have

$$-1 - \frac{5}{2048} < c < 5 - \frac{5}{2048} \,.$$

All points (U_{i-1}, U_i) lie on only six straight lines, with $c \in \{-1, 0, 1, 2, 3, 4\}$, see Figure 2.2. On the "lowest" straight line ($c = -1$) there is only one point. The distance between straight lines measured along the vertical U_i–axis is $\frac{1}{z_1} = \frac{1}{5}$.

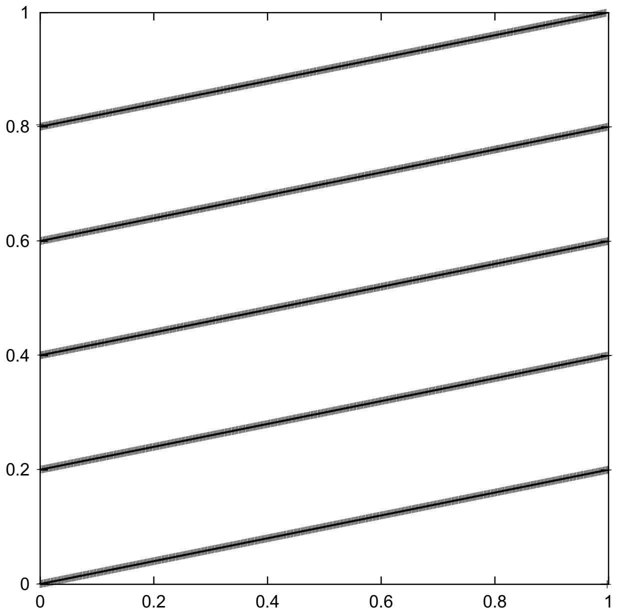

Fig. 2.2. The points (U_{i-1}, U_i) of Example 2.5

Higher-dimensional vectors ($m > 2$) are analyzed analogously. The generator called RANDU

$$N_i = a N_{i-1} \bmod M \,, \quad \text{with } a = 2^{16} + 3,\ M = 2^{31}$$

may serve as example. Its random points in the cube $[0,1)^3$ lie on only 15 planes ($\longrightarrow$ Exercise 2.2). For many applications this must be seen as a severe defect.

In Example 2.4 we asked what the maximum gap between the parallel straight lines is. In other words, we have searched for strips of maximum size in which no point (U_{i-1}, U_i) falls. Alternatively we can directly analyze the lattice formed by consecutive points. For illustration consider again Figure 2.1. We follow the points starting with $(\frac{1}{11}, \frac{2}{11})$. By vectorwise adding an appropriate multiple of $(1, a) = (1, 2)$ the next two points are obtained. Proceeding in this way one has to take care that upon leaving the unit square each component with value ≥ 1 must be reduced to $[0, 1)$ to observe $\bmod M$. The reader may verify this with Example 2.4 and numerate the points of the lattice in Figure 2.1 in the correct sequence. In this way the lattice can be defined. This process of defining the lattice can be generalized to higher dimensions $m > 2$. ($\longrightarrow$ Exercise 2.3)

A disadvantage of the linear congruential generators of Algorithm 2.2 is the boundedness of the period by M and hence by the word length of the computer. The situation can be improved by *shuffling* the random numbers in a random way. For practical purposes, the period gets close enough to infinity. (The reader may test this on Example 2.5.) For practical advice we refer to [PrTVF92].

2.1.4 Fibonacci Generators

The original Fibonacci recursion motivates trying the formula

$$N_{i+1} := (N_i + N_{i-1}) \bmod M \,.$$

It turns out that this first attempt of a three-term recursion is not suitable for generating random numbers ($\longrightarrow$ Exercise 2.15). The modified approach

$$N_{i+1} := (N_{i-\nu} - N_{i-\mu}) \bmod M \tag{2.5}$$

for suitable ν, $\mu \in \mathbb{N}$ is called *lagged Fibonacci* generator. For many choices of ν, μ the approach (2.5) leads to recommendable generators.

Example 2.6

$$\begin{aligned} & U_i := U_{i-17} - U_{i-5}\,, \\ & \text{in case } U_i < 0 \text{ set } U_i := U_i + 1.0 \end{aligned}$$

The recursion of Example 2.6 immediately produces floating-point numbers $U_i \in [0, 1)$. This generator requires a prologue in which 17 initial U's are generated by means of another method. The generator can be run with varying lags ν, μ. [KaMN89] recommends

Algorithm 2.7 (Fibonacci generator)

```
Repeat: $\zeta := U_i - U_j$
        if $\zeta < 0$, set $\zeta := \zeta + 1$
        $U_i := \zeta$
        $i := i - 1$
        $j := j - 1$
        if $i = 0$, set $i := 17$
        if $j = 0$, set $j := 17$
```

Initialization: Set $i = 17,\ j = 5$, and calculate $U_1, ..., U_{17}$ with a congruential generator, for instance with $M = 714025,\ a = 1366,\ b = 150889$. Set the seed $N_0 =$ your favorite dream number, possibly inspired by the system clock of your computer.

Figure 2.3 depicts 10000 random points calculated by means of Algorithm 2.7. Visual inspection suggests that the points are not arranged in some apparent structure. The points appear to be sufficiently random. But the generator provided by Example 2.6 is not sophisticated enough for ambitious applications; its pseudo-random numbers are rather correlated.

A generator of uniform deviates that can be highly recommended is the Mersenne twister [MaN98], it has a truly remarkable long period.

2.2 Extending to Random Variables From Other Distributions

Frequently normal variates are needed. Their generation is based on uniform deviates. The simplest strategy is to calculate

$$X := \sum_{i=1}^{12} U_i - 6, \quad \text{for } \ U_i \sim \mathcal{U}[0,1]\,.$$

X has expectation 0 and variance 1. The Central Limit Theorem ($\longrightarrow$ Appendix B1) assures that X is approximately normally distributed ($\longrightarrow$ Exercise 2.4). But this crude attempt is not satisfying. Better methods calculate non uniformly distributed random variables, for example, by a suitable transformation out of a uniformly distributed random variable [Dev86]. But the most obvious approach inverts the distribution function.

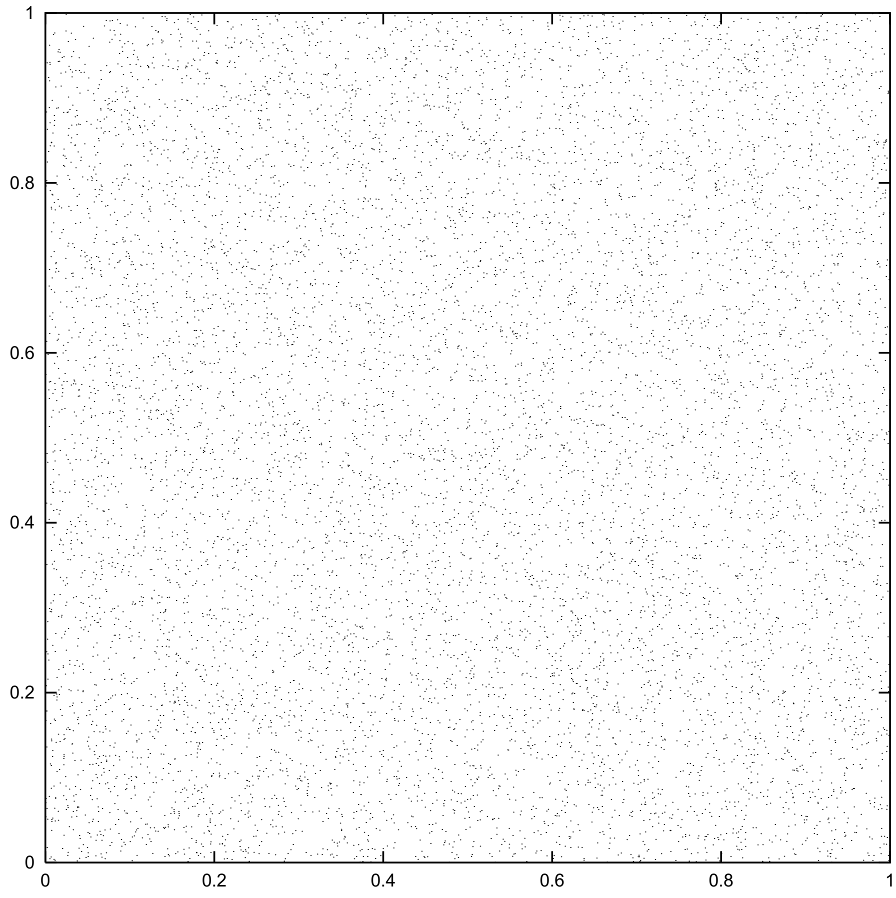

Fig. 2.3. 10000 (pseudo-)random points (U_{i-1}, U_i), calculated with Algorithm 2.7

2.2.1 Inversion

The following theorem is the basis for inversion methods.

Theorem 2.8 (inversion)

Suppose $U \sim \mathcal{U}[0,1]$ and F be a continuous strictly increasing distribution function. Then $F^{-1}(U)$ is a sample from F.

Proof: Let P denote the underlying probability.
$U \sim \mathcal{U}[0,1]$ means $\mathsf{P}(U \le \xi) = \xi$ for $0 \le \xi \le 1$.
Consequently

$$\mathsf{P}(F^{-1}(U) \le x) = \mathsf{P}(U \le F(x)) = F(x)\,.$$

Application

Following Theorem 2.8, the inversion method takes uniform deviates $u \sim \mathcal{U}[0,1]$ and sets $x = F^{-1}(u)$ ($\longrightarrow$ Exercises 2.5, 2.16). To judge the inversion method we consider the normal distribution as the most important example. Neither for its distribution function F nor for its inverse F^{-1} there is a closed-form expression ($\longrightarrow$ Exercise 1.3). So numerical methods are used. We discuss two approaches.

Numerical inversion means to calculate iteratively a solution x of the equation $F(x) = u$ for prescribed u. This iteration requires tricky termination criteria, in particular when x is large. Then we are in the situation $u \approx 1$, where tiny changes in u lead to large changes in x (Figure 2.4). The approximation of the solution x of $F(x) - u = 0$ can be calculated with bisection, or Newton's method, or the secant method ($\longrightarrow$ Appendix C1).

Alternatively the inversion $x = F^{-1}(u)$ can be approximated by a suitably constructed function $G(u)$,

$$G(u) \approx F^{-1}(u)\,.$$

Then only $x = G(u)$ needs to be evaluated. Constructing such an approximation formula G, it is important to realize that $F^{-1}(u)$ has "vertical" tangents at $u = 1$ (horizontal in Figure 2.4). This pole behavior must be reproduced correctly by the approximating function G. This suggests to use rational approximation ($\longrightarrow$ Appendix C1). For the Gaussian distribution one incorporates the point symmetry with respect to $(u, x) = (\frac{1}{2}, 0)$, and the pole at $u = 1$ (and hence at $u = 0$) in the *ansatz* for G ($\longrightarrow$ Exercise 2.6). Rational approximation of $F^{-1}(u)$ with a sufficiently large number of terms leads to high accuracy [Moro95]. The formulas are given in Appendix D2.

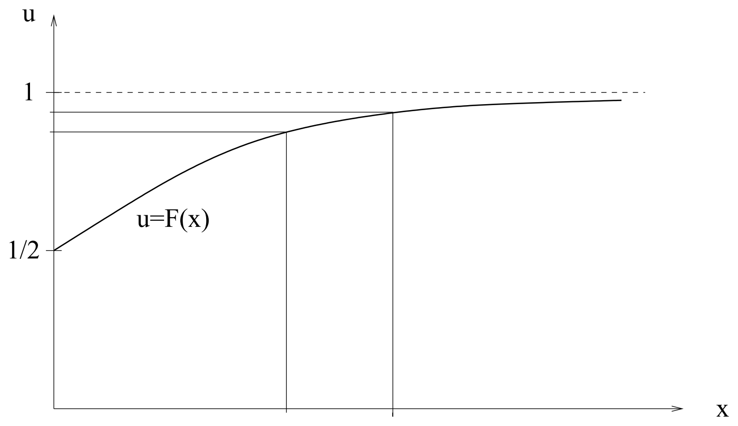

Fig. 2.4. Normal distribution; small changes in u can lead to large changes in x

2.2.2 Transformations in $\mathbb{R}^1$

Another class of methods uses transformations between random variables. We start the discussion with the scalar case. If we have a random variable X with *known* density and distribution, what can we say about the density and distribution of a transformed $h(X)$?

Theorem 2.9

Suppose X is a random variable with density $f(x)$ and distribution $F(x)$. Further assume $h : S \longrightarrow B$ with $S, B \subseteq \mathbb{R}$, where S is the support[2] of $f(x)$, and let h be strictly monotonous.

(a) Then $Y := h(X)$ is a random variable. Its distribution F_Y is

$$F_Y(y) = F(h^{-1}(y)) \quad \text{in case } h' > 0$$
$$F_Y(y) = 1 - F(h^{-1}(y)) \quad \text{in case } h' < 0.$$

(b) If h^{-1} is absolutely continuous then for almost all y the density of $h(X)$ is

$$f(h^{-1}(y)) \left| \frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y} \right| . \tag{2.6}$$

Proof:

(a) For $h' > 0$ we have $\mathsf{P}(h(X) \le y) = \mathsf{P}(X \le h^{-1}(y)) = F(h^{-1}(y))$.

(b) For absolutely continuous h^{-1} the density of $Y = h(X)$ is equal to the derivative of the distribution function almost everywhere. Evaluating the derivative $\frac{\mathrm{d}F(h^{-1}(y))}{\mathrm{d}y}$ with the chain rule implies the assertion. The absolute value in (2.6) is necessary such that a positive density comes out in case $h' < 0$. (See for instance [Fisz63], § 2.4 C.)

Application

Since we are able to calculate uniform deviates, we start from $X \sim \mathcal{U}[0,1]$ with f being the density of the uniform distribution,

$$f(x) = 1 \ \text{ for } \ 0 \le x \le 1, \quad \text{otherwise} \ \ f = 0 \, .$$

Here the support S is the unit interval. What we need are random numbers Y matching a prespecified target density $g(y)$. It remains to find a transformation h such that the density in (2.6) is identical to $g(y)$,

$$1 \cdot \left| \frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y} \right| = g(y) \, .$$

Then we only evaluate $h(X)$.

[2] f is zero outside S. (In this section, S is no asset price.)

Example 2.10 (exponential distribution)

The exponential distribution with parameter $\lambda > 0$ has the density

$$g(y) = \begin{cases} \lambda \mathrm{e}^{-\lambda y} & \text{for } y \geq 0 \\ 0 & \text{for } y < 0\,. \end{cases}$$

Here the range B consists of the nonnegative real numbers. The aim is to generate an exponentially distributed random variable Y out of a $\mathcal{U}[0,1]$-distributed random variable X. To this end we define the monotonous transformation from the unit interval $S = [0,1]$ into B by the decreasing function

$$y = h(x) := -\frac{1}{\lambda}\log x$$

with the inverse function $h^{-1}(y) = \mathrm{e}^{-\lambda y}$ for $y \geq 0$. For this h verify

$$f(h^{-1}(y))\left|\frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y}\right| = 1 \cdot \left|(-\lambda)\mathrm{e}^{-\lambda y}\right| = \lambda \mathrm{e}^{-\lambda y} = g(y)$$

as density of $h(X)$. Hence $h(X)$ is distributed exponentially.

Application:

In case $U_1, U_2, \ldots$ are nonzero uniform deviates, the numbers $h(U_i)$

$$-\frac{1}{\lambda}\log(U_1), \quad -\frac{1}{\lambda}\log(U_2), \; \ldots$$

are distributed exponentially. ($\longrightarrow$ Exercise 2.17)

Attempt to Generate a Normal Distribution

Starting from the uniform distribution ($f = 1$) a transformation $y = h(x)$ is searched such that its density equals that of the standard normal distribution,

$$1 \cdot \left|\frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y}\right| = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\,.$$

This is a differential equation for h^{-1} without analytical solution. As we will see, a transformation can be applied successfully in $\mathbb{R}^2$. To this end we need a generalization of the scalar transformation of Theorem 2.9 into $\mathbb{R}^n$.

2.2.3 Transformations in $\mathbb{R}^n$

The generalization of Theorem 2.9 to the vector case is

Theorem 2.11

Suppose X is a random variable in $\mathbb{R}^n$ with density $f(x) > 0$ on the support S. The transformation $h : S \to B,\ S, B \subseteq \mathbb{R}^n$ is assumed to be invertible and the inverse be continuously differentiable on B. $Y := h(X)$ is the transformed random variable. Then Y has the density

$$f(h^{-1}(y)) \left| \frac{\partial(x_1, ..., x_n)}{\partial(y_1, ..., y_n)} \right|, \quad y \in B\,, \tag{2.7}$$

where $x = h^{-1}(y)$ and $\frac{\partial(x_1,...,x_n)}{\partial(y_1,...,y_n)}$ is the determinant of the Jacobian matrix of all first-order derivatives of $h^{-1}(y)$.

(Theorem 4.2 in [Dev86])

2.3 Normally Distributed Random Variables

In this section the focus is on applying the transformation method in $\mathbb{R}^2$ to generate Gaussian random numbers. We describe the classical approach of Box and Muller. Inversion is one of several valid alternatives.[3]

2.3.1 Method of Box and Muller

To apply Theorem 2.11 we start with the unit square $S := [0,1]^2$ and the density (2.7) of the bivariate uniform distribution. The transformation is

$$\begin{cases} y_1 = \sqrt{-2\log x_1}\,\cos 2\pi x_2 =: h_1(x_1, x_2) \\ y_2 = \sqrt{-2\log x_1}\,\sin 2\pi x_2 =: h_2(x_1, x_2)\,, \end{cases} \tag{2.8}$$

$h(x)$ is defined on $[0,1]^2$ with values in $\mathbb{R}^2$. The inverse function h^{-1} is given by

$$\begin{cases} x_1 = \exp\left\{-\frac{1}{2}(y_1^2 + y_2^2)\right\} \\ x_2 = \dfrac{1}{2\pi} \arctan \dfrac{y_2}{y_1} \end{cases}$$

where we take the main branch of arctan. The determinant of the Jacobian matrix is

$$\begin{aligned} \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} &= \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} = \\ &= \frac{1}{2\pi} \exp\left\{-\tfrac{1}{2}(y_1^2 + y_2^2)\right\} \left(-y_1 \frac{1}{1 + \frac{y_2^2}{y_1^2}} \frac{1}{y_1} - y_2 \frac{1}{1 + \frac{y_2^2}{y_1^2}} \frac{y_2}{y_1^2} \right) \\ &= -\frac{1}{2\pi} \exp\left\{-\tfrac{1}{2}(y_1^2 + y_2^2)\right\}\,. \end{aligned}$$

This shows that $\left| \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} \right|$ is the density (2.7) of the bivariate standard normal distribution. Since this density is the product of the two one-dimensional densities,

[3] See also the Notes on this section.

$$\left|\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right| = \left[\frac{1}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}y_1^2\right)\right]\cdot\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}y_2^2\right)\right],$$

the two components of the vector y are independent. So, when the components of the vector X are $\sim \mathcal{U}[0,1]$, the vector $h(X)$ consists of two independent standard normal variates. Let us summarize the application of this transformation:

Algorithm 2.12 (Box–Muller)

```
(1) generate $U_1 \sim \mathcal{U}[0,1]$ and $U_2 \sim \mathcal{U}[0,1]$.
(2) $\theta := 2\pi U_2, \quad \rho := \sqrt{-2\log U_1}$
(3) $Z_1 := \rho\cos\theta$ is a normal variate
      (as well as $Z_2 := \rho\sin\theta$).
```

The variables U_1, U_2 stand for the components of X. Each application of the algorithm provides two standard normal variates. Note that a line structure in $[0,1]^2$ as in Example 2.5 is mapped to curves in the (Z_1, Z_2)-plane. This underlines the importance of excluding an evident line structure.

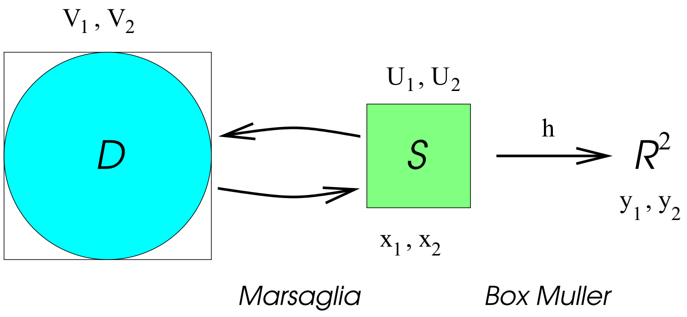

Fig. 2.5. Transformations of the Box–Muller–Marsaglia approach, schematically

2.3.2 Variant of Marsaglia

The variant of Marsaglia prepares the input in Algorithm 2.12 such that trigonometric functions are avoided. For $U \sim \mathcal{U}[0,1]$ we have $V := 2U - 1 \sim \mathcal{U}[-1,1]$. (Temporarily we misuse also the financial variable V for local purposes.) Two values V_1, V_2 calculated in this way define a point in the (V_1, V_2)-plane. Only points within the unit disk are accepted:

$$\mathcal{D} := \{\,(V_1,V_2) \mid V_1^2 + V_2^2 < 1\,\}\,;\ \text{accept only } (V_1,V_2)\in\mathcal{D}\,.$$

In case of rejection both values V_1, V_2 must be rejected. As a result, the surviving (V_1, V_2) are uniformly distributed on $\mathcal{D}$ with density $f(V_1, V_2) = \frac{1}{\pi}$ for $(V_1, V_2) \in \mathcal{D}$. A transformation from the disk $\mathcal{D}$ into the unit square $S := [0,1]^2$ is defined by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} V_1^2 + V_2^2 \\ \frac{1}{2\pi} \arg((V_1, V_2)) \end{pmatrix} .$$

That is, the Cartesian coordinates V_1, V_2 on $\mathcal{D}$ are mapped to the squared radius and the normalized angle.[4] For illustration, see Figure 2.5. These "polar coordinates" (x_1, x_2) are uniformly distributed on S ($\longrightarrow$ Exercise 2.7).

Application

For input in (2.8) use $V_1^2 + V_2^2$ as x_1 and $\frac{1}{2\pi} \arctan \frac{V_2}{V_1}$ as x_2. With these variables the relations

$$\cos 2\pi x_2 = \frac{V_1}{\sqrt{V_1^2 + V_2^2}} , \quad \sin 2\pi x_2 = \frac{V_2}{\sqrt{V_1^2 + V_2^2}} ,$$

hold, which means that it is no longer necessary to evaluate trigonometric functions. The resulting algorithm of Marsaglia has modified the Box–Muller method by constructing input values x_1, x_2 in a clever way.

Algorithm 2.13 (polar method)

```
(1) Repeat: generate $U_1, U_2 \sim \mathcal{U}[0,1]$;  $V_1 := 2U_1 - 1$,
    $V_2 := 2U_2 - 1$,  until  $W := V_1^2 + V_2^2 < 1$.
(2) $Z_1 := V_1 \sqrt{-2 \log(W)/W}$
    $Z_2 := V_2 \sqrt{-2 \log(W)/W}$
    are both standard normal variates.
```

The probability that $W < 1$ holds is given by the ratio of the areas, $\pi/4 = 0.785...$ So in about 21% of all $\mathcal{U}[0,1]$ drawings the (V_1, V_2)-tuple is rejected because of $W \geq 1$. Nevertheless the savings of the trigonometric evaluations makes Marsaglia's polar method more efficient than the Box–Muller method. Figure 2.6 illustrates normally distributed random numbers ($\longrightarrow$ Exercise 2.8).

[4] $\arg((V_1, V_2)) = \arctan(V_2/V_1)$ with the proper branch

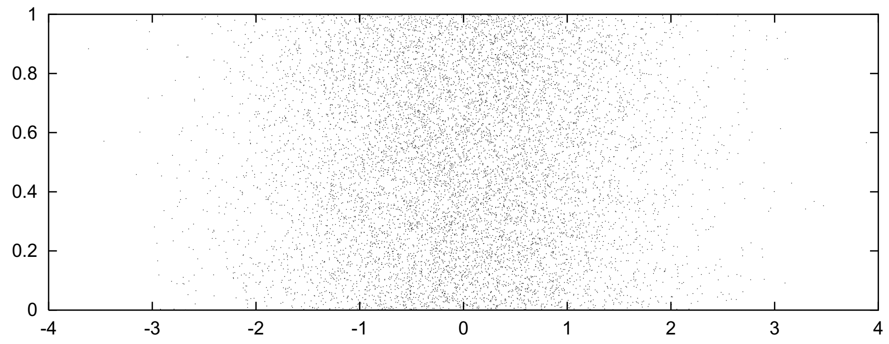

Fig. 2.6. 10000 numbers $\sim \mathcal{N}(0,1)$ (values entered horizontally and separated vertically with distance 10^{-4})

2.3.3 Correlated Random Variables

The above algorithms provide independent normal deviates. In many applications random variables are required that depend on each other in a prescribed way. Let us first recall the general n-dimensional density function.

Multivariate normal distribution (notations):

$$X = (X_1, ..., X_n), \quad \mu = \mathsf{E}X = (\mathsf{E}X_1, ..., \mathsf{E}X_n)$$

The covariance matrix (B1.8) of X is denoted Σ, and has elements

$$\Sigma_{ij} = (\mathsf{Cov}X)_{ij} := \mathsf{E}\left((X_i - \mu_i)(X_j - \mu_j)\right) , \quad \sigma_i^2 = \Sigma_{ii} ,$$

for $i, j = 1, \dots, n$. Using this notation, the correlation coefficients are

$$\rho_{ij} := \frac{\Sigma_{ij}}{\sigma_i \sigma_j} \qquad (\Rightarrow \ \rho_{ii} = 1) , \tag{2.9}$$

which set up the correlation matrix. The correlation matrix is a scaled version of Σ. The density function $f(x_1, ..., x_n)$ corresponding to $\mathcal{N}(\mu, \Sigma)$ is

$$f(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\det \Sigma)^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^{tr} \Sigma^{-1} (x-\mu) \right\} . \tag{2.10}$$

By theory, a covariance matrix (or correlation matrix) Σ is symmetric, and positive semidefinite. If in practice a matrix $\tilde{\Sigma}$ is corrupted by insufficient data, a close matrix Σ can be calculated with the features of a covariance matrix [Hig02]. In case $\det \Sigma \neq 0$ the matrix Σ is positive definite, which we assume now.

Below we shall need a factorization of Σ into $\Sigma = AA^{tr}$. From numerical mathematics we know that for symmetric positive definite matrices Σ the Cholesky decomposition $\Sigma = LL^{tr}$ exists, with a lower triangular matrix L ($\longrightarrow$ Appendix C1). There are numerous factorizations $\Sigma = AA^{tr}$ other than

Cholesky. A more involved factorization of Σ is the principal component analysis, which is based on eigenvectors ($\longrightarrow$ Exercise 2.18).

Transformation

Suppose $Z \sim \mathcal{N}(0, I)$ and $x = Az$, $A \in \mathbb{R}^{n\times n}$, where z is a realization of Z, 0 is the zero vector, and I the identity matrix. We see from

$$\exp\left\{-\frac{1}{2}z^{tr}z\right\} = \exp\left\{-\frac{1}{2}(A^{-1}x)^{tr}(A^{-1}x)\right\} = \exp\left\{-\frac{1}{2}x^{tr}A^{-tr}A^{-1}x\right\}$$

and from $\mathrm{d}x = |\det A|\mathrm{d}z$ that

$$\frac{1}{|\det A|}\exp\left\{-\frac{1}{2}x^{tr}(AA^{tr})^{-1}x\right\}\mathrm{d}x = \exp\left\{-\frac{1}{2}z^{tr}z\right\}\mathrm{d}z$$

holds for arbitrary nonsingular matrices A. To complete the transformation, we need a matrix A such that $\Sigma = AA^{tr}$. Then $|\det A| = (\det \Sigma)^{1/2}$, and the densities with the respect to x and z are converted correctly. In view of the general density $f(x)$ recalled in (2.10), AZ is normally distributed with $AZ \sim \mathcal{N}(0, AA^{tr})$, and hence the factorization $\Sigma = AA^{tr}$ implies

$$AZ \sim \mathcal{N}(0, \Sigma)\,.$$

Finally, translation with vector μ implies

$$\mu + AZ \sim \mathcal{N}(\mu, \Sigma)\,. \tag{2.11}$$

Application

Suppose we need a normal variate $X \sim \mathcal{N}(\mu, \Sigma)$ for given mean vector μ and covariance matrix Σ. This is most conveniently based on the Cholesky decomposition of Σ. Accordingly, the desired random variable can be calculated with the following algorithm:

Algorithm 2.14 (correlated random variable)

```
(1) Calculate the Cholesky decomposition $AA^{tr} = \Sigma$
(2) Calculate $Z \sim \mathcal{N}(0, I)$ componentwise
      by $Z_i \sim \mathcal{N}(0,1)$, $i = 1, ..., n$, for instance,
      with Marsaglia's polar algorithm
(3) $\mu + AZ$ has the desired distribution $\sim \mathcal{N}(\mu, \Sigma)$
```

Special case $n = 2$: In this case, in view of (2.9), only one correlation number is involved, namely, $\rho := \rho_{12} = \rho_{21}$, and the covariance matrix must be of the form

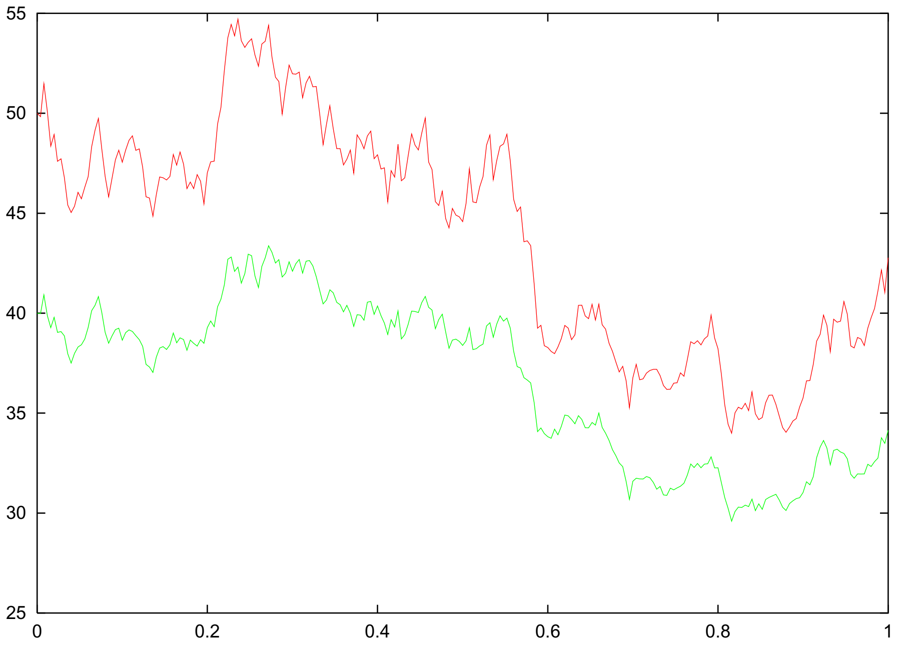

Fig. 2.7. Simulation of a correlated vector process with two components, and $\mu = 0.05$, $\sigma_1 = 0.3$, $\sigma_2 = 0.2$, $\rho = 0.85$, $\Delta t = 1/250$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} . \tag{2.12}$$

In this two-dimensional situation it makes sense to carry out the Cholesky decomposition analytically ($\longrightarrow$ Exercise 2.9). Figure 2.7 illustrates a highly correlated two-dimensional situation, with $\rho = 0.85$. An example based on (2.12) is (3.28).

2.4 Monte Carlo Integration

A classical application of random numbers is the Monte Carlo integration. The discussion in this section will serve as background for Quasi Monte Carlo, a topic of the following Section 2.5.

Let us begin with the one-dimensional situation. Assume a probability distribution with density g. Then the expectation of a function f is

$$\mathsf{E}(f) = \int_{-\infty}^{\infty} f(x)g(x)\,\mathrm{d}x\,,$$

compare (B1.4). For a definite integral on an interval $\mathcal{D} = [a, b]$, we use the uniform distribution with density

$$g = \frac{1}{b-a} \cdot 1_{\mathcal{D}} = \frac{1}{\lambda_1(\mathcal{D})} \cdot 1_{\mathcal{D}} \,,$$

where $\lambda_1(\mathcal{D})$ denotes the length of the interval $\mathcal{D}$. This leads to

$$\mathsf{E}(f) = \frac{1}{\lambda_1(\mathcal{D})} \int_a^b f(x)\,\mathrm{d}x \,,$$

or

$$\int_a^b f(x)\,\mathrm{d}x = \lambda_1(\mathcal{D}) \cdot \mathsf{E}(f) \,.$$

This equation is the basis of *Monte Carlo integration.* It remains to approximate $\mathsf{E}(f)$. For independent samples $x_i \sim \mathcal{U}[a, b]$ the law of large numbers ($\longrightarrow$ Appendix B1) establishes the estimator

$$\frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

as approximation to $\mathsf{E}(f)$. The approximation improves as the number of trials N goes to infinity; the error is characterized by the Central Limit Theorem.

This principle of the Monte Carlo Integration extends to the higher-dimensional case. Let $\mathcal{D} \subset \mathbb{R}^m$ be a domain on which the integral

$$\int_{\mathcal{D}} f(x)\,\mathrm{d}x$$

is to be calculated. For example, $\mathcal{D} = [0, 1]^m$. Such integrals occur in finance, for example, when mortgage-backed securities (CMO, collateralized mortgage obligations) are valuated [CaMO97]. The classical or *stochastic Monte Carlo integration* draws random samples $x_1, ..., x_N \in \mathcal{D}$ which should be independent and uniformly distributed. Then

$$\theta_N := \lambda_m(\mathcal{D}) \frac{1}{N} \sum_{i=1}^{N} f(x_i) \tag{2.13}$$

is an approximation of the integral. Here $\lambda_m(\mathcal{D})$ is the volume of $\mathcal{D}$ (or the m-dimensional Lebesgue measure [Nie92]). We assume $\lambda_m(\mathcal{D})$ to be finite. From the law of large numbers follows convergence of θ_N to $\lambda_m(\mathcal{D})\mathsf{E}(f) = \int_{\mathcal{D}} f(x)\,\mathrm{d}x$ for $N \to \infty$. The variance of the error

$$\delta_N := \int_{\mathcal{D}} f(x)\,\mathrm{d}x - \theta_N$$

satisfies

$$\mathsf{Var}(\delta_N) = \mathsf{E}(\delta_N^2) - (\mathsf{E}(\delta_N))^2 = \frac{\sigma^2(f)}{N}(\lambda_m(\mathcal{D}))^2\,, \tag{2.14a}$$

with the variance of f

$$\sigma^2(f) := \int_{\mathcal{D}} f(x)^2\,\mathrm{d}x - \left(\int_{\mathcal{D}} f(x)\,\mathrm{d}x\right)^2 . \tag{2.14b}$$

Hence the standard deviation of the error δ_N tends to 0 with the order $O(N^{-1/2})$. This result follows from the Central Limit Theorem or from other arguments ($\longrightarrow$ Exercise 2.10). The deficiency of the order $O(N^{-1/2})$ is the slow convergence ($\longrightarrow$ Exercise 2.11 and the second column in Table 2.1). To reach an absolute error of the order ε, equation (2.14a) tells that the sample size is $N = O(\varepsilon^{-2})$. To improve the accuracy by a factor of 10, the costs (that is the number of trials, N) increase by a factor of 100. Another disadvantage is the lack of a genuine error *bound*. The probabilistic error of (2.14) does not rule out the risk that the result may be completely wrong. The $\sigma^2(f)$ in (2.14b) is not known and must be approximated, which adds to the uncertainty of the error. And the Monte Carlo integration responds sensitively to changes of the initial state of the used random-number generator. This may be explained by the potential clustering of random points.

In many applications the above deficiencies are balanced by two good features of Monte Carlo integration: A first advantage is that the order $O(N^{-1/2})$ of the error holds independently of the dimension m. Another good feature is that the integrands f need not be smooth, square integrability suffices ($f \in \mathcal{L}^2$, see Appendix C3).

So far we have described the basic version of Monte Carlo integration, stressing the slow decline of the probabilistic error with growing N. The variance of the error δ can also be diminished by decreasing the numerator in (2.14a). This variance of the problem can be reduced by suitable methods. (We will come back to this issue in Section 3.5.4.)

We conclude the excursion into the stochastic Monte Carlo integration with the variant for those cases in which $\lambda_m(\mathcal{D})$ is hard to calculate. For $\mathcal{D} \subseteq [0,1]^m$ and $x_1, ..., x_N \sim \mathcal{U}[0,1]^m$ use

$$\int_{\mathcal{D}} f(x)\,\mathrm{d}x \approx \frac{1}{N} \sum_{\substack{i=1 \\ x_i \in \mathcal{D}}}^{N} f(x_i)\,. \tag{2.15}$$

For the integral (1.50) with density f_{GBM} see Section 3.5.

2.5 Sequences of Numbers with Low Discrepancy

One difficulty with random numbers is that they may fail to distribute uniformly. Here, "uniform" is not meant in the stochastic sense of a distribution $\sim \mathcal{U}[0,1]$, but has the meaning of an equidistributedness that avoids extreme clustering or holes. The aim is to generate numbers for which the deviation from uniformity is minimal. This deviation is called "discrepancy." Another objective is to obtain good convergence for some important applications.

2.5.1 Discrepancy

The bad convergence behavior of the stochastic Monte Carlo integration is not inevitable. For example, for $m = 1$ and $\mathcal{D} = [0,1]$ an equidistant x-grid with mesh size $1/N$ leads to a formula (2.13) that resembles the trapezoidal sum ((C1.2) in Appendix C1). For smooth f, the order of the error is at least $O(N^{-1})$. (Why?) But such a grid-based evaluation procedure is somewhat inflexible because the grid must be prescribed in advance and the number N that matches the desired accuracy is unknown beforehand. In contrast, the free placing of sample points with Monte Carlo integration can be performed until some termination criterion is met. It would be desirable to find a compromise in placing sample points such that the fineness advances but clustering is avoided. The sample points should fill the integration domain $\mathcal{D}$ as uniformly as possible. To this end we require a measure of the equidistributedness.

Let $Q \subseteq [0,1]^m$ be an arbitrary axially parallel m-dimensional rectangle in the unit cube $[0,1]^m$ of $\mathbb{R}^m$. That is, Q is a product of m intervals. Suppose a set of points $x_1, ..., x_N \in [0,1]^m$. The decisive idea behind discrepancy is that for an evenly distributed point set, the fraction of the points lying within the rectangle Q should correspond to the volume of the rectangle (see Figure 2.8). Let # denote the number of points, then the goal is

$$\frac{\#\text{ of } x_i \in Q}{\#\text{ of all points in } [0,1]^m} \approx \frac{\text{vol}(Q)}{\text{vol}([0,1]^m)}$$

for as many rectangles as possible. This leads to the following definition:

Definition 2.15 (discrepancy)

The discrepancy of the point set $\{x_1, ..., x_N\} \subset [0,1]^m$ is

$$D_N := \sup_Q \left| \frac{\#\text{ of } x_i \in Q}{N} - \text{vol}(Q) \right| .$$

Analogously the variant D_N^* *(star discrepancy)* is obtained when the set of rectangles is restricted to those Q^*, for which one corner is the origin:

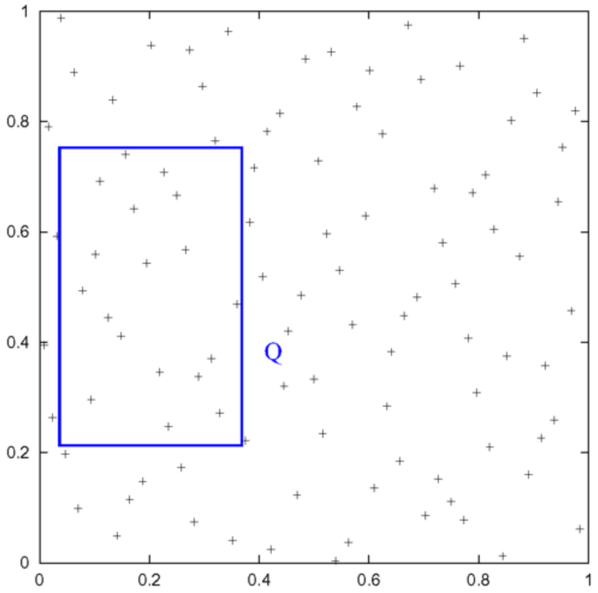

Fig. 2.8 On the idea of discrepancy

Table 2.1 Comparison of different convergence rates to zero

N	$\frac{1}{\sqrt{N}}$	$\sqrt{\frac{\log\log N}{N}}$	$\frac{\log N}{N}$	$\frac{(\log N)^2}{N}$	$\frac{(\log N)^3}{N}$
10^1	.31622777	.28879620	.23025851	.53018981	1.22080716
10^2	.10000000	.12357911	.04605170	.21207592	.97664572
10^3	.03162278	.04396186	.00690776	.04771708	.32961793
10^4	.01000000	.01490076	.00092103	.00848304	.07813166
10^5	.00316228	.00494315	.00011513	.00132547	.01526009
10^6	.00100000	.00162043	.00001382	.00019087	.00263694
10^7	.00031623	.00052725	.00000161	.00002598	.00041874
10^8	.00010000	.00017069	.00000018	.00000339	.00006251
10^9	.00003162	.00005506	.00000002	.00000043	.00000890

$$Q^* = \prod_{i=1}^{m} [0, y_i)$$

where $y \in \mathbb{R}^m$ denotes the corner diagonally opposite the origin.

The more evenly the points of a sequence are distributed, the closer the discrepancy D_N is to zero. Here D_N refers to the first N points of a sequence of points (x_i), $i \geq 1$. The discrepancies D_N and D_N^* satisfy ($\longrightarrow$ Exercise 2.12b)

$$D_N^* \leq D_N \leq 2^m D_N^* \,.$$

The discrepancy allows to find a deterministic bound on the error δ_N of the Monte Carlo integration,

$$|\delta_N| \le v(f) D_N^* ; \tag{2.16}$$

here $v(f)$ is the variation of the function f with $v(f) < \infty$, and the domain of integration is $\mathcal{D} = [0,1]^m$ [Nie92], [TrW92], [MoC94]. This result is known as Theorem of Koksma and Hlawka. The bound in (2.16) underlines the importance to find numbers $x_1, ..., x_N$ with small value of the discrepancy D_N. After all, a set of N *randomly* chosen points satisfies

$$\mathsf{E}(D_N) = O\left(\sqrt{\frac{\log\log N}{N}}\right) .$$

This is in accordance with the $O(N^{-1/2})$ law. The order of magnitude of these numbers is shown in Table 2.1 (third column).

Definition 2.16 (low-discrepancy point sequence)

A sequence of points or numbers $x_1, x_2, ..., x_N, ... \in [0,1]^m$ is called low-discrepancy sequence if

$$D_N = O\left(\frac{(\log N)^m}{N}\right) . \tag{2.17}$$

Deterministic sequences of numbers satisfying (2.17) are also called *quasi*-random numbers, although they are fully deterministic. Table 2.1 reports on the orders of magnitude. Since $\log(N)$ grows only modestly, a low discrepancy essentially means $D_N \approx O(N^{-1})$ as long as the dimension m is small. The equation (2.17) expresses some dependence on the dimension m, contrary to Monte Carlo methods. But the dependence on m in (2.17) is less stringent than with classical quadrature.

2.5.2 Examples of Low-Discrepancy Sequences

In the one-dimensional case ($m = 1$) the point set

$$x_i = \frac{2i-1}{2N} , \quad i = 1, ..., N \tag{2.18}$$

has the value $D_N^* = \frac{1}{2N}$; this value can not be improved ($\longrightarrow$ Exercise 2.12c). The monotonous sequence (2.18) can be applied only when a reasonable N is known and fixed; for $N \to \infty$ the x_i would be newly placed and an integrand f evaluated again. Since N is large, it is essential that the previously calculated results can be used when N is growing. This means that the points $x_1, x_2, ...$ must be placed "dynamically" so that they are preserved and the fineness improves when N grows. This is achieved by the sequence

$$\frac{1}{2}, \ \frac{1}{4}, \frac{3}{4}, \ \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \ \frac{1}{16}, \ldots$$

This sequence is known as van der Corput sequence. To motivate such a dynamical placing of points imagine that you are searching for some item in the interval $[0, 1]$ (or in the cube $[0, 1]^m$). The searching must be fast and successful, and is terminated as soon as the object is found. This defines N dynamically by the process.

The formula that defines the van der Corput sequence can be formulated as algorithm. Let us study an example, say, $x_6 = \frac{3}{8}$. The index $i = 6$ is written as binary number

$$6 = (110)_2 =: (d_2 \ d_1 \ d_0)_2 \quad \text{with} \quad d_i \in \{0, 1\} .$$

Then reverse the binary digits and put the radix point in front of the sequence:

$$(.\ d_0 \ d_1 \ d_2)_2 = \frac{d_0}{2} + \frac{d_1}{2^2} + \frac{d_3}{2^3} = \frac{1}{2^2} + \frac{1}{2^3} = \frac{3}{8}$$

If this is done for all indices $i = 1, 2, 3, \ldots$ the van der Corput sequence $x_1, x_2, x_3, \ldots$ results. These numbers can be defined with the following function:

Definition 2.17 (radical-inverse function)

For $i = 1, 2, \ldots$ let j be given by the expansion in base b (integer ≥ 2)

$$i = \sum_{k=0}^{j} d_k b^k ,$$

with digits $d_k \in \{0, 1, \ldots, b-1\}$, which depend on b, i. Then the radical-inverse function is defined by

$$\phi_b(i) := \sum_{k=0}^{j} d_k b^{-k-1} .$$

The function $\phi_b(i)$ is the digit-reversed fraction of i. This mapping may be seen as reflecting with respect to the radix point. To each index i a rational number $\phi_b(i)$ in the interval $0 < x < 1$ is assigned. Every time the number of digits j increases by one, the mesh becomes finer by a factor $1/b$. This means that the algorithm fills all mesh points on the sequence of meshes with increasing fineness ($\longrightarrow$ Exercise 2.13). The above classical van der Corput sequence is obtained by

$$x_i := \phi_2(i) .$$

The radical-inverse function can be applied to construct points x_i in the m-dimensional cube $[0, 1]^m$. The simplest construction is the Halton sequence.

Definition 2.18 (Halton sequence)

Let $p_1, ..., p_m$ be pairwise prime integers. The Halton sequence is defined as the sequence of vectors

$$x_i := (\phi_{p_1}(i), ..., \phi_{p_m}(i)) \ , \quad i = 1, 2, ...$$

Usually one takes $p_1, ..., p_m$ as the first m prime numbers. Figure 2.9 shows for $m = 2$ and $p_1 = 2$, $p_2 = 3$ the first 10000 Halton points. Compared to the pseudo-random points of Figure 2.3, the Halton points are distributed more evenly.

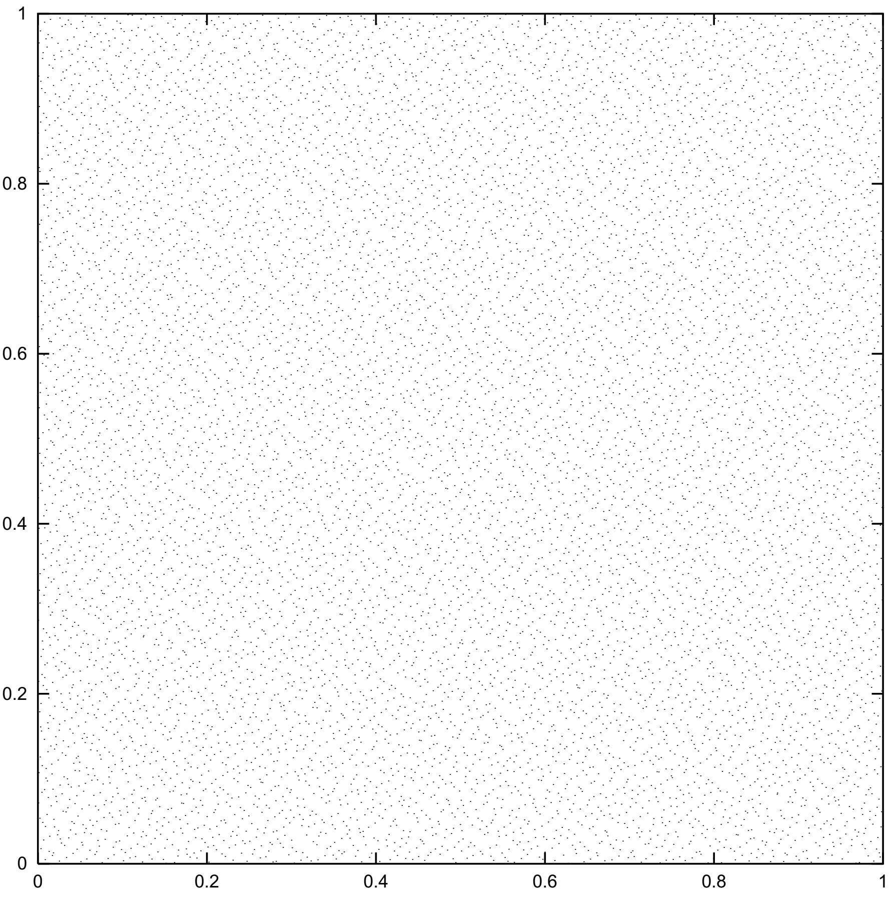

Fig. 2.9. 10000 Halton points from Definition 2.18, with $p_1 = 2$, $p_2 = 3$

Halton sequences x_i of Definition 2.18 are easily constructed, but fail to be uniform when the dimension m is high, see Section 5.2 in [Gla04]. Then correlations between the radical-inverse functions for different dimensions are observed. This problem can be cured with a simple modification of the Halton sequence, namely, by using only every lth Halton number [KoW97]. The leap l is a prime different from all bases $p_1, \ldots, p_m$. The result is the "Halton sequence leaped"

$$x_k := (\phi_{p_1}(lk), ..., \phi_{p_m}(lk)) \ , \quad k = 1, 2, ... \tag{2.19}$$

This modification has shown good performance for dimensions at least up to $m = 400$. As reported in [KoW97], $l = 409$ is one example of a good leap value.

Other sequences have been constructed out of the van der Corput sequence. These include the sequences developed by Sobol, Faure and Niederreiter, see [Nie92], [MoC94], [PrTVF92]. All these sequences are of low discrepancy, with

$$N \cdot D_N^* \leq C_m (\log N)^m + O\left((\log N)^{m-1}\right) \, .$$

The Table 2.1 shows how fast the relevant terms $(\log N)^m / N$ tend to zero. If m is large, extremely large values of the denominator N are needed before the terms become small. But it is assumed that the bounds are unrealistically large and overestimate the real error. For the Halton sequence in case $m = 2$ the constant is $C_2 = 0.2602$.

Quasi Monte Carlo (QMC) methods approximate the integrals with the arithmetic mean θ_N of (2.13), but use low-discrepancy numbers x_i instead of random numbers. QMC is a *deterministic* method. Practical experience with low-discrepancy sequences are better than might be expected from the bounds known so far. This also holds for the bound (2.16) by Koksma and Hlawka; apparently a large class of functions f satisfy $|\delta_N| \ll v(f) D_N^*$, see [SpM94].

Notes and Comments

on Section 2.1:

The linear congruential method is sometimes called Lehmer generator. Easily accessible and popular generators are RAN1 and RAN2 from [PrTVF92]. Further references on linear congruential generators are [Mar68], [Rip87], [Nie92], [LEc99]. Example 2.4 is from [Fis96], and Example 2.5 from [Rip87]. Nonlinear congruential generators are of the form

$$N_i = f(N_{i-1}) \mod M \, .$$

Hints on the algorithmic implementation are found in [Gen98]. Generally it is advisable to run the generator in integer arithmetic in order to avoid rounding errors that may spoil the period, see [Lehn02]. For Fibonacci generators we refer to [Bre94]. The version of (2.5) is a subtractive generator. Additive versions (with a plus sign instead of the minus sign) are used as well [Knu95], [Gen98]. The codes in [PrTVF92] are recommendable. For simple statistical tests with illustrations see [Hig04].

There are multiplicative Fibonacci generators of the form

$$N_{i+1} := N_{i-\nu} N_{i-\mu} \mod M \, .$$

Hints on parallelization are given in [Mas99]. For example, parallel Fibonacci generators are obtained by different initializing sequences. Note that computer systems and software packages often provide built-in random number generators. But often these generators are not clearly specified, and should be handled with care.

on Sections 2.2, 2.3:

The inversion result of Theorem 2.8 can be formulated placing less or no restrictions on F, see [Rip87], p. 59, [Dev86], p. 28, or [Lan99], p. 270. There are numerous other methods to calculate normal and non normal variates; for a detailed overview with many references see [Dev86]. The Box–Muller approach was suggested in [BoM58]. Marsaglia's modification was published in a report quoted in [MaB64]. Several algorithms are based on the rejection method [Dev86], [Fis96]. Fast algorithms include the "ziggurat" generator, which works with precomputed tables [MaT00], and the Wallace algorithm [Wal96], which works with a pool of random numbers and suitable transformations. Platform-dependent implementation details place emphasis on the one or the other advantage. A survey on Gaussian random number generators is [ThLLV07]. For simulating Lévy processes, see [ConT04]. For singular symmetric positive *semi*definite matrices Σ ($x^{tr}\Sigma x \geq 0$ for all x), the Cholesky decomposition can be cured, see [GoV96], or [Gla04].

on Section 2.4:

The bounds on errors of the Monte Carlo integration refer to arbitrary functions f; for smooth functions better bounds can be expected. In the one-dimensional case the variation is defined as the supremum of $\sum_j |f(t_j) - f(t_{j-1})|$ over all partitions, see Section 1.6.2. This definition can be generalized to higher-dimensional cases. A thorough discussion is [Nie78], [Nie92].

An advanced application of Monte Carlo integration uses one or more methods of *reduction of variance*, which allows to improve the accuracy in many cases [HaH64], [Rub81], [Nie92], [PrTVF92], [Fis96], [Kwok98], [Lan99]. For example, the integration domain can be split into subsets *(stratified sampling)* [RiW03]. Another technique is used when for a *control variate* g with $g \approx f$ the exact integral is known. Then f is replaced by $(f - g) + g$ and Monte Carlo integration is applied to $f - g$. Another alternative, the method of *antithetic variates*, will be described in Section 3.5.4 together with the control-variate technique.

on Section 2.5:

Besides the supremum discrepancy of Definition 2.15 the $\mathcal{L}^2$-analogy of an integral version is used. Hints on speed and preliminary comparison are found in [MoC94]. For application on high-dimensional integrals see [PaT95]. For large values of the dimension m, the error (2.17) takes large values, which might suggest to discard its use. But the notion of an *effective dimension* and practical results give a favorable picture at least for CMO applications of order $m = 360$ [CaMO97]. The error bound of Koksma and Hlawka (2.16) is not necessarily recommendable for practical use, see the discussion in [SpM94]. The analogy of the equidistant lattice in (2.18) in higher-dimensional space has unfavorable values of the discrepancy, $D_N = O\left(\frac{1}{\sqrt[m]{N}}\right)$. For $m > 2$ this is worse than Monte Carlo, compare [Rip87]. — Monte Carlo does not take advantage of smoothness of integrands. In the case of smooth integrands, sparse-grid approaches are highly competitive. These most refined quadrature methods moderate the *curse of the dimension*, see [GeG98], [GeG03], [Rei04].

Van der Corput sequences can be based also on other bases. Halton's paper is [Hal60]. Computer programs that generate low-discrepancy numbers are available. For example, Sobol numbers are calculated in [PrTVF92] and Sobol- and Faure numbers in the computer program FINDER [PaT95] and in [Tez95]. At the current state of the art it is open which point set has the smallest discrepancy in the m-dimensional cube. There are generalized Niederreiter sequences, which include Sobol- and Faure sequences as special cases [Tez95]. In several applications deterministic Monte Carlo seems to be superior to stochastic Monte Carlo [PaT96]. A comparison based on finance applications has shown good performance of Sobol numbers; in [Jon11] Sobol numbers are outperformed by Halton sequences leaped (2.19). [NiJ95] and Chapter 5 in [Gla04] provide more discussion and many references.

Besides volume integration, Monte Carlo is needed to integrate over possibly high-dimensional probability distributions. Drawing samples from the required distribution can be done by running a cleverly constructed Markov chain. This kind of method is called Markov Chain Monte Carlo (MCMC). That is, a chain of random variables $X_0, X_1, X_2, \ldots$ is constructed where for given X_j the next state X_{j+1} does not depend on the history of the chain $X_0, X_1, X_2, \ldots, X_{j-1}$. By suitable construction criteria, convergence to any chosen target distribution is obtained. For MCMC we refer to the literature, for example to [GiRS96], [Lan99], [Beh00], [Tsay02], [Häg02].

Exercises

Exercise 2.1

Consider the random number generator $N_i = 2N_{i-1} \bmod 11$. For $(N_{i-1}, N_i) \in \{0, 1, ..., 10\}^2$ and integer tuples with $z_0 + 2z_1 = 0 \bmod 11$ the equation

$$z_0 N_{i-1} + z_1 N_i = 0 \bmod 11$$

defines families of parallel straight lines, on which all points (N_{i-1}, N_i) lie. These straight lines are to be analyzed. For which of the families of parallel straight lines are the gaps maximal?

Exercise 2.2 Deficient Random Number Generator

For some time the generator

$$N_i = aN_{i-1} \bmod M, \quad \text{with } a = 2^{16} + 3,\ M = 2^{31}$$

was in wide use. Show for the sequence $U_i := N_i/M$

$$U_{i+2} - 6U_{i+1} + 9U_i \quad \text{is integer!}$$

What does this imply for the distribution of the triples (U_i, U_{i+1}, U_{i+2}) in the unit cube?

Exercise 2.3 Lattice of the Linear Congruential Generator

a) Show by induction over j

$$N_{i+j} - N_j = a^j (N_i - N_0) \bmod M$$

b) Show for integer $z_0, z_1, ..., z_{m-1}$

$$\begin{pmatrix} N_i \\ N_{i+1} \\ \vdots \\ N_{i+m-1} \end{pmatrix} - \begin{pmatrix} N_0 \\ N_1 \\ \vdots \\ N_{m-1} \end{pmatrix} = (N_i - N_0) \begin{pmatrix} 1 \\ a \\ \vdots \\ a^{m-1} \end{pmatrix} + M \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{m-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a & M & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a^{m-1} & 0 & \cdots & M \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{m-1} \end{pmatrix}$$

Exercise 2.4 Coarse Approximation of Normal Deviates

Let $U_1, U_2, ...$ be independent random numbers $\sim \mathcal{U}[0, 1]$, and

$$X_k := \sum_{i=k}^{k+11} U_i - 6\,.$$

Calculate mean and variance of the X_k.

Exercise 2.5 Cauchy-Distributed Random Numbers

A Cauchy-distributed random variable has the density function

$$f_c(x) := \frac{c}{\pi} \frac{1}{c^2 + x^2} .$$

Show that its distribution function F_c and its inverse F_c^{-1} are

$$F_c(x) = \frac{1}{\pi} \arctan \frac{x}{c} + \frac{1}{2} \quad , \quad F_c^{-1}(y) = c \tan(\pi(y - \frac{1}{2})) .$$

How can this be used to generate Cauchy-distributed random numbers out of uniform deviates?

Exercise 2.6 Inverting the Normal Distribution

Suppose $F(x)$ is the standard normal distribution function. Construct a rough approximation $G(u)$ to $F^{-1}(u)$ for $0.5 \leq u < 1$ as follows:

a) Construct a rational function $G(u)$ ($\longrightarrow$ Appendix C1) with correct asymptotic behavior, point symmetry with respect to $(u, x) = (0.5, 0)$, using only one parameter.
b) Fix the parameter by interpolating a given point $(x_1, F(x_1))$.
c) What is a simple criterion for the error of the approximation?

Exercise 2.7 Uniform Distribution

For the uniformly distributed random variables $(V_1,\, V_2)$ on $[-1, 1]^2$ consider the transformation

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} V_1^2 + V_2^2 \\ \frac{1}{2\pi} \arg((V_1, V_2)) \end{pmatrix}$$

where $\arg((V_1, V_2))$ denotes the corresponding angle. Show that $(X_1,\, X_2)$ is distributed uniformly.

Exercise 2.8 Programming Assignment: Normal Deviates

a) Write a computer program that implements the *Fibonacci generator*

$$\begin{aligned} U_i :=& U_{i-17} - U_{i-5} \\ U_i :=& U_i + 1 \text{ in case } U_i < 0 \end{aligned}$$

in the form of Algorithm 2.7.
Tests: Visual inspection of 10000 points in the unit square.
b) Write a computer program that implements *Marsaglia's Polar Algorithm* (Algorithm 2.13). Use the uniform deviates from a).

Tests:

1.) For a sample of 5000 points calculate estimates of mean and variance.

2.) For the discretized SDE

$$\Delta x = 0.1\Delta t + Z\sqrt{\Delta t}, \quad Z \sim \mathcal{N}(0,1)$$

calculate some trajectories for $0 \leq t \leq 1,\ \Delta t = 0.01,\ x_0 = 0$.

Exercise 2.9 Correlated Distributions

Suppose we need a two-dimensional random variable (X_1, X_2) that must be normally distributed with mean 0, and given variances σ_1^2, σ_2^2 and prespecified correlation ρ. How is X_1, X_2 obtained out of $Z_1, Z_2 \sim \mathcal{N}(0,1)$?

Exercise 2.10 Error of the Monte Carlo Integration

The domain for integration is $Q = [0,1]^m$. For

$$\Theta_N := \frac{1}{N}\sum_{i=1}^{N} f(x_i)\ , \quad \mathsf{E}(f) := \int f\,\mathrm{d}x\ , \quad g := f - \mathsf{E}(f)$$

and $\sigma^2(f)$ from (2.14b) show

a) $\mathsf{E}(g) = 0$

b) $\sigma^2(g) = \sigma^2(f)$

c) $\sigma^2(\delta_N) = \mathsf{E}(\delta_N^2) = \frac{1}{N^2}\int(\sum g(x_i))^2\,\mathrm{d}x = \frac{1}{N}\sigma^2(f)$
Hint on (c): When the random points x_i are i.i.d. (independent identical distributed), then also $f(x_i)$ and $g(x_i)$ are i.i.d. A consequence is $\int g(x_i)g(x_j)\,\mathrm{d}x = 0$ for $i \neq j$.

Exercise 2.11 Experiment on Monte Carlo Integration

To approximate the integral

$$\int_0^1 f(x)\,\mathrm{d}x$$

calculate a Monte Carlo sum

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

for $f(x) = 5x^4$ and, for example, $N = 100000$ random numbers $x_i \sim \mathcal{U}[0,1]$. The absolute error behaves like $cN^{-1/2}$. Compare the approximation with the exact integral for several N and seeds to obtain an estimate of c.

Exercise 2.12 Bounds on the Discrepancy

(Compare Definition 2.15) Show

a) $0 \leq D_N \leq 1$,

b) $D_N^* \leq D_N \leq 2^m D_N^*$ (show this at least for $m \leq 2$),

c) $D_N^* \geq \frac{1}{2N}$ for $m = 1$.

Exercise 2.13 Algorithm for the Radical-Inverse Function

Use the idea

$$i = \left(d_k b^{k-1} + ... + d_1\right) b + d_0$$

to formulate an algorithm that obtains $d_0, d_1, ..., d_k$ by repeated division by b. Reformulate $\phi_b(i)$ from Definition 2.17 into the form $\phi_b(i) = z/b^{j+1}$ such that the result is represented as rational number. The numerator z should be calculated in the same loop that establishes the digits $d_0, ..., d_k$.

Exercise 2.14 Testing the Distribution

Let X be a random variate with density f and let $a_1 < a_2 < ... < a_l$ define a partition of the support of f into subintervals, including the unbounded intervals $x < a_1$ and $x > a_l$. Recall from (B1.1), (B1.2) that the probability of a realization of X falling into $a_k \leq x < a_{k+1}$ is given by

$$p_k := \int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x \ , \ k = 1, 2, \ldots, l-1 \, ,$$

which can be approximated by $(a_{k+1} - a_k) f\left(\frac{a_k + a_{k+1}}{2}\right)$. Perform a sample of j realizations $x_1, \ldots, x_j$ of a random number generator, and denote j_k the number of samples falling into $a_k \leq x < a_{k+1}$. For normal variates with density f from (B1.9) design an algorithm that performs a simple statistical test of the quality of the $x_1, \ldots, x_j$.

Hints: See Section 2.1 for the special case of uniform variates. Argue for what choices of a_1 and a_l the probabilities p_0 and p_l may be neglected. Think about a reasonable relation between l and j.

Exercise 2.15 Quality of Fibonacci-Generated Numbers

Analyze and visualize the planes in the unit cube, on which all points fall that are generated by the Fibonacci recursion

$$U_{i+1} := (U_i + U_{i-1}) \bmod 1 \, .$$

Exercise 2.16

Use the inversion method and uniformly distributed $U \sim \mathcal{U}[0, 1]$ to calculate a stochastic variable X with distribution

$$F(x) = 1 - \mathrm{e}^{-2x(x-a)} \ , \ x \geq a \, .$$

Exercise 2.17 Time-Changed Wiener Process

For a time-changing function $\tau(t)$ set $\tau_j := \tau(j\,\Delta t)$ for some time increment Δt.

a) Argue why Algorithm 1.8 changes to $W_j = W_{j-1} + Z\sqrt{\tau_j - \tau_{j-1}}$ (last line).

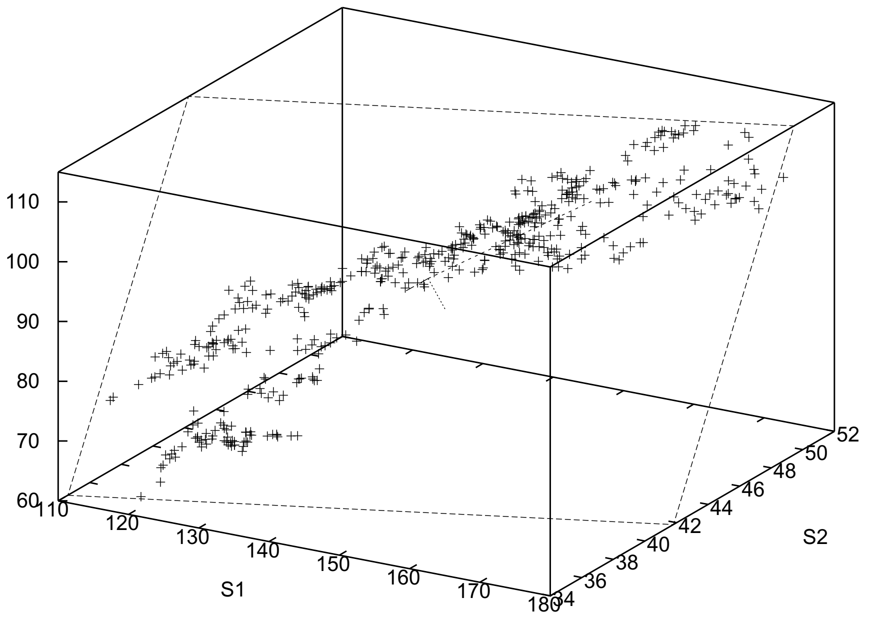

Fig. 2.10. Prices of the DAX assets Allianz (S1), BMW (S2), and HeidelbergCement; 500 trading days from Nov 5, 2005; eigenvalues of the covariance matrix are 400.8, 25.8, 2.73; eigenvectors centered at the mean point and scaled by $\sqrt{\lambda}$ are shown, and the plane spanned by $v^{(1)}, v^{(2)}$.

b) Let τ_j be the exponentially distributed jump instances of a Poisson experiment, see Section 1.9 and Property 1.20e. How should the jump intensity λ be chosen such that the expectation of the $\Delta\tau$ is Δt? Implement and test the algorithm, and visualize the results. Experiment with several values of the jump intensity λ.

Exercise 2.18 Spectral Decomposition of a Covariance Matrix

For symmetric positive definite $n \times n$ matrices Σ there exists a set of orthonormal eigenvectors $v^{(1)}, \ldots, v^{(n)}$ and eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n > 0$ such that

$$\Sigma v^{(j)} = \lambda_j v^{(j)}, \quad j = 1, \ldots, n.$$

Arrange the n eigenvector columns into the $n \times n$ matrix $B := (v^{(1)}, \ldots, v^{(n)})$, and the eigenvalues into the diagonal matrices $\Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\Lambda^{\frac{1}{2}} := \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$.

a) Show $\Sigma B = B\Lambda$.

b) Show that

$$A := B\Lambda^{\frac{1}{2}}$$

factorizes Σ in the sense $\Sigma = AA^{tr}$.

c) Show

$$AZ = \sum_{j=1}^{n} \sqrt{\lambda_j}\, Z_j\, v^{(j)}$$

d) And the reversal of Section 2.3.3 holds: For a random vector $X \sim \mathcal{N}(0, \Sigma)$ the transformed random vector $A^{-1}X$ has uncorrelated components: Show $\mathsf{Cov}(A^{-1}X) = I$ and $\mathsf{Cov}(B^{-1}X) = \Lambda$.

e) For the 2×2 matrix

$$\Sigma = \begin{pmatrix} 5 & 1 \\ 1 & 10 \end{pmatrix}$$

calculate the Cholesky decomposition and $B\Lambda^{\frac{1}{2}}$.

Hint: The above is the essence of the *principal component analysis*. Here Σ represents a covariance matrix or a correlation matrix. (For an example see Figure 2.10.) The matrix B and the eigenvalues in Λ reveal the structure of the data. B defines a linear transformation of the data to a rectangular coordinate system, and the eigenvalues λ_j measure the corresponding variances. In case $\lambda_{k+1} \gg \lambda_k$ for some index k, the sum in c) can be truncated after the kth term in order to reduce the dimension. The computation of B and Λ (and hence A) is costly, but a dominating λ_1 allows for a simple approximation of $v^{(1)}$ by the power method.