Approximation and simulation of processes and distributions

T. Mikosch

Contents

1 Preliminaries

In this section we list some results on approximations to distribution for sums of independent random variables which we will frequently refer to. Some of the results you might know from a course in statistics or elementary probability theory.

1.1 Asymptotic results

1.1.1 The law of large numbers

Let $X_1, X_2, \ldots$ be an iid sequence of random variables with common distribution function F and denote by X a generic element of this sequence. One of the fundamental results of probability theory (and actually one of the columns on which this theory stands) is *Kolmogorov's strong long of large numbers* (SLLN). It says that the sequence of sample means

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges to a finite constant a a.s. if and only if $E|X| < \infty$, and then $a = EX$ necessarily. The sufficiency part in the SLLN is also referred to as *ergodic theorem*; it remains valid for finite mean stationary ergodic sequences (X_n).

1.1.2 The Glivenko–Cantelli result

One of the basic tools in statistics is the *empirical distribution function*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}} = n^{-1}\,\#\{i \le n \ : X_i \le x\}\,, \quad x \in \mathbb{R}\,.$$

For given $X_1(\omega), X_2(\omega), \ldots$, F_n is indeed a probability distribution function with atoms at $X_1(\omega)$, $X_2(\omega), \ldots, X_n(\omega)$, but since it depends on the outcomes $\omega \in \Omega$ it is a random distribution function. The empirical distribution function contains major information about the underlying distribution function F of iid observations. As $n \to \infty$, the SLLN implies that

$$F_n(x) \overset{\text{a.s.}}{\to} F(x) \tag{1.1}$$

for every fixed x. For continuous F, using the uniform continuity and monotonicity of F, it is not difficult to show the Glivenko–Cantelli result which says that (1.1) holds uniformly in x:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \overset{\text{a.s.}}{\to} 0\,.$$

This result can even be shown to be true for any distribution function F. In combination with the so-called *empirical functional central limit theorem* it is a major justification for working with the empirical distribution function F_n. As we will see later, the bootstrap, an important approximation techniques for distributions, uses only information contained in the empirical distribution function to approximate the distribution of functionals of an iid sample.

1.1.3 The central limit theorem

The rate of convergence in the SLLN is described, in a sense, by the *central limit theorem* (CLT) which says that

$$\sqrt{n}\,\left(\overline{X}_n - EX\right) \xrightarrow{d} N(0,\sigma^2)\,,$$

if and only if $\sigma^2 = \mathrm{var}(X) < \infty$, where $N(\mu,\sigma^2)$ stands for the normal distribution with mean μ and variance σ^2. Here and in what follows, $Y_n \xrightarrow{d} Y$ denotes convergence in distribution:

$$P(Y_n \le x) \to P(Y \le x) \tag{1.2}$$

for all continuity points of the distribution F_Y of Y. Sometimes (as for $F_Y = N(0,1)$) we write $Y_n \xrightarrow{d} F_Y$ meaning $Y_n \xrightarrow{d} Y$. Notice that convergence in (1.2) is uniform if F_Y is continuous. In particular, the CLT implies

$$\Delta_n = \sup_{x\in\mathbb{R}} \left| P\left(\frac{\sqrt{n}}{\sigma}(\overline{X}_n - EX) \le x\right) - \Phi(x)\right| \to 0 \quad \text{as } n\to\infty, \tag{1.3}$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-y^2/2}\,dy\,, \quad x\in\mathbb{R}\,.$$

denotes the standard normal $N(0,1)$ distribution function.

Remark 1.1 It is sometimes important to know that the CLT implies convergence of some moments of the sample mean. Indeed, one can show (e.g. Gut [30]) that the CLT and the existence of a pth moment of X_1 together imply that

$$\sigma^{-1}\,E\left|\sqrt{n}\,\left(\overline{X}_n - EX\right)\right|^p \to E|Y|^p \tag{1.4}$$

provided Y has an $N(0,1)$ distribution. An analogous statement is true without the absolute values provided p is a positive integer. These properties follow as an exercise in *uniform integrability* of sequences of random variables.

1.1.4 The Berry-Esséen inequality

The rate of convergence in the CLT can be described in a more precise way. Recall the definition of Δ_n from (1.3). Then, if $E|X|^3 < \infty$ and $\sigma^2 = \mathrm{var}(X)$,

$$\Delta_n \le C_{\mathrm{BE}}\,\frac{1}{\sqrt{n}}\,\frac{E|X-EX|^3}{\sigma^3}\,,$$

where $C_{\mathrm{BE}} \le 0.7655...$ is a universal constant. Thus the rate of convergence in the CLT depends on the value of the ratio $E|X-EX|^3/\sigma^3$ and is of the order $1/\sqrt{n}$. In general, the order is not improvable as the example of an iid Bernoulli sequence (X_i) with $P(X_i = \pm 1) = 0.5$ shows. (Check this by considering the probability $P(\overline{X}_n = 0)$ which is of the order $1/\sqrt{n}$.) Indeed, in this case a bound of Δ_n from below of the order $1/\sqrt{n}$ holds. The poor approximation of the distribution of $\overline{X}_n$ through the normal distribution is due to the discreteness of the distribution of X_i. The sums $X_1 + \cdots + X_n$ are again discrete random variables, living on the integers. Only the normalization $1/\sqrt{n}$ in the CLT makes the gaps in the support of the distribution smaller and let them eventually

disappear when passing to the limit. As a rule of thumb, the approximation of the distribution of the sample mean is the better the smoother the distribution of the X_i. For example, the existence of a density can improve the rate in the CLT substantially.

Write

$$G_n(x) = P\left(\frac{\sqrt{n}}{\sigma}(\overline{X}_n - EX) \le x\right), \quad x \in \mathbb{R}.$$

The rate of the approximation in the CLT can be further improved if one does not consider uniform approximations. There exist local approximations of the absolute error in the CLT

$$G_n(x) - \Phi(x) \tag{1.5}$$

or local approximations of the relative error in the CLT

$$\frac{1 - G_n(x)}{1 - \Phi(x)} - 1. \tag{1.6}$$

If $x = x_n \to \infty$ the latter approximations are often called *large deviations*; if x_n grows too fast compared to n, (1.6) does not converge to zero.

1.1.5 Asymptotic expansions

One way to construct approximations to the absolute error in the CLT is to use *asymptotic* or *Edgeworth expansions*. The idea behind it can be seen from a Fourier approximation argument. Write f_n for the characteristic function of $\sqrt{n}(\overline{X}_n - EX)/\sigma$ and recall that G_n is its distribution function. Also write f for the characteristic function of X. Assume without loss of generality that $EX = 0$ and $\sigma^2 = 1$. A formal Taylor expansion yields

$$\log f(t) = \log E\mathrm{e}^{\,itX} = \sum_{j=2}^{\infty} \frac{\gamma_j}{j!}(it)^j, \quad t \in \mathbb{R}.$$

The coefficient γ_j is called the *cumulant of order j of F*. The cumulants can be expressed through the moments of X and vice versa. For example,

$$\gamma_1 = EX, \quad \gamma_2 = \sigma^2, \quad \gamma_3 = E(X - EX)^3.$$

By independence of the X_i we immediately conclude that

$$\begin{aligned} \log f_n(t) &= n \log f(t/\sqrt{n}) = \sum_{j=2}^{\infty} \frac{\gamma_j}{j!}(it)^j n^{(-j+2)/2} \\ &= -\frac{t^2}{2} + \sum_{j=1}^{\infty} \frac{\gamma_{j+2}}{(j+2)!}\,(it)^{j+2} n^{-j/2}. \end{aligned}$$

Another formal Taylor expansion yields

$$\exp\left\{\sum_{j=1}^{\infty} \frac{\gamma_{j+2}}{(j+2)!}\,u^{j+2} z^j\right\} = 1 + \sum_{\nu=1}^{\infty} P_\nu(u)\, z^\nu,$$

where P_ν is a polynomial with coefficients depending on the cumulants γ_j. Finally,

$$f_n(t) \quad = \quad \mathrm{e}^{\,-t^2/2} + \sum_{\nu=1}^{\infty} P_\nu(it)\,\mathrm{e}^{\,-t^2/2}\,n^{-\nu/2}\,.$$

Using a formal inverse Fourier transform for each term in the latter expansion, one can switch back to the probabilities. Notice that the characteristic function of Φ is just $\mathrm{e}^{\,-t^2/2}$ and the characteristic function of G_n is f_n. Hence

$$G_n(x) \quad = \quad \Phi(x) + \sum_{\nu=1}^{\infty} Q_\nu(x)\,n^{-\nu/2}\,,$$

where

$$\int_{\mathbb{R}} \mathrm{e}^{\,itx}\,dQ_\nu(x) = P_\nu(it)\,\mathrm{e}^{\,-t^2/2}\,.$$

Clearly, the above expansions have to be shown to make sense, but then one needs some additional assumptions. We give a particular result of this kind.

Theorem 1.2 *If F has a density, $E|X|^k < \infty$ for some integer $k \geq 3$ and $EX = 0$, then*

$$\sup_{x\in\mathbb{R}} (1+|x|^k) \left| G_n(x) - \Phi(x) - \sum_{\nu=1}^{k-2} Q_\nu(x)\,n^{-\nu/2} \right| = o(n^{-(k-2)/2})\,.$$

A direct comparison with the Berry-Esséen inequality shows a significant improvement in the approximation to G_n. However, notice that the expansion

$$\Phi(x) + \sum_{\nu=1}^{k-2} Q_\nu(x)\,n^{-\nu/2} \tag{1.7}$$

is not a distribution function any more. Indeed, it can even assume negative values. Nevertheless, the asymptotic expansions of G_n are much preciser approximations than a naive application of the CLT can provide, at least in the center of the distribution. In the far out tails the distribution of G_n is in general not well approximated by (1.7) and therefore different techniques are called for. It is in general very difficult to make statements about the precision of the approximation to $G_n(x)$ in a given x-area.

Theorem 1.2 can be extended to densities p_n of G_n. The results for densities are obtained by formal differentiation in (1.7).

In order to improve upon the approximation in the CLT asymptotic expansions with 2-3 terms are often sufficient. For completeness we give here the first three terms. Write

$$\varphi(x) = \frac{\mathrm{e}^{\,-0.5x^2}}{\sqrt{2\pi}}$$

for the standard normal density. Then (recall $\sigma^2 = 1$)

$$\begin{aligned}
Q_1(x) &= -\varphi(x)\,(x^2-1)\,\frac{EX^3}{6}\,,\\
Q_2(x) &= -\varphi(x)\left[(x^5-10x^3+15x)\,\frac{(EX^3)^2}{72}+(x^3-3x)\,\frac{EX^4-3}{24}\right]\\
Q_3(x) &= -\varphi(x)\left[(x^8-28x^6+210x^4-420x^2+105)\,\frac{(EX^3)^3}{1296}\right.\\
&\qquad \left.+(x^6-15x^4+45x^2-15)\,\frac{EX^3\,(EX^4-3)}{144}+(x^4-6x^2+3)\,\frac{EX^5-10EX^3}{120}\right].
\end{aligned}$$

The higher the order of the terms the more complicated the formulae. For general expressions of the expansions; see Petrov [43], Section VI, 1. The polynomials appearing in the Q_ν are the *Hermite polynomials*, given by the recursions

$$H_0(x) = 1 \quad \text{and} \quad -\varphi(x)\,H_m(x) = [\varphi(x)\,H_{m-1}(x)]'\,.$$

They appear very frequently in different disguises in higher-order expansions of distributions, random variables and stochastic processes which are somewhat related to the normal distribution or Gaussian processes (such as Brownian motion).

1.1.6 The continuous mapping theorem

In what follows, we will frequently make use of the continuous mapping theorem. Assume that Y_n, Y are random elements assuming values in a separable metric space (S,d) (we will usually consider a subset of the Euclidean space endowed with the usual distance). Then one says that (Y_n) converges in distribution to Y ($Y_n \xrightarrow{d} Y$) if $Ef(Y_n) \to Ef(Y)$ for all continuous bounded real-valued functions f on S. If h is a continuous function from (S,d) to another separable metric space (S',d') (we usually consider a mapping from $\mathbb{R}^d \to \mathbb{R}^k$) then $f(h)$ is again bounded and continuous, provided f is defined on (S',d') and bounded, continuous. Hence $Ef(h(Y_n)) \to Ef(h(Y))$ for all bounded continuous f.

Thus we have proved the continuous mapping theorem:

$$Y_n \xrightarrow{d} Y \quad \text{implies} \quad h(Y_n) \xrightarrow{d} h(Y)\,.$$

It is possible to show that convergence in distribution is nothing but convergence of the underlying distributions, also referred to as *weak convergence*. For example, $Y_n \xrightarrow{d} Y$ is equivalent to

$$F_{Y_n}(A) \to F_Y(A)\,,$$

for all Borel sets in S satisfying $F_Y(\partial A) = 0$.

An excellent treatment of weak convergence and convergence in distribution is Billingsley's classic of probability theory *Convergence of Probability Measures* [10].

Comments

Asymptotic expansions have been intensively studied in the 20th century at times when powerful computers were not available. They are still popular in asymptotic statistics when it comes to

showing that certain statistical tools have a better asymptotic precision than others. For example, asymptotic expansions have been used to show that the bootstrap is superior to the normal approximation when the CLT is applicable; see Hall [32] and Section 2.2.6. Asymptotic expansions for the distribution of non-linear functionals of iid data are prominent in the work of van Zwet; see for example [6] for a recent contribution.

Good introductions to asymptotic expansions and asymptotic theory in general are given in the books of Petrov [43, 44]. Extensions to the multivariate case can be found in Bhattacharya and Rao [7]. The case of asymptotic expansions for dependent data in abstract spaces was treated in a path-breaking paper by Götze and Hipp [28].

1.2 L_p-minimal metrics d_p

1.2.1 Definition

In what follows, it will prove useful to introduce metrics on sets of distributions on $\mathbb{R}$ (the restriction to $\mathbb{R}$ is irrelevant in most cases and extensions to more general spaces are possible; we assume it for convenience only). Write Γ_p for the distributions on $\mathbb{R}$ with finite pth moment for some $p \geq 1$. For $F, G \in \Gamma_p$, let H be any distribution on $\mathbb{R}^2$ with marginal distribution F in the first component and G in the second component and write $\mathcal{H}$ for the class of all such distributions H. Define

$$d_p(F,G) = \inf_{H \in \mathcal{H}} \left(\int_{\mathbb{R}^2} |x-y|^p \, H(dx,dy) \right)^{1/p} . \tag{1.8}$$

It can be shown (using a weak compactness argument) that the infimum in (1.8) is actually attained. For the purposes of illustration it is more convenient to think in terms of random variables with values in $\mathbb{R}$. Indeed, let (X,Y) be random vectors on the same probability space with marginal distributions $F_X = F$ and $F_Y = G$, $E|X|^p < \infty$, $E|Y|^p < \infty$ such that $F_{XY} = H$. Then

$$d_p(F,G) = \inf_{\text{all vectors } (X,Y) \text{ with marginals } F = F_X,\, G = F_Y} (E|X-Y|^p)^{1/p} .$$

Here we interpret X and Y as elements of the space L_p, i.e., they are equivalence classes of random variables in the sense that X and Z are identified if $X = Z$ a.s. With this interpretation it is often convenient to write $d_p(X,Y)$ instead of $d_p(F,G)$. This is clearly an abuse of notation since we do not consider a metric on the space L_p of random variables with finite pth moment but on the space Γ_p of distributions with finite pth moment. Nevertheless, convenience is often stronger than mathematical correctness, and we will follow the incorrect patterns, knowing precisely that we are doing something wrong.

Now notice that

$$d_p(X,Y) = 0 \text{ if and only if } F_X = F_Y \text{ if and only if } d_p(F_X, F_Y) = 0.$$

Moreover, it is clear from the definition that

$$d_p(X,Y) = d_p(Y,X), \text{ i.e., } d_p(F,G) = d_p(G,F).$$

(The latter relation also follows from an application of Fubini's theorem.) In order to make $d_p(F,G)$ a *probability metric on* Γ_p it thus suffices to show the triangle inequality: for distributions F, G, M on $\mathbb{R}$:

$$d_p(F,G) \leq d_p(F,M) + d_p(M,G) .$$

To prove this is a bit tricky since one has to find a vector (X, Y, Z) with marginals $F = F_X$, $G = F_Y$ and $M = F_Z$ which satisfies some particular conditional independence condition; see Bickel and Freedman [8] for such a construction. They also prove:

Lemma 1.3 *The random variables $X, X_1, X_2, \ldots$ (defined on the same probability space) satisfy $d_p(X_n, X) \to 0$ if and only if*

$$X_n \xrightarrow{d} X \text{ and } E|X_n|^p \to E|X|^p.$$

Remark 1.4 Do not forget that $d_p(X_n, X) \to 0$ means $d_p(F_n, F) \to 0$, where $F_n = F_{X_n}$ and $F = F_X$. Thus, in terms of weak convergence of probability distributions on $\mathbb{R}$, $d_p(F_n, F) \to 0$ if and only if

$$F_n \xrightarrow{w} F \text{ and } \int_{\mathbb{R}} |x|^p \, dF_n(x) \to \int_{\mathbb{R}} |x|^p \, dF(x)\,.$$

Here $\xrightarrow{w}$ stands for the weak convergence of probability measures.

Remark 1.5 In the literature (see Zolotarev [61], Rachev [46]) d_p is sometimes referred to as *L_p-minimal metric* (for obvious reasons) or as *Wasserstein metric* (see Bickel and Freedman [8]). Wasserstein considered the special case $p = 2$.

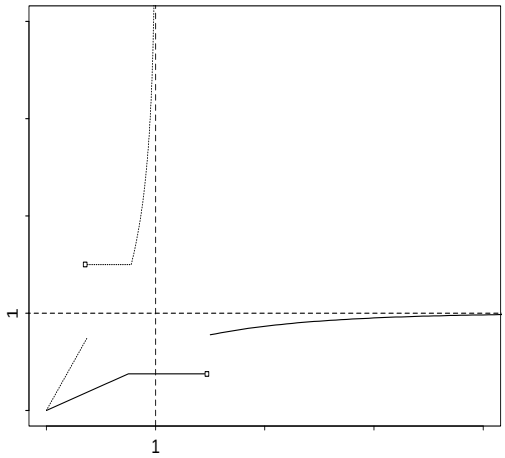

Figure 1.6 *A distribution function F on $[0, \infty)$ and its quantile function $F^{\leftarrow}$. In a sense, $F^{\leftarrow}$ is the mirror image of F with respect to the line $x = y$.*

Remark 1.7 The fact that the minimum in (1.8) is really achieved can be made more transparent. Let for any distribution function M,

$$M^{\leftarrow}(t) = \inf\{x : M(x) \geq t\}\,, \quad t \in (0, 1)\,,$$

be the corresponding *quantile function*. Clearly, if M^{-1} exists, $M^{\leftarrow} = M^{-1}$ and $M^{\leftarrow}$ has very much the same properties as an inverse function; see Resnick [50] for more properties of generalized

inverse functions. In particular, for $x \in \mathbb{R}$, $M^{\leftarrow}(t) \leq x$ if and only if $t \leq M(x)$. Therefore, for a uniform on $(0,1)$ random variable U,

$$\{M^{\leftarrow}(U) \leq x\} = \{U \leq M(x)\}$$

and hence

$$P(M^{\leftarrow}(U) \leq x) = P(U \leq M(x)) = M(x)\,, \quad x \in \mathbb{R}\,.$$

This means that $M^{\leftarrow}(U)$ has distribution M.

One can show (e.g. Rachev [46]) that

$$\begin{aligned}[d_p(F,G)]^p &= \int_0^1 |F^{\leftarrow}(t) - G^{\leftarrow}(t)|^p \, dt \\ &= E|F^{\leftarrow}(U) - G^{\leftarrow}(U)|^p\,,\end{aligned}$$

i.e., the minimum in (1.8) is achieved for the random vector $(F^{\leftarrow}(U), G^{\leftarrow}(U))$.

The case $d_1(F,G)$ is then particularly illuminating. Since the quantile function of a distribution function is nothing but the mirror image of the distribution function at the line $x = y$, the area between the graphs of F, G is exactly the same as between the quantile functions $F^{\leftarrow}, G^{\leftarrow}$. This means that

$$d_1(F,G) = \int_0^1 |F^{\leftarrow}(t) - G^{\leftarrow}(t)| \, dt = \int_{\mathbb{R}} |F(x) - G(x)| \, dx\,.$$

Thus, we do not want to forget: $d_p(X_n, X) \to 0$ holds if and only if

- (X_n) converges in distribution to X and
- The underlying pth moments converge: $E|X_n|^p \to E|X|^p$.

This is a convenient notion for many purposes in statistics where one is often not only interested in convergence of the distributions but also of the moments of certain statistics. d_p-convergence yields this for free.

1.2.2 Convergence of the empirical distribution function to the true distribution function in the d_p-distance

Let $X_1, X_2, \ldots$ be iid random variables with common distribution function F and finite pth moment for some $p \geq 1$. Write F_n for the empirical distribution function:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x\}}\,, \quad x \in \mathbb{R}\,.$$

We know from the Glivenko-Cantelli result (Section 1.1.2) that

$$F_n(x) \to F(x) \text{ uniformly for all } x \in \mathbb{R} \text{ with probability 1.} \tag{1.9}$$

Moreover, consider the pth moment with respect to F_n:

$$\int_{\mathbb{R}} |x|^p \, dF_n(x) = \frac{1}{n} \sum_{i=1}^n |X_i|^p\,.$$

Since $E|X_i|^p < \infty$ and the $|X_i|^p$ are iid, the SLLN gives

$$\int_{\mathbb{R}} |x|^p \, dF_n(x) \to E|X_1|^p = \int_{\mathbb{R}} |x|^p \, dF(x) \quad \text{a.s.} \tag{1.10}$$

If we imagine a random variable Y_n with distribution F_n and a random variable X with distribution F, it seems that we have shown that $Y_n \xrightarrow{d} X$ and $E|Y_n|^p \to E|X|^p$, and according to Lemma 1.3 the latter should be equivalent to $d_p(Y_n, X) \to 0$ and $d_p(F_n, F) \to 0$. However, we must be careful: relations (1.10) and (1.9) hold a.s. This means, $d_p(F_n, F) \to 0$ holds only on a set of probability 1. Indeed, the empirical distribution function is a random function, and therefore $d_p(F_n, F)$ is a random variable as well.

Moreover, since the Glivenko-Cantelli lemma holds for any distribution F and the SLLN (1.10) holds if and only if $E|X_1|^p < \infty$ we thus have shown:

- $d_p(F_n, F) \to 0$ a.s. if and only if $E|X_i|^p < \infty$.

1.2.3 Properties of the d_p-distance

In this section we collect some elementary properties of the d_p-metric.

Scaling property

Since $(E|cX - cY|^p)^{1/p} = |c| \, (E|X - Y|^p)^{1/p}$ it is immediate from the definition of d_p that

$$d_p(cX, cY) = |c| \, d_p(X, Y) \,, \quad c \in \mathbb{R} \,.$$

Bounds for sums

In what follows, we will have to compare the distributions of sums of independent random variables in the d_p-metric for $p \geq 1$. So let (X_i) and (Y_i) be two sequences of independent random variables on the same probability space, the dependence between (X_i) and (Y_i) being unspecified. It will be convenient to use the L_p-norm $|X|_p = (E|X|^p)^{1/p}$. From Minkowski's inequality we know that

$$\left| \sum_{i=1}^n (X_i - Y_i) \right|_p \leq \sum_{i=1}^n |X_i - Y_i|_p \,.$$

Now, first take the infimum with respect to all joint distributions of $(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i)$ on the left-hand side, resulting in $d_p(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i)$, and then take the infimum with respect to all joint distributions of the (X_i, Y_i) on the right-hand side, resulting in $\sum_{i=1}^n d_p(X_i, Y_i)$. Thus we have proved

$$d_p\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \right) \leq \sum_{i=1}^n d_p(X_i, Y_i) \,.$$

In particular, for iid (X_i, Y_i),

$$d_p\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \right) \leq \, n \, d_p(X_1, Y_1) \,. \tag{1.11}$$

The latter bound can be substantially improved for $p = 2$ provided $EX_i = EY_i$ for all i. Assume for the moment that the pairs (X_i, Y_i) are mutually independent. Then the $X_i - Y_i$ are mutually independent and have mean zero. Hence, using these facts,

$$\left| \sum_{i=1}^{n} (X_i - Y_i) \right|_2^2 = \sum_{i=1}^{n} |X_i - Y_i|_2^2 \,.$$

Now, first take the infimum with respect to all joint distributions of $(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i)$ on the left-hand side and then with respect to all joint distributions of the (X_i, Y_i) on the right-hand side, resulting in

$$\left[d_2 \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i \right) \right]^2 \leq \sum_{i=1}^{n} \left[d_2(X_i, Y_i) \right]^2 .$$

For iid (X_i) and (Y_i) we have

$$d_2 \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i \right) \leq \sqrt{n}\, d_2(X_1, Y_1) \,. \tag{1.12}$$

Compared to (1.11) we gained an improvement of the order $\sqrt{n}$ on the right-hand side. Using the scaling property of d_p, we can rewrite (1.12) as follows:

$$\sqrt{n}\, d_2(\overline{X}_n, \overline{Y}_n) \leq d_2(X_1, Y_1) \,, \tag{1.13}$$

where, as usual, $\overline{X}_n$ and $\overline{Y}_n$ stand for the sample means of the corresponding X- and Y-sequences. The left-hand side can be written as

$$\inf \left(E \left| \sqrt{n}(\overline{X}_n - \overline{Y}_n) \right|^2 \right)^{1/2} ,$$

where the infimum is taken over all joint distributions of $(\overline{X}_n, \overline{Y}_n)$. If we assume the Y_i to be iid $N(0,1)$ random variables, then $\sqrt{n}\overline{Y}_n$ is $N(0,1)$ distributed, and so it looks as if we could prove the CLT in the d_2-metric by using (1.13). However, the right-hand side in (1.13) does not depend on n and therefore the CLT is not straightforward.

2 The bootstrap

2.1 Efron's naive bootstrap

In 1979 the very influential paper by Bradley Efron [22] *Bootstrap methods: another look at the jackknife* was published. Since then hundreds of research papers and at least a dozen of books have been written about the "bootstrap", and Efron himself claimed that this method could solve all problems of applied statistics.

Pull yourself up by your own bootstraps is a US-American saying which actually says something which is impossible. In the novel *Baron Münchhausen* of Gottfried August Bürger the Lügenbaron got himself and his horse out of a swamp by pulling his own pony-tail, and therefore some Germans call the method Münchhausen method. Anyway, these two references seem to indicate that there is something unkosher with the method, but we will see that Efron had a simple idea that can be fruitfully used as a concept in many statistical applications. The idea became very popular in the

80s and 90s, mostly because its effects could be visualized using the power of modern computers, but also its pitfalls could be seen, and some of the pitfalls could be repaired. Efron himself made the bootstrap idea very popular in a series of papers and through the books Efron [23] and Efron and Tibshirani [24].

Thus,

What is the "bootstrap" about?

The main idea is a very simple one. Suppose you have an iid sample with common distribution function F_θ:

$$X_1, \ldots, X_n \sim F_\theta\,, \quad \theta \in \Theta\,,$$

where $\Theta \subset \mathbb{R}^d$, say. Let

$$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$$

be an estimator of θ. In order to say something reasonable about the quality of the estimator $\widehat{\theta}_n$ one needs to measure the spread of the distribution of $\widehat{\theta}_n$ around θ. For example, it would be helpful to know something about $\mathrm{var}(\widehat{\theta}_n)$. However, we are given only one sample, and so we can calculate only one value $\widehat{\theta}_n$ which cannot be used to determine the variance of $\widehat{\theta}_n$. One way out of this situation would be to use asymptotic theory as for example given by the CLT: given that the CLT $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ holds and that one can estimate σ by some $\widehat{\sigma}_n$, $[-1.96\widehat{\sigma}_n/\sqrt{n}; 1.96\widehat{\sigma}_n/\sqrt{n}]$ could be taken as asymptotic 95% confidence interval for θ.

We learned in Section 1.1.4 that the rate in the CLT for the sample mean can be rather slow, in particular when one deals with discrete distributions, and that the asymptotic normal theory has to be used with care for "small" sample sizes. Efron [22] promised in particular that his method would work also for small sample sizes, and actually he was mostly interested in this case because, otherwise, the asymptotic theory applies.

Efron's bootstrap is based on the Glivenko–Cantelli idea (Section 1.1.2) that the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}}\,, \quad x \in \mathbb{R}\,,$$

is a consistent estimator of the underlying distribution function F_θ:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_\theta(x)| \xrightarrow{\text{a.s.}} 0\,.$$

Therefore the distribution function F_θ is reasonably approximated by F_n. Denote the values of a concrete sample by

$$x_1, \ldots, x_n\,,$$

i.e., we interpret them as realizations of the iid random variables $X_1, \ldots, X_n$:

$$x_1 = X_1(\omega)\,, \ldots, x_n = X_n(\omega)\,.$$

The bootstrap idea consists of everywhere replacing the random variables X_i with distribution function F_θ in the statistic $\widehat{\theta}_n(X_1, \ldots, X_n)$ by random variables X_i^* with distribution function F_n

constructed from $x_1, \ldots, x_n$. In contrast to the one sample $x_1, \ldots, x_n$, we can generate as many independent *bootstrap samples*

$$X_{n1}^*(i), \ldots, X_{nn}^*(i),$$

with common distribution function F_n as we like. This is at least in principle true: by using the computer we can draw independently with replacement from the set of numbers $x_1, \ldots, x_n$ as often as we like. This procedure is also called a (particular) *resampling method*. In what follows, for the ease of presentation, we suppress the dependence of the $X_{nj}^*(i)$ on the sample size n, i.e., we simply write $X_j^*(i)$, and we often also suppress the index i and write X_j^* when it is clear what it is intended to say.

Thus, suppose you have drawn $n \times B$ times independently with replacement from the empirical distribution function F_n, where B is a very large number. This means we have the array of iid random variables

$$X_1^*(1), \ldots, X_n^*(1), \ldots, X_1^*(B), \ldots, X_n^*(B), \tag{2.1}$$

with common distribution function F_n *given the values* $x_1, \ldots, x_n$ *of the iid sample* $X_1, \ldots, X_n$.

A next step would be to plug the samples (2.1) in the estimator $\widehat{\theta}_n$, i.e., with the help of the computer we calculate the values

$$\widehat{\theta}_n^*(1) = \widehat{\theta}_n(X_1^*(1), \ldots, X_n^*(1)), \ldots, \widehat{\theta}_n^*(B) = \widehat{\theta}_n(X_1^*(B), \ldots, X_n^*(B))$$

Now, at least in principle, we can determine the distribution of

$$\widehat{\theta}_n^* = \widehat{\theta}_n(X_1^*, \ldots, X_n^*)$$

and its characteristics such as the expectation, variance, moments, quantiles etc. However, this may create some trouble, as the following example shows.

Example 2.1 (Bootstrap sample mean)
Consider the class of distributions with finite mean and take $\theta = EX_1$. The sample mean $\overline{X}_n$ is an unbiased estimator of θ. The bootstrap version of the sample mean is the *bootstrap sample mean*

$$\overline{X}_n^* = \frac{1}{n} \sum_{i=1}^{n} X_i^*,$$

where $X_1^*, \ldots, X_n^*$ is an iid sample with distribution F_n for the given values $x_1, \ldots, x_n$ of the iid sample $X_1, \ldots, X_n$ with distribution $F = F_\theta$. Since the X_i^* have a discrete distribution with mass n^{-1} at the atoms x_j of the jth element in the sample (with this interpretation all x_j are distinct even if some of the values x_j coincide). In this sense, there are n^n distinct realizations $x_{i_1}, \ldots, x_{i_n}$ of the sample $X_1^*, \ldots, X_n^*$, and so it might become tedious to calculate the distribution of the bootstrap sample mean or some of its distributional characteristics. Indeed, for the bootstrap sample mean we would have to calculate all values

$$\frac{1}{n} \sum_{j=1}^{n} x_{i_j}, \tag{2.2}$$

for the n^n realizations $x_{i_1}, \ldots, x_{i_n}$. The distribution of these values would give us the exact distribution of $\overline{X}_n^*$. The mean of $\overline{X}_n^*$ would be given by the average of the n^n values (2.2), the moments could be calculated accordingly, etc.

But this approach is not appealing from a practical point of view.

Before we give Efron's idea of how to circumvent the above calculation problem, it is worthwhile mentioning that the moments of the bootstrap sample mean can be easily calculated. Indeed, X_i^* has distribution function F_n, and we know how to calculate the expectation and variance of a sample mean. Let E^* and var^* denote the expectation and variance with respect to the measure induced by the empirical distribution function. Then,

$$\begin{aligned} E^*\left(\overline{X}_n^*\right) &= \frac{1}{n}\sum_{i=1}^n E^*(X_i^*) = E^*(X_1^*) = \overline{x}_n\,, \\ \mathrm{var}^*(\overline{X}_n^*) &= \frac{1}{n}\,\mathrm{var}^*(X_1^*) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n (x_i - E^*(X_1^*))^2\right] = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n (x_i-\overline{x}_n)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n (x_i-\overline{x})^2\,. \end{aligned}$$

Thus the mean of the bootstrap sample mean is the sample mean and the variance of the bootstrap sample mean, up to a small change of normalization, is the usual sample variance. We mention that the trick of recalling that X_i^* has distribution F_n is often useful when it comes to the calculation of moments for "regular statistics". □

Remark 2.2 Above we used the notation E^* and var^*. Obviously, the latter can be reduced to E^*. In order to make things precise, E^* is always understood as the expectation with respect to the *conditional* probability measure

$$P^*(\cdot) = P(\cdot \mid X_1 = x_1, X_2 = x_2, \ldots)\,.$$

If you have difficulties with this infinite condition in the expectation, it is convenient (and often correct) to imagine P^* as the probability measure associated with a finite segment $(X_1,\ldots,X_n)$. For example, since $\widehat{\theta}_n^* = \widehat{\theta}_n(X_1^*,\ldots,X_n^*)$ and X_i^* depends on $X_1,\ldots,X_n$, it is clear that

$$P^*(\theta_n^* \in A) = P(\theta_n^* \in A \mid X_1 = x_1, \ldots, X_n = x_n)\,.$$

In the previous example we saw that it is in general not feasible to calculate the distribution or distributional characteristics of the bootstrap statistics $\widehat{\theta}_n^*$ explicitly. Efron proposed to circumvent these problems simply by approximating the distribution of $\widehat{\theta}_n^* = \widehat{\theta}_n(X_1^*,\ldots,X_n^*)$ by producing a large number B, the *bootstrap sample size*, of independent realizations

$$\widehat{\theta}_n^*(1) = \widehat{\theta}_n^*(X_1^*(1),\ldots,X_n^*(1))\,,\ldots,\widehat{\theta}_n^*(B) = \widehat{\theta}_n^*(X_1^*(B),\ldots,X_n^*(B))\,.$$

Then use the SLLN and Glivenko–Cantelli. For example, the distribution function of $\widehat{\theta}_n^*$ is approximated according to

$$\sup_{x\in\mathbb{R}}\left|\frac{1}{B}\sum_{i=1}^B I_{\{\widehat{\theta}_n^*(i)\le x\}} - P^*(\widehat{\theta}_n^*\le x)\right| \overset{\text{a.s.}}{\to} 0\,, \quad B\to\infty\,.$$

Here we have to mention that a.s. convergence refers to the probability measure P^* explained in Remark 2.2. This means in particular that this convergence depends on the realization $x_1,\ldots,x_n$ of the sample. Note that one can in particular derive confidence bands for the distribution of $\widehat{\theta}_n^*$ by

choosing the corresponding quantiles from the distribution of the empirical distribution function of the values $\widehat{\theta}_n^*(i)$, $i = 1, \ldots, B$. The median of the latter distribution can be considered as an estimator of the median of the distribution of $\widehat{\theta}_n^*$.

Similarly, in P^*-probability,

$$\frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_n^*(i) \overset{\text{a.s.}}{\to} E^*(\widehat{\theta}_n^*) \quad \text{and} \quad \frac{1}{B}\sum_{i=1}^{B}[\widehat{\theta}_n^*(i)]^2 \overset{\text{a.s.}}{\to} E^*[\widehat{\theta}_n^*)]^2\,,$$

and also

$$\frac{1}{B-1}\sum_{i=1}^{B}\left([\widehat{\theta}_n^*(i)]^2 - \left[\frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_n^*(i)\right]^2\right) \overset{\text{a.s.}}{\to} \operatorname{var}^*(\widehat{\theta}_n^*)\,. \tag{2.3}$$

Since we can choose B very large by using enough computer power, the approximation of the bootstrap distribution and of the bootstrap moments does not present any practical problems.

We summarize *Efron's naive bootstrap method* as follows:

- Draw a large number B of independent iid samples of size n from the empirical distribution function F_n of the concrete sample $X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$:

  $$X_1^*(1), \ldots, X_n^*(1), \ldots, X_1^*(B), \ldots, X_n^*(B)\,.$$

  This means you draw with replacement any of the values $x_1, \ldots, x_n$ with equal probability $1/n$.

- Calculate the corresponding estimators from the B bootstrap samples:

  $$\widehat{\theta}_n^*(1) = \widehat{\theta}_n(X_1^*(1), \ldots, X_n^*(1)), \ldots, \widehat{\theta}_n^*(B) = \widehat{\theta}_n(X_1^*(B), \ldots, X_n^*(B))\,.$$

- Approximate moments and probabilities of the quantity $\widehat{\theta}_n^* = \widehat{\theta}(X_1^*, \ldots, X_n^*)$ by using the SLLN applied to $\widehat{\theta}_n^*(1), \ldots, \widehat{\theta}_n^*(B)$. For example, (2.3) is an approximation to the bootstrap variance.

In sum,

What have we gained from Efron's bootstrap?

The basic idea is to replace in the estimator $\widehat{\theta}_n(X_1, \ldots, X_n)$ of θ the X_i-values (of which we have only one realization $x_1, \ldots, x_n$) by random variables $X_1^*, \ldots, X_n^*$ with distribution F_n. Since we can (in principle) generate arbitrarily many values $\widehat{\theta}_n(X_1^*(i), \ldots, X_n^*(i))$ we are (in principle) able to determine the distribution of $\widehat{\theta}_n(X_1^*, \ldots, X_n^*)$ arbitrarily accurately as well as distributional characteristics such as moments of these quantities through the SLLN. Thus the bootstrap offers an alternative to the asymptotic theory based on CLT-related methods.

2.2 A discussion of Efron's bootstrap

There are various problems related to the bootstrap which we want to address now.

2.2.1 How large has B to be chosen?

This is a question which depends on the kind of problem one wants to solve, and on the estimator $\widehat{\theta}_n$. Efron and Tibshirani [24] give various rules of thumb "based on personal experience" how large to choose the bootstrap sample size. For "simple" statistics such as the sample variance or standard deviation they observed that $B = 200$ can be sufficient. For the complicated problem of determining high or low order quantiles such as the 98% or 99% quantile of the distribution of a "complicated" statistic their experience shows that a bootstrap sample size of several ten thousand was not unusual in order to get sufficiently high accuracy.

In general, only experimenting helps. If the calculation of the bootstrapped statistic is not costly, one should take B as big as possible or at least as big that the bootstrap variance does not change significantly, when the bootstrap sample size further increases.

In what follows, we do not consider the problem of choosing B anymore. In theory we will always assume that we can calculate the distribution of $\widehat{\theta}_n^*$ *exactly.*

2.2.2 Does the bootstrap work?

This is a naive, but interesting question and not easy to answer. When we consider an estimator $\widehat{\theta}_n$ of a parameter θ, we are usually interested in the question about the quality of this estimator. The construction of the bootstrap statistic $\widehat{\theta}_n(X_1^*, \ldots, X_n^*)$ is born out of the idea that the empirical distribution function F_n is close to the distribution function F of the data and therefore it seems reasonable that the distributions of $\widehat{\theta}_n(X_1^*, \ldots, X_n^*)$ and $\widehat{\theta}_n(X_1, \ldots, X_n)$ are close as well. However, the distance between F and F_n, when measured for example in the uniform distance $\sup_x |F_n(x) - F(x)|$, is close only if n is large. Moreover, our experience from calculus tells us that we are dealing with a complicated continuity problem. Indeed, the closeness of F_n and F shall imply that the distributions of $\widehat{\theta}_n^*$ and $\widehat{\theta}_n$ are close in a sense. This is by no means obvious even for a simple object such as the bootstrap sample mean, and therefore further investigations are needed.

One possible answer to the above problem is the following: for large n, the bootstrap statistic $\widehat{\theta}_n^*$ should have the same asymptotic properties as $\widehat{\theta}_n$ itself. For example, if $\widehat{\theta}_n$ is asymptotically normal so should $\widehat{\theta}_n^*$ be. Otherwise, we may conclude that $\widehat{\theta}_n$ and $\widehat{\theta}_n^*$ behave in totally different ways. Or to say it more dramatically, $\widehat{\theta}_n^*$ would not estimate θ at all.

The idea that the distributions of $\widehat{\theta}_n$ and $\widehat{\theta}_n^*$ should be close for large n is referred to as *bootstrap consistency* or as to the fact that *the bootstrap works.* It is often convenient to use certain metrics to describe the closeness of distributions quantitatively. For example, the L_p-minimal probability metric d_p (Section 1.2) is one such tool.

Notice that the question as to whether the bootstrap works is as to whether a plug-in argument in convergence works. Indeed, on the one hand, $F_n \to F$ in some sense as $n \to \infty$. On the other hand, the distribution of the plug-in estimator $\widehat{\theta}_n^*$ (whose distribution is determined through F_n) and the distribution of $\widehat{\theta}_n$ (whose distribution is determined through F) should be close. Thus, both the distribution F_n of the plugged-in random variables X_i^* *and* the plug-in statistic $\widehat{\theta}_n$ change in dependence on n.

The bootstrap can be shown to work by theoretical means for smooth functionals of the data. The general theory would require to study some complicated statistical differentiability properties; see for example Mammen [39]. In Section 2.2.3 we will be able to show this for the bootstrap sample mean. As a matter of fact, the bootstrap is too general a method and therefore it might not come as a surprise that the bootstrap fails in various situations. We will consider pitfalls of the bootstrap in Section 2.2.4. In Section 2.2.5 we will consider a method which sometimes works in order to overcome pitfalls.

2.2.3 An example: the bootstrap for the sample mean works

In this section we make heavily use of the L_2-minimal metric d_2 introduced in Section 1.2. It is therefore useful to recall its definition and properties.

Let $X_1, \ldots, X_n$ be iid random variables with distribution function F and unit variance. By $X_1(\omega) = x_1, X_2(\omega) = x_2, \ldots$ denote a particular realization of the sequence (X_i). Recall that F_n stands for the empirical distribution function of the sample. Let $X_1^*, \ldots, X_n^*$ be iid random variables with distribution function F_n. They constitute one *bootstrap sample* for the given sample $X_1 = x_1, \ldots, X_n = x_n$. Since F_n is a distribution with finite support, it is trivial that the second moment of X_i^* with respect to F_n is finite. Recall that

$$E^*(X_j^*) = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}_n \,.$$

Consider an iid sequence $(\widetilde{X}_i)$ with the same distribution as (X_i). For fixed n, both the sequences $\sqrt{n}(\widetilde{X}_i - EX_i)$, $i = 1, \ldots, n$, and $\sqrt{n}(X_i^* - \overline{x}_n)$, $i = 1, \ldots, n$, consist of iid zero-mean random variables. An application of (1.13) and the triangle inequality yield

$$\begin{aligned} d_2\left(\sqrt{n}(\overline{X}_n^* - \overline{x}_n), \sqrt{n}(\overline{\widetilde{X}}_n - EX_1)\right) &\leq d_2(X_1^* - \overline{x}_n, \widetilde{X}_1 - EX_1) \\ &\leq d_2(X_1^*, \widetilde{X}_1) + d_2(\overline{x}_n, EX_1)\,. \end{aligned} \tag{2.4}$$

Since the x_i are realizations of a random sequence, i.e., we fix one particular ω, we can treat them as constants. Hence

$$d_2(\overline{x}_n, EX_1) = |\overline{x}_n - EX_1| \to 0 \quad P - \text{a.s.} \tag{2.5}$$

The convergence to zero for P-almost every realization $(X_n(\omega))$ follows from the SLLN. Since $\widetilde{X}_1 \stackrel{d}{=} X_1$ and X_1^* has distribution F_n,

$$d_2(X_1^*, \widetilde{X}_1) = d_2(F_n, F) \to 0 \quad P - \text{a.s.} \tag{2.6}$$

according to Section 1.2.2. Now combine (2.4)–(2.6) to obtain

$$d_2\left(\sqrt{n}(\overline{X_n^*} - \overline{x}_n), \sqrt{n}(\overline{\widetilde{X}}_n - EX_1)\right) \to 0 \quad \text{a.s.}$$

The "a.s." on the right-hand side refers to the fact that we live on a realization $x_1 = X_1(\omega), x_2 = X_2(\omega), \ldots$, of the iid sequence (X_i). This means that we have proved that the normalized bootstrap sample mean $\sqrt{n}(\overline{X}_n^* - \overline{x}_n)$ and the sample mean $\sqrt{n}(\overline{\widetilde{X}}_n - EX_1)$ are close in distribution in the d_2-sense. Hence,

The bootstrap works for the sample mean.

In particular, since by the CLT $\sqrt{n}(\overline{\widetilde{X}}_n - EX_1) \stackrel{d}{\to} Y$ for an $N(0,1)$-distributed random variable Y and also $E[\sqrt{n}(\overline{\widetilde{X}}_n - EX_1)]^2 \to EY^2$ (see Remark 1.1),

$$d_2\left(\sqrt{n}(\overline{X_n^*} - \overline{x}_n), Y\right) \leq d_2\left(\sqrt{n}(\overline{X_n^*} - \overline{x}_n), \sqrt{n}(\overline{\widetilde{X}}_n - EX_1)\right) + d_2\left(\sqrt{n}(\overline{\widetilde{X}}_n - EX_1), Y\right) \to 0 \, P - \text{a.s.}$$

This means that the bootstrapped sample mean satisfies the CLT in the d_2-metric a.s. But we know from Lemma 1.3 that the latter convergence implies convergence in distribution and convergence of the second moments. In particular, the bootstrap sample mean satisfies the CLT along almost every sample path $x_1 = X_1(\omega), x_2 = X_2(\omega), \ldots$.

2.2.4 Pitfalls of the bootstrap method: degenerate U-statistics and extremes

Early on it has been observed that the bootstrap may fail even in some seemingly simple situations. Several such examples were given in Bickel and Freedman [8] which was one of the first papers that dealt with the bootstrap in a rigorous way. In what follows we give two of their counter-examples.

Example 2.3 (The bootstrap fails for non-linear statistics)
Bickel and Freedman [8] showed the failure of the bootstrap for certain U-statistics (of order 2), i.e., statistics of the form

$$H_n = H_n(X_1, \dots, X_n) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j)\,,$$

where $h(x, y) = h(y, x)$ is a symmetric real-valued kernel function. The "U" refers to the fact that H_n is an Unbiased estimator of the expectation $Eh(X_1, X_2)$, provided the latter is well defined. U-statistics were introduced by Hoeffding [34] as a natural generalization of the sample mean to functions of several variables. Various statistics important for estimating and testing hypotheses can be written as U-statistics or have a structure which is close to a U-statistic. They include the sample variance, Gini's mean difference, the K-function from spatial statistics, the empirical correlation integral. U-statistics are the simplest examples of non-linear functionals of a sample and their study was the basis for the understanding of more complicated non-linear objects. We refer to the books of Serfling [52] and van der Vaart [58] for more information.

We do not want to consider the most complicated case but focus on one very simple example. Consider the kernel

$$h(x, y) = xy\,.$$

The resulting U-statistic is of the form

$$H_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} X_i X_j = \frac{1}{n(n-1)} \left[\left(\sum_{i=1}^n X_i \right)^2 - \sum_{i=1}^n X_i^2 \right] .$$

Assume that $EX_i = 0$ and $\mathrm{var}(X_1) = 1 < \infty$. Clearly, $EH_n = 0$, and the SLLN implies that $H_n \stackrel{\text{a.s.}}{\to} Eh(X_1, X_2) = 0$, i.e., H_n is a consistent estimator of the mean. Moreover, (H_n) converges in distribution to a somewhat unusual limit: we know by the CLT and the SLLN that

$$\frac{1}{\sqrt{n-1}} \sum_{i=1}^n X_i \quad \stackrel{d}{\to} \quad Y \sim N(0, 1)\,,$$

$$\frac{1}{n-1} \sum_{i=1}^n X_i^2 \quad \stackrel{\text{a.s.}}{\to} \quad 1\,.$$

Those two combined with the continuous mapping theorem (Section 1.1.6) give

$$n\, H_n = \left(\frac{1}{\sqrt{n-1}} \sum_{i=1}^n X_i \right)^2 - \frac{1}{n-1} \sum_{i=1}^n X_i^2 \ \stackrel{d}{\to}\ Y^2 - 1\,.$$

Thus the limit is a centered χ^2-distribution. As a matter of fact, this kind of limit is typical for so-called *degenerate U-statistics*, i.e., U-statistics with kernel satisfying $E[h(X_1, X_2)|X_2] = \text{const}$ a.s. See Dynkin and Mandelbaum [21] or Lee [38] for more information.

Now consider the naive bootstrap version of nH_n:

$$n\,H_n^* = \frac{2}{n-1}\sum_{1\le i<j\le n} X_i^* X_j^* = \left(\frac{1}{\sqrt{n-1}}\sum_{i=1}^n X_i^*\right)^2 - \frac{1}{n-1}\sum_{i=1}^n [X_i^*]^2\,. \tag{2.7}$$

It can be shown (Athreya [3]) that

$$\frac{1}{n-1}\sum_{i=1}^n [X_i^*]^2 \stackrel{\text{a.s.}}{\to} 1\,,$$

for almost every ω. As to the first term, consider

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i^* = \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i^* - \overline{X}_n)\right] + \sqrt{n}\,\overline{X}_n\,. \tag{2.8}$$

By the bootstrap CLT of Section 2.2.3, the sequence in square brackets converges in P^*-distribution to an $N(0,1)$ distribution for almost every realization $(X_i(\omega))$. Since we live in the bootstrap world on one realization $(X_i(\omega))$, it remains to consider the behavior of $\sqrt{n}\,\overline{X}_n(\omega)$. Since $(\sqrt{n}\,\overline{X}_n)$ satisfies the CLT in P-probability, it does not converge with P-probability 1. Actually, since $EX_1 = 0$, $(\sqrt{n}\,\overline{X}_n)$ crosses 0 infinitely often and contains subsequencees converging to $\pm\infty$. This follows from the general theory of random walks (e.g. Spitzer [55]) or the law of the iterated logarithm. An application of the continuous mapping theorem (Section 1.1.6) tells one that (2.7) does not converge in distribution in P^*-probability for almost every realization $(X_i(\omega))$. Thus,

The bootstrap does not work for this degenerate U-statistic.

This fact remains valid for general degenerate U-statistics as mentioned in Bretagnolle [11] and Bickel and Freedman [8].

This problem was solved several years later by Arcones and Giné [2]; see also Dehling and Mikosch [18] for a solution by using d_p-metrics. The key to its solution is quite simple. Indeed, from the argument after (2.8) it might have become clear that the centering of the X_i^* represents a major difficulty. Indeed, if one recalls that, because of $EX_i = 0$, we could also write

$$n\,H_n = \frac{2}{n-1}\sum_{1\le i<j\le n} (X_i - EX_i)\,(X_j - EX_j)$$

It is natural to define the following bootstrap version of H_n:

$$\begin{aligned} n\,\widetilde{H}_n^* &= \frac{2}{n-1}\sum_{1\le i<j\le n} (X_i^* - E^*(X_i^*))\,(X_j^* - E^*(X_j^*)) \\ &= \frac{2}{n-1}\sum_{1\le i<j\le n} (X_i^* - \overline{X}_n)\,(X_j^* - \overline{X}_n) \\ &= \left(\frac{1}{\sqrt{n-1}}\sum_{i=1}^n (X_i^* - \overline{X}_n)\right)^2 - \frac{1}{n-1}\sum_{i=1}^n (X_i^* - \overline{X}_n)^2\,. \end{aligned}$$

Using similar arguments as above, it can be shown that for P-a.e. ω,

$$P^*(n\,\widetilde{H}_n^* \le x) = P(n\,\widetilde{H}_n^* \le x \mid X_1, X_2, \ldots) \to P(Y^2 - 1 \le x)\,, \quad x \in \mathbb{R}\,,$$

for some $N(0,1)$ random variable Y, i.e., both (nH_n) and $(n\widetilde{H}_n^*)$ have the same limit distribution and, hence, the bootstrap works for the modified bootstrap version $\widetilde{H}_n^*$, whereas it does not for H_n^*. □

Example 2.4 (The bootstrap fails for extremes)
Let $X_1, \ldots, X_n$ be an iid sample with uniform distribution on $(0, \theta)$ for some unknown $\theta > 0$. The maximum likelihood estimator for θ is the maximum M_n of the sample. It is not difficult to see that

$$P(n\,(\theta - M_n)/\theta \le x) \to 1 - \mathrm{e}^{\,-x}\,, \quad x > 0\,,$$

i.e., M_n has an asymptotic exponential distribution. In order to use the latter relation for the construction of asymptotic confidence bands one would have to know the unknown θ. If we think of θ as the upper end point of the support of F it is natural to bootstrap $n\,(\theta - M_n)/\theta$ by $n\,(M_n^* - M_n)/M_n^*$, where M_n^* is the maximum of the bootstrap sample $X_1^*, \ldots, X_n^*$. However, this approach does not work since

$$\begin{aligned} P^*(n\,(M_n - M_n^*) = 0) &= P^*(M_n^* = M_n) = 1 - P^*(M_n^* < M_n) \\ &= 1 - [P^*(X_1^* < M_n)]^n = 1 - [1 - n^{-1}]^n \to 1 - \mathrm{e}^{\,-1} \approx 0.63\,. \end{aligned} \tag{2.9}$$

Thus, the bootstrapped maximum does not converge to an exponential distribution in P^*-probability, hence it has an asymptotic behavior different from M_n and therefore the bootstrap fails. □

Bickel and Freedman [8] also report that the bootstrap fails for the spacings of an iid sample. A method how to overcome this pitfall is discussed in Section 2.2.5.

2.2.5 Subsampling

Another pitfall of the bootstrap was mentioned by Athreya [4] and Hall [31] for the sample mean of an iid sample when the variance of the summands is infinite. This case is very much related to the upper order statistics in a sample, as the following discussion will show.

When $\mathrm{var}(X_1) = \infty$ typically so-called *stable distributions* occur for the (normalized and centered) sample mean. The stable distributions are rather unfamiliar in statistics (with the exception of the Cauchy distribution which occurs as the distribution of the ratio of two independent mean zero Gaussian random variables) since their densities cannot be expressed in a simple way through elementary functions. Notice that for an iid finite variance sequence (X_i) the strong law of large numbers implies that

$$\frac{\max_{i=1,\ldots,n} |X_i|}{\left(\sum_{i=1}^n (X_i - \overline{X}_n)^2|\right)^{1/2}} \to 0 \quad P - \text{a.s.}$$

This means that the largest value in the sample is negligible compared to the sample standard deviation. For an infinite variance sequence this relation does not remain valid any more (O'Brien [41]), i.e., the maximum term then makes an essential contribution to the sample standard deviation and cannot be neglected any more. It can be shown to be of the same order as the (centered) sums $X_1 + \cdots + X_n$; see Embrechts et al. [25], Section 8.2.4 for details. As in Example 2.4 for the maximum of an iid sample the Hall example for the sample mean shows that the bootstrap and the extremes in a sample do not really fit. When extreme values are involved in the analysis, Example 2.4 shows that the maximum value M_n of a sample is drawn too often compared to the other values in the sample. It is not difficult to see that relation (2.9) remains valid for *any* distribution F with a density, and it can be modified for other distributions as well. Hence, if we were to draw a large number B of bootstrap samples from the empirical distribution, in about 63% of the B cases the bootstrap maximum $M_n^*(i)$ would be equal to the maximum M_n of the original

sample. This shows that there is no real variability for the bootstrap distribution of the maxima and, more generally, for the upper order statistics in a sample. Therefore the naive bootstrap distribution of the bootstrapped upper order statistics cannot be taken as an approximation to the distributions of the upper order statistics.

Another glance at (2.9) also gives us some hint how we could modify the bootstrap procedure in order to overcome this extreme value pitfall of the naive bootstrap. Indeed, the too large proportion of appearances of the value M_n is due to the choice of the sample size in the naive bootstrap! Indeed, so far we argued that we should choose bootstrap samples $X_1^*, \ldots, X_n^*$ of size n and plug them in the statistic $\widehat{\theta}_n(X_1, \ldots, X_n)$ of interest: $\widehat{\theta}_n^* = \widehat{\theta}_n(X_1^*, \ldots, X_n^*)$. However, there is no real good reason for this procedure. For example, we could generate bootstrap samples $X_1^*, \ldots, X_{k_n}^*$ for any subsequence (k_n) of the integers satisfying $k_n \to \infty$. Now imagine that you bootstrap the maximum from such a subsample and write as before M_n^* for $\max(X_1^*, \ldots, X_{k_n}^*)$. Then (2.9) turns into

$$P^*(M_n^* = M_n) = 1 - [P^*(X_1^* < M_n]^{k_n} = 1 - [1 - n^{-1}]^{k_n}\,.$$

Now it it clear what we have to do in order to get rid of the dominance of the M_n-value: we have to choose the sample size k_n in such a way that $k_n/n \to 0$, i.e., smaller than the original sample size n. This procedure can indeed be shown to work in many situations of interest. It is called *subsample bootstrap* and can be recommended as a procedure in those situations when the extreme order statistics play an important role in the statistic to be bootstrapped. The bootstrap problem for the sample maximum was solved by using the subsample method by Bretagnolle [11], Swanepoel [56] and more recently by Deheuvels et al. [17] who derived a natural range for $k_n \to \infty$.

The subsample bootstrap is also useful for estimators of the tail index of a Pareto-like distribution or, more generally, of the extreme value index, but also for tail and high quantile estimation in extreme value statistics. We consider one such statistic in more detail. The tail parameter α of a Pareto-like distribution with $1 - F(x) = cx^{-\alpha}$ for some $c > 0$, $x \geq x_0$, is typically estimated by statistics which involve an increasing number of upper order statistics. Writing $X_{(1)} \leq \cdots \leq X_{(n)}$ for the order statistics of the iid sample $X_1, \ldots, X_n$, the quantity

$$\widehat{\alpha}_H = \left(m^{-1} \sum_{i=1}^{m} \log X_{(n-i+1)} - \log X_{(n-m)} \right)^{-1}$$

estimates α consistently provided $m = m_n \to \infty$ and $m/n \to 0$. The statistic $\widehat{\alpha}_H$ is the *Hill estimator* of the tail parameter α. The Hill estimator and its ramifications have caused a flood of literature on asymptotic properties; see Chapter 6 in Embrechts et al. [25] for a discussion. Among it, a few papers deal with the bootstrapped Hill estimator. As mentioned above, the authors encounter exactly the problems mentioned above since $\widehat{\alpha}_H$ is clearly very much dependent on the upper order statistics in the sample and they propose the subsample bootstrap as an alternative since the naive bootstrap does not work. We refer to Guillou [29] and the references therein for some recent work on the subsample bootstrap and estimation of α.

As a matter of fact, the mentioned problem of the bootstrap for the sample mean with infinite variance was also solved by using the subsample bootstrap; see Arcones and Giné [1].

The bootstrap subsampling method is also discussed in a recent paper by Bickel et al. [9]

2.2.6 Does the bootstrap work better than the CLT?

This is a good question. Which means: there is no precise answer possible.

In his appraisal of the bootstrap technology Efron frequently pointed at the fact that this method is also good when the sample sizes are small. Clearly, the bootstrap algorithm can be

applied for *any* sample size but then the question arises as to what one actually achieves with this algorithm. There is no theoretical answer so far.

A natural question, for example, would be as to whether the bootstrap approximation to the distribution of some reasonable statistics such as the sample mean or the sample median is better than the approximation by the CLT. We have seen in Section 1.1.4 that the approximation through the CLT can be quite bad; the Berry-Esséen inequality gives the pessimistic rate $1/\sqrt{n}$ which is attained by a sum of symmetric Bernoulli random variables.

The bootstrap can indeed be shown to do a better asymptotic job than the CLT under some smoothness conditions on the underlying distributions. To illustrate this, we will, as before, focus on our war horse, the sample mean $\overline{X}_n$ of an iid sequence of iid random variables $X, X_1, X_2, \ldots$ with common distribution F. We assume that $\sigma^2 = \mathrm{var}(X) < \infty$. The main technique for showing that the bootstrap is advantageous is the use of asymptotic or Edgeworth expansions; see Section 1.1.5. In the context of the bootstrap it was early on propagated by Singh [54]. Afterwards Peter Hall did an enormous amount of work and gained an almost complete picture of the bootstrap by using expansions of different kinds. His contributions are summarized in his monograph [32]. In what follows we want to explain the main ideas of his approach, following Chapter 3 in [32].

As before, let

$$G_n(x) = P\left(\sqrt{n}\,(\overline{X}_n - EX) \le x\right) = P\left(\frac{\sqrt{n}}{\sigma}\,(\overline{X}_n - EX) \le x/\sigma\right), \quad x \in \mathbb{R},$$

denote the standard normal distribution by Φ and its density by φ. Assume that we can use the theory of Section 1.1.5 to obtain the first order asymptotic expansion:

$$\begin{aligned} G_n(x) - \Phi(x/\sigma) &= n^{-1/2} Q_1(x/\sigma) + O(n^{-1}) \\ &= -n^{-1/2}\,\varphi(x/\sigma)\,((x/\sigma)^2 - 1)\,\frac{E(X-EX)^3}{6\,\sigma^3} + O(n^{-1}), \end{aligned} \tag{2.10}$$

uniformly in $x \in \mathbb{R}$. This expansion holds if F has a density and $E|X|^4 < \infty$ (Theorem 1.2). The corresponding asymptotic expansion for the bootstrapped sample mean is obtained simply by replacing the X_i's everywhere by their bootstrap analogs and by replacing the moments of X with respect to F by the corresponding moments of X^* with respect to the empirical distribution function F_n, i.e., by the sample moments. Recall that for a realization $x_1 = X_1(\omega), x_2 = X_2(\omega), \ldots$ of the (X_i) sequence,

$$\begin{aligned} E^*(X^*) &= \overline{x}_n, \\ (\sigma^*)^2 &= \mathrm{var}^*(X_1^*) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2 = s_n^2, \end{aligned}$$

$$E^*((X^* - E^*(X_1^*))^3) = E^*((X^* - \overline{x}_n)^3) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^3 =: \overline{x_n^3}.$$

The first order bootstrap asymptotic expansion then takes on the form

$$\begin{aligned} G_n^*(x) &= P^*\left(\sqrt{n}\,(\overline{X}_n^* - \overline{x}_n) \le x\right) \\ &= P^*\left(\frac{\sqrt{n}}{s_n}(\overline{X}_n^* - \overline{x}_n) \le x/s_n\right) \\ &= \Phi(x/s_n) - n^{-1/2}\,\varphi(x/s_n)\,((x/s_n)^2 - 1)\,\frac{\overline{x_n^3}}{6\,s_n^3} + O_P(n^{-1}). \end{aligned} \tag{2.11}$$

Under smoothness conditions (such as the existence of a density of F, existence of moments and further conditions; see [32], Chapter 5) the latter expansion can be shown to make sense as $n \to \infty$, uniformly for $x \in \mathbb{R}$. Notice one important difference between (2.10) and (2.11): the remainder term in (2.10) is of the order $O_P(n^{-1})$. This means $O_P(n^{-1}) = C_n(\omega)n^{-1}$, where

$$\lim_{\epsilon \to \infty} \limsup_{n \to \infty} P(|C_n| > \epsilon) = 0\,. \tag{2.12}$$

We do not know how big $C_n(\omega)$ actually is, except that it is *stochastically bounded* in the sense of (2.12). A sufficient condition for $C_n = O_P(1)$ is $C_n \xrightarrow{d} Y$ for some random variable Y.

In order to judge the accuracy of the approximation of $G_n(x)$ by $G_n^*(x)$ we consider the difference

$$G_n(x) - G_n^*(x) \;=\; \Phi(x/\sigma) - \Phi(x/s_n) + n^{-1/2}[Q_1(x/\sigma) - \widehat{Q}_1(x/s_n)] + O_P(n^{-1})\,. \tag{2.13}$$

Here $\widehat{Q}_1$ is obtained from Q_1 by replacing all moments by the corresponding sample moments. A Taylor expansion argument yields for some ξ_n between $1/\sigma$ and $1/s_n$,

$$\begin{aligned} \Phi(x/\sigma) - \Phi(x/s_n) &= x\,\varphi(x\,\xi_n)\,[\sigma^{-1} - s_n^{-1}] \\ &= x\,\varphi(x\,\xi_n)\,\frac{s_n^2 - \sigma^2}{\sigma s_n(s_n + \sigma)}\,. \end{aligned}$$

Multiplying the right-hand side by $\sqrt{n}$, the CLT, the SLLN and the continuous mapping theorem (see Section 1.1.6) together yield

$$\xrightarrow{d}\; x\,\varphi(x/\sigma)\,\frac{1}{2\sigma^3}\,N(0, \mathrm{var}((X_1 - EX_1)^2))\,,$$

provided the moments on the right-hand side are finite. In other words, the approximation (2.13) can never be better than $O_P(1/\sqrt{n})$ even if the remaining terms in the expansion (2.13) converge to zero faster. This is a slightly disappointing fact: the bootstrap approximation $G_n^*(x)$ to $G_n(x)$ is of the same rate as the worst rate in the CLT described by the Berry-Esséen inequality.

Re-inspecting the previous formulae, it is clear where the rate $O_P(1/\sqrt{n})$ comes from: it is the CLT $\sqrt{n}\,(s_n^2 - \sigma^2) \xrightarrow{d} N(0, \mathrm{var}((X_1 - EX_1)^2))$ which cannot be improved. Or more informatively, it comes from the difference $\Phi(x/\sigma) - \Phi(x/s_n)$. This yields a hint how we could improve upon the rate in the asymptotic expansion. Indeed, replacing $G_n(x)$ and $G_n^*(x)$ by the distribution functions of the corresponding studentized versions, the following asymptotic expansions can be shown to hold (Section 2.6 and Chapter 3 in [32]):

$$\widetilde{G}_n(x) \;=\; P\left(\frac{\sqrt{n}}{s_n}\,(\overline{X}_n - EX) \le x\right) = \Phi(x) + n^{-1/2}\,\varphi(x)\,(2x^2+1)\,\frac{E(X-EX)^3}{6\sigma^3} + O(n^{-1})\,,$$

$$\widetilde{G}_n^*(x) \;=\; P^*\left(\frac{\sqrt{n}}{s_n^*}\,(\overline{X}^*_n - \overline{x}_n) \le x\right) = \Phi(x) + n^{-1/2}\,\varphi(x)\,(2x^2+1)\,\frac{\overline{x_n^3}}{6\,s_n^3} + O_P(n^{-1})\,,$$

where $(s_n^*)^2$ is the bootstrap estimator of the sample variance s_n^2:

$$(s_n^*)^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i^* - \frac{1}{n}\sum_{i=1}^n X_i^*\right)^2\,.$$

Compare the last expansions with those in (2.10) and (2.11). Notice that the studentization of the sample mean leads to a different form of the polynomials in the asymptotic expansion; this is due to a Δ-method argument as outlined in Section 2.6 of [32].

Now, we get an improvement through the approximation

$$\widetilde{G}_n(x) - \widetilde{G}_n^*(x) \;=\; n^{-1/2}\phi(x)\,(2x^2+1)\left[\frac{E(X-EX)^3}{6\sigma^3} - \frac{\overline{x_n^3}}{6\,s_n^3}\right] + O_P(n^{-1})\,.$$

An argument, similar to the one above, but more tedious, shows that the difference between the ratio of the moments on the right-hand side and their sample counterpart, when multiplied with $\sqrt{n}$, satisfies the CLT, hence it is of the order $O_P(1/\sqrt{n})$. In view of the pre-factor $n^{-1/2}$ this amounts to an order $O_P(n^{-1})$.

This can be understood as an improvement upon the asymptotic rate in the CLT, and in the literature it is celebrated as an indication of the superior behavior of the bootstrap compared to the CLT. Although this rate does not solve the small sample problem, it shows that, in an asymptotic sense, the distribution of the studentized bootstrap sample mean gives a better approximation to the distribution of the studentized sample mean than the normal distribution.

Peter Hall points at another advantage of the bootstrap approximation which is actually another consequence of the bootstrap asymptotic expansion. It turns out that the asymptotic accuracy of the approximation of the α-quantiles of various important statistics is better than given by the normal approximation. Approximation of the quantiles of statistics is bread and butter in statistics, for example, if one is interested in confidence bands or critical regions for testing. A procedure to improve upon the normal approximation is called *Cornish-Fisher expansion* which we want to explain now. Suppose that t_α is the α-quantile of the standard normal distribution: $\Phi(t_\alpha) = \alpha$. Assume for simplicity we are interested in the distribution of the centered and normalized sample mean of iid random variables $X, X_1, X_2, \ldots$, i.e., $EX = 0$, $\mathrm{var}(X) = 1$ and $G_n(x) = P(\sqrt{n}\,\overline{X}_n \le x)$. Recall the notion of quantile function from page 7 and write $g_\alpha = G_n^{\leftarrow}(\alpha)$ for the α-quantile of the distribution function G_n. Consider the *formal* Edgeworth expansion

$$G_n(x) = \Phi(x) + \sum_{\nu=1}^{\infty} Q_\nu(x)\; n^{-\nu/2} = \Phi(x) + \sum_{\nu=1}^{\infty} P_\nu(x)\,\varphi(x)\; n^{-\nu/2}\,, \tag{2.14}$$

where we write $Q_\nu(x) = P_\nu(x)\varphi(x)$. Recall from Section 1.1.5 that P_ν is a polynomial whose coefficients depend on the moments of X. The idea of a Cornish-Fisher expansion is to use the latter formal asymptotic expansion to get a formal Taylor expansion for the quantile g_α at the corresponding quantile t_α of the standard normal distribution:

$$g_\alpha = t_\alpha + n^{-1/2}p_1(t_\alpha) + n^{-1}p_2(t_\alpha) + \cdots\,. \tag{2.15}$$

Here the p_i's are polynomials whose coefficients depend on the moments of X. To determine these polynomials, one has to insert (2.15) into the following equation (2.14) (here we assume that this equation has a unique solution):

$$\begin{aligned} &\Phi(\alpha) = \alpha = G_n(g_\alpha) \\ = \;& \Phi\left(t_\alpha + \sum_{i=1}^{\infty} p_i(t_\alpha)n^{-i/2}\right) + \sum_{\nu=1}^{\infty} P_\nu\left(t_\alpha + \sum_{i=1}^{\infty} p_i(t_\alpha)n^{-i/2}\right)\,\varphi\left(t_\alpha + \sum_{i=1}^{\infty} p_i(t_\alpha)n^{-i/2}\right) n^{-\nu/2}\,. \end{aligned}$$

Then, Taylor expanding the right-hand side at t_α, one can derive the form of the p_i's. We refer to [32], Section 2.5, for details. The first two polynomials are then of the form

$$\begin{aligned} p_1(x) &= -P_1(x)\,, \\ p_2(x) &= P_1(x)P_1'(x) - \frac{1}{2}\,x\,[P_1(x)]^2 - P_2(x)\,. \end{aligned}$$

Hall [32], Theorem 2.4, gives conditions for the validity of the Cornish-Fisher expansion

$$g_\alpha = t_\alpha + n^{-1/2}p_1(t_\alpha) + n^{-1}p_2(t_\alpha) + \cdots + n^{-j/2}p_j(t_\alpha) + o(n^{-j/2})\,, \tag{2.16}$$

uniformly for $\alpha \in (\epsilon, 1-\epsilon)$, $0 < \epsilon < 0.5$. This assumption is crucial; for α too close to 0 (for example $\alpha = \alpha_n \to 0$) or 1 the expansion (2.16) does not work. This expansion shows the quality of the approximation of the α-quantile of the distribution G_n of the sample mean to the corresponding quantile of the standard normal distribution. In particular, the traditional normal approximation of g_α via the CLT gives us uniformly for $\alpha \in (\epsilon, 1-\epsilon)$

$$g_\alpha - t_\alpha = n^{-1/2}p_1(t_\alpha) + O(n^{-1}) = O(n^{-1/2})\,,$$

i.e., the error is of the order $O(1/\sqrt{n})$.

Hall used similar ideas for the bootstrap. Again replacing the X_i's by the X_i^*'s and the moments of X by the sample moments, one obtains a bootstrap Cornish-Fisher expansion for the distribution G_n^* of the bootstrapped sample mean:

$$\widehat{g}_\alpha = t_\alpha + n^{-1/2}\widehat{p}_1(t_\alpha) + n^{-1}\widehat{p}_2(t_\alpha) + \cdots + n^{-j/2}\widehat{p}_j(t_\alpha) + o_P(n^{-j/2})\,,$$

where the polynomials $\widehat{p}_i$ are obtained from the p_i's by replacing the moments of X by the sample moments. We mention that these expansions remain valid for the studentized versions of the sample mean as well although the form of the polynomials then changes when compared to the sample mean; for a similar case see the asymptotic expansion of the studentized sample mean in the beginning of this section. Considering the bootstrap Cornish-Fisher expansion for the studentized sample mean $\sqrt{n}\,(\overline{X}_n - EX)/s_n$ which is based on the asymptotic expansion of $\sqrt{n}\,(\overline{X}_n^* - \overline{X}_n)/s_n^*$, one can conclude that the discrepancy between g_α and $\widehat{g}_\alpha$ is of the order

$$g_\alpha - \widehat{g}_\alpha = n^{-1/2}[p_i(t_\alpha) - \widehat{p}_i(t_\alpha)] + O_P(n^{-1})\,,$$

and since $p_i(t_\alpha) - \widehat{p}_i(t_\alpha)$ only depends on the differences between certain moments of X and the corresponding sample moments, the CLT applied to these differences, as above, gives the order $O_P(1/\sqrt{n})$. Together with the pre-factor $n^{-1/2}$ this gives the order $O_P(n^{-1})$ which compares favorably to the rate $O(1/\sqrt{n})$ achieved by the CLT.

Finally, the above reasoning also works for smooth functions of the sample mean and its studentized version. See Hall [32] for details.

Comments

The bootstrap of smooth functions of the sample mean can give better *asymptotic* approximations than the standard normal approximation via the CLT. Rates are of the order $O_P(n^{-1})$ instead of $O(n^{-1/2})$ for the CLT. The discussion at the beginning of the section showed that one has to be careful; studentization was crucial for the improvement on the rates. Two critical points may be mentioned here. 1. The above theory is again a large sample theory, whereas in many

practical applications one has sample sizes of a couple of dozen observations, if at all. 2. The rate $O_P(n^{-1})$ depends on the realization of the sample, i.e., on ω. This means that it is not a precise approximation.

Both mentioned problems cannot be solved by theoretical means.

Only experience and experimenting with simulations can help.

The mentioned book by Hall [32] covers the whole range of asymptotic results for the bootstrap in the case of smooth functions of the sample mean, i.e., when the sample mean gives a good approximation to this function. It also refers to many tricks one can use in one or the other situation (improvement on bias or variance of estimators, coverage of confidence intervals, etc.). The book, however, requires basic knowledge of asymptotic statistics and techniques such as asymptotic expansions. It is therefore not an introductory course on the topic.

2.3 Dependence and the bootstrap

Early on it was tried to generalize the bootstrap to dependent data. A natural class of dependent sequences consists of the *stationary ergodic sequences* (X_n). We first want to recall what this means. (Strict) *stationarity* means that

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{1+h}, \ldots, X_{n+h}), \quad n \geq 1, h \geq 0. \tag{2.17}$$

The symbol $\stackrel{d}{=}$ refers to identity in law. Stationarity thus means that the finite-dimensional distributions of the process are invariant under shifts in time and, according to Kolmogorov's consistency theorem, the finite-dimensional distributions determine the distribution of the underlying process. Stationarity does not mean that the strong law of large numbers holds. For example, let $A, Y_1, Y_2, \ldots$ be independent non-degenerate random variables and such that (Y_i) is iid. For $X_i = AY_i$ we then obtain

$$\frac{1}{n}\sum_{i=1}^{n} X_i \stackrel{\text{a.s.}}{\to} A\,EY_1,$$

provided EY_1 is well defined. This means that the limit is random. If one wants to avoid this situation one has to assume more, namely that (X_t) is *ergodic*. This means, roughly speaking, that for all Borel functions f for which $Ef(X_1)$ is defined, the strong law of large numbers holds:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \stackrel{\text{a.s.}}{\to} Ef(X_1).$$

In particular, if (X_i) is stationary ergodic, then the empirical distribution function converges pointwise for every $x \in \mathbb{R}$:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \stackrel{\text{a.s.}}{\to} F(x), \tag{2.18}$$

where F is the distribution function of X_1. The same proof as in the iid case (which only uses the strong law of large numbers (2.18) and the monotonicity of F and F_n) shows that the Glivenko-Cantelli theorem holds (cf. Section 1.1.2):

$$\sup_x |F_n(x) - F(x)| \stackrel{\text{a.s.}}{\to} 0.$$

In what follows, we will always assume that the sequence (X_i) is stationary and ergodic. Then the Glivenko-Cantelli result holds. The latter result was one of the bases for the bootstrap in the iid case and so one might be tempted to believe that in the dependent case the naive bootstrap works as before. This, however, is a wrong belief, as the following simple example shows.

Example 2.5 (The bootstrap of the sample mean fails for dependent data.)
As before, let $\overline{X}_n$ be the sample mean. We assume that $\mathrm{var}(X_1) < \infty$. Then it is not difficult to see that

$$\begin{aligned} \mathrm{var}(\overline{X}_n) &= \frac{1}{n^2} \sum_{1 \le i,j \le n} \mathrm{cov}(X_i, X_j) \\ &= \frac{1}{n} \left[\mathrm{var}(X_1) + 2 \sum_{h=1}^{n-1} \left(1 - \frac{h}{n}\right) \mathrm{cov}(X_1, X_{1+h}) \right] . \end{aligned} \tag{2.19}$$

For independent random variables the second term on the right-hand side would disappear, but is does not in general for dependent random variables. It is not difficult to see that

$$n \,\mathrm{var}(\overline{X}_n) \to \tau^2 := \mathrm{var}(X_1) + 2 \sum_{h=1}^{\infty} \mathrm{cov}(X_1, X_{1+h}) \,,$$

provided the right hand infinite series is absolutely summable. On the other hand, the naive bootstrapped sample mean has variance

$$\mathrm{var}^*(\overline{X}^*_n) = \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \,,$$

where, as usual, $x_1 = X_1(\omega), x_2 = X_2(\omega), \ldots$. Since we have assumed that the sequence (X_i) is ergodic, the strong law of large numbers yields

$$n \,\mathrm{var}^*(\overline{X}^*_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \to \sigma^2 := \mathrm{var}(X_1)$$

for P-a.e. ω. A comparison of the two asymptotic variances τ^2 and σ^2 indicates that the limit theory for the sample mean and its bootstrap version could be completely different. This is indeed correct. It is well known (but not trivial to prove) that the CLT holds for the sample mean of a stationary ergodic sequence (X_n) provided some additional *mixing conditions* hold and $E|X_1|^{2+\delta} < \infty$ for some $\delta > 0$, i.e.,

$$\sqrt{n} \, \left[\overline{X}_n - EX_1 \right] \xrightarrow{d} N(0, \tau^2) \,.$$

We do not want to discuss mixing conditions here. We only mention that the sequences $(X_1, \ldots, X_h)$ and $(X_{h+n}, X_{h+n+1}, \ldots)$ become "asymptotically independent" for any h as $n \to \infty$ and that one can "measure the rate" at which this happens. For the CLT to hold one needs that these rates decay to zero sufficiently fast. The interested reader is referred to the classical books by Ibragimov and Linnik [35] and Billingsley [10] or the more recent overview by Doukhan [19].

For the CLT in the case of stationary ergodic sequences observe that we can apply the Berry–Esséen inequality from Section 1.1.4 (Φ is the standard normal distribution):

$$\left| P^* \left(\frac{\sqrt{n}}{\sqrt{\mathrm{var}^*(X_1^*)}} \left(\overline{X}^*_n - E^*(X_1^*) \right) \le x \right) - \Phi(x) \right| \le C_{\mathrm{BE}} \, \frac{1}{\sqrt{n}} \frac{E^* |X_1^* - E^*(X_1^*)|^3}{[\mathrm{var}^*(X_1^*)]^{3/2}} \,.$$

The moments on the right-hand side are with respect to the empirical distribution function, hence they are the corresponding sample moments. Notice that

$$\mathrm{var}^*(X_1^*) = E^*[(X_1^*)^2] - [E^*(X_1^*)]^2$$

and, by the triangle inequality in L_3,

$$\left[E^*|X_1^* - E^*(X_1^*)|^3\right]^{1/3} \le \left[E^*|X_1^*|^3\right]^{1/3} + |E^*(X_1^*)|\,.$$

Using this knowledge, one can see that the right-hand side in the Berry–Esséen inequality can be estimated by quantities which only depend on the sample power moments up to order 3. Then apply the strong law of large numbers to the sample moments to conclude that *the CLT holds for the bootstrap sample mean for any stationary ergodic sequence* (X_i) *with finite 3rd moment.* As a matter of fact, it is well known that the latter statement is in general wrong for the sample mean of a stationary ergodic sequence. We are certainly not too surprised about this result. Indeed, the naive bootstrap *destroys* the dependence structure of the underlying sequence (X_i) completely — the bootstrap mechanism of independently drawing from the sample $x_1, \ldots, x_n$ does not make a difference between dependent and independent data. □

Now the task is quite clear: we have to change the bootstrap algorithm in a way such that the dependence structure in the data is kept intact in some sense. This is possible if one has a particular model for the stationary sequence (X_i) which prescribes how the X_t's depend on an iid sequence (Z_t). A simple example would be an ARMA(p, q) (*autoregressive moving average process of order* (p, q)) process given by the difference equations:

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}\,, \quad t \in \mathbb{Z}\,,$$

for given integers $p \ge 0$ and $q \ge 0$ and an iid noise sequence (Z_t). For example, consider the AR(1) equation

$$X_t = \varphi X_{t-1} + Z_t\,, \quad t \in \mathbb{Z}\,,$$

which can be shown to have a stationary solution if $|\varphi| < 1$ and a mild moment condition on Z_t holds (see Brockwell and Davis [12], Chapter 3). Clearly, (Z_t) is unobservable, but we could get an approximation to the Z_t's if we were to estimate φ by some standard method (least squares, maximum likelihood, etc.; see [12], Chapter 8) first and then to calculate the *residuals*

$$\widehat{Z}_t = X_t - \widehat{\varphi}\, X_{t-1}\,, \quad t = 2, \ldots, n\,,$$

from the observed values $X_1 = X_1(\omega), \ldots, X_n = X_n(\omega)$. Clearly, the $\widehat{Z}_t$'s are then not independent (why?), but if the model fit were good we would expect them to be "almost independent". Now, one could base a bootstrap method for functionals of the X_t's (such as the mean, variance, etc.) on the empirical distribution function of the residuals:

$$\widehat{F}_n(x) = \frac{1}{n-1} \sum_{i=2}^{n} I_{(-\infty, x]}(\widehat{Z}_i)\,, \quad x \in \mathbb{R}\,.$$

Indeed, suppose you draw iid $Z_1^*, \ldots, Z_n^*$ with distribution $\widehat{F}_n$, you can then calculate the bootstrap sample

$$X_1^* = Z_1^*\,, \quad X_2^* = \widehat{\varphi}\, X_1^* + Z_2^*\,, \ldots\,, \quad X_n^* = \widehat{\varphi}\, X_{n-1}^* + Z_n^*\,.$$

The bootstrap sample mimics the dependence structure of the original data and also keeps the distributional structure somewhat intact because we may hope that the residuals have very much the same distribution as the iid sequence (Z_t). Clearly, we have used a lot of approximation on the way to the definition of the bootstrap sample $X_1^*, \ldots, X_n^*$. Moreover, one is interested in the distribution of statistics $\widehat{\theta}_n(X_1^*, \ldots, X_n^*)$ such as the sample mean, the sample variance, the sample autocovariances and sample autocorrelations and one wants to know whether the distribution of these statistics approximates the corresponding distributions of $\widehat{\theta}_n(X_1, \ldots, X_n)$. This is the question as to whether the bootstrap works or whether the bootstrap is consistent in this case. It can be shown that the bootstrap works for such models; we refer to Heimann and Kreiss [33] and Kreiss and Franke [36] to get some taste of the method.

The basic idea in this approach was to use some hidden iid property for constructing the bootstrap sequence based on the empirical distribution function of the "residuals". A similar approach is possible for Markov chains when one uses the so-called *regenerative* or *renewal property*, i.e., these processes contain blocks of random lengths which are independent. The *renewal bootstrap* is based on the empirical distribution function of these independent blocks. We refer to unpublished work by Radulović [48] and Paparoditis and Politis [42]. An alternative way to bootstrap Markov chains is to estimate the transition probabilities first and then to simulate from these probabilities independent realizations of the Markov chain. (Clearly, there is a problem with the choice of the initial values in the Markov chain.) We refer to Datta and McCormick [15] and Athreya and Fuh [5] for this approach.

The above approaches require some special structure of the stationary process. A completely different approach was taken by Künsch [37] which covers more general classes of stationary processes. He introduced the *moving blocks bootstrap* a version of which we want to discuss now.

Consider an integer b which is supposed to be small compared to n and fixed for the moment. For ease of presentation we always assume that $m = n/b$ is an integer. Introduce the overlapping blocks of size b:

$$\mathbf{X}_1 = \{X_1, \ldots, X_b\}, \ldots, \mathbf{X}_{n-b+1} = \{X_{n-b+1}, \ldots, X_n\}.$$

The new iid samples are now drawn with replacement from the set $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n-b+1}\}$, i.e., we draw from the empirical measure F_n with atoms at the vectors $\mathbf{X}_1, \ldots, \mathbf{X}_{n-b+1}$. Or in other words, any of the values $\mathbf{X}_i$, $i = 1, \ldots, n-b+1$, is chosen with the same probability $(n-b+1)^{-1}$. The notion *moving blocks bootstrap* clearly gained its name from the construction of the $\mathbf{X}_i$ sequence. Notice that the dependence structure inside any block remains untouched.

As in the naive bootstrap case, we label the iid bootstrap sample with common distribution F_n with a star:

$$\mathbf{X}_1^*, \ldots, \mathbf{X}_m^*.$$

The latter sequence can now be used to construct bootstrap versions of standard statistics. As before, we focus on the sample mean. Write $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^b$ and

$$\overline{X}_n^* = \frac{1}{n} \sum_{i=1}^{m} (\mathbf{1}, \mathbf{X}_i^*),$$

where $(\cdot, \cdot)$ stands for the usual inner product. Notice that each summand

$$Y_i^* = (\mathbf{1}, \mathbf{X}_i^*), \quad i = 1, \ldots, m$$

is the sum of b successive values X_i, i.e., $\overline{X}_n^*$ is indeed an average of $n = b \times m$ random variables. The original dependence structure of the blocks of X_k's is kept in each subsum Y_i^*, but the Y_i^*'s are clearly (conditionally) iid.

As before, we want to compare the variances of the bootstrapped sample mean and the sample mean. Given that they yield similar results, we may take this as an indication of the fact that the bootstrap might work for the sample mean. Write

$$Y_i = (\mathbf{1}, \mathbf{X}_i)\,, \quad i = 1, \ldots, n-b+1\,.$$

As before we write P^*, E^*, var^* for the quantities related to the moving blocks bootstrap procedure. Then

$$\begin{aligned} E^*(Y_i^*) &= \frac{1}{n-b+1} \sum_{i=1}^{n-b+1} Y_i\,, \\ \mathrm{var}^*(Y_i^*) &= \frac{1}{n-b+1} \sum_{i=1}^{n-b+1} Y_i^2 - [E^*(Y_i^*)]^2\,. \end{aligned}$$

Since the Y_i^*'s are iid given the data and $b = n/m$,

$$\begin{aligned} n\,\mathrm{var}^*(\overline{X}_n^*) &= \frac{m}{n}\,\mathrm{var}^*(Y_1^*) \\ &= \frac{1}{b}\left[\frac{1}{n-b+1}\sum_{i=1}^{n-b+1} Y_i^2 - \left(\frac{1}{n-b+1}\sum_{i=1}^{n-b+1} Y_i\right)^2\right]. \end{aligned}$$

For fixed b, by virtue of the ergodic theorem, the right-hand side converges a.s. as $n \to \infty$ to

$$\begin{aligned} \frac{1}{b}\,\mathrm{var}(Y_1) &= \frac{1}{b}\,\mathrm{var}(X_1 + \cdots + X_b) \\ &= \mathrm{var}(X_1) + 2\sum_{h=1}^{b-1}\left(1 - \frac{h}{b}\right)\mathrm{cov}(X_1, X_{1+h})\,, \end{aligned} \tag{2.20}$$

where we used formula (2.19) in the last step. Now recall that the CLT for mixing (X_i) had the asymptotic variance

$$\tau^2 = \mathrm{var}(X_1) + 2\sum_{h=1}^{\infty}\mathrm{cov}(X_1, X_{1+h})\,. \tag{2.21}$$

Thus, in order to get the right asymptotic variance in the CLT, i.e., in order to switch from (2.20) to (2.21), we have to assume that $b = b_n \to \infty$ at some slow rate. The latter fact is usually expressed as $m = m_n = b_n/n \to 0$. This means that the block length b has to be chosen "small" with respect to the sample size n but "not too small". For real life data these rules are clearly very vague, and so one has to depend on simulations in order to check how large b has to be chosen. There is a trade-off between b and m. If we choose b small, then we have m large, but if b is too small we destroy the dependence structure in the data, i.e., we are too close to the naive bootstrap procedure. On the other hand, if b is big we keep the dependence structure in larger subsamples intact, but we have only a small sample size m which may make the use of asymptotic methods (CLT, SLLN) questionable and which leads to an increase of the variance of the estimators.

Comments

Künsch [37] showed that his moving blocks bootstrap procedure works for various kinds of statistics beyond the sample mean and the sample variance, including M-estimators (certain "robust" estimators). An alternative to the moving blocks bootstrap was suggested by Carlstein [14]. He proposed to base the bootstrap on non-overlapping blocks. The obvious disadvantage is that one gets less blocks for the same number b as in Künsch's bootstrap. Moreover, the moving blocks bootstrap captures the "moving" dependence in the sample, thus also the dependence if one moved from one block (in Carlstein's method) to another. Carlstein's bootstrap breaks the dependence between two successive blocks in quite an unnatural way. Künsch [37] gives a careful comparison between his and Carlstein's method. Among others, in his Remark 3.3, he mentions that Carlstein's bootstrap can yield a larger variance of the bootstrapped statistics. Improvements on the bootstrap for dependent sequences were given by Bühlmann [13] and Radulović [47]. A recent overview of bootstrap methods in the dependent case can be found in Radulović [49].

2.4 Extensions of Efron's bootstrap

2.4.1 The wild bootstrap

We start with iid random variables $X, X_1, X_2, \ldots$ and denote by $x_1 = X_1(\omega), x_2 = X_2(\omega), \ldots$ one realization of this sequence. The iid naive bootstrap sequence $X_1^*, X_2^*, \ldots$ has representation

$$X_i^* = \sum_{k=1}^{n} x_k I_{\{n^{-1}(k-1,k]\}}(U_i)\,, \quad i = 1, 2, \ldots\,, \tag{2.22}$$

where (U_i) is a sequence of iid random variables with a uniform on $(0, 1)$ distribution. Indeed, the latter construction says that, given the data $x_1, \ldots, x_n$, the X_i^* are iid (since a function of U_i only) and X_i^* has a uniform distribution on $x_1, \ldots, x_n$. This construction shows nicely the two sources of randomness which are involved in the bootstrap sequence: the original sample of $x_i = X_i(\omega)$ values and, independently of the latter, an iid sequence of uniform $U(0, 1)$ random variables. Whereas the x_i's represent the randomness of the real world, the U_i's generate the randomness in the bootstrap world.

Switching again to the bootstrap sample mean and using representation (2.22), we get an interesting representation of $\overline{X}_n^*$:

$$\overline{X}_n^* \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^* = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{n} x_k I_{\{n^{-1}(k-1,k]\}}(U_i)\right] = \sum_{k=1}^{n} x_k \left[\frac{1}{n} M_{n,k}\right]\,,$$

where

$$M_{n,k} = \sum_{i=1}^{n} I_{\{n^{-1}(k-1,k]\}}(U_i)\,, \quad k = 1, \ldots, n\,. \tag{2.23}$$

The vector $\mathbf{M}_n = (M_{n,1}, \ldots, M_{n,n})$ has a multinomial distribution $\mathrm{Mult}(n; n^{-1}, \ldots, n^{-1})$. Indeed, the vector is formed from n draws on n equally likely cells, independent of $x_1, \ldots, x_n$. Notice that the vector $\mathbf{M}_n$ has *exchangeable* components: you can permute the components in any way and the distribution of the vector does not change. This is immediate from the representation (2.23).

The vector $\mathbf{M}_n$ has the following properties:

$$M_{n,k} \geq 0\,, \quad k = 1, \ldots, n\,, \tag{2.24}$$

$$\frac{1}{n}\sum_{k=1}^{n} M_{n,k} = 1\,, \tag{2.25}$$

$$n\sum_{k=1}^{n}(n^{-1}M_{n,k} - n^{-1})^2 \xrightarrow{P} c\,, \tag{2.26}$$

for some positive constant $c > 0$, where $c = 1$ in this case. Conditions (2.25) and (2.26) mean that, on average, the weights $M_{n,i}$ are close to 1.

Thus the naive bootstrap of the sample mean, up to the factor n^{-1}, is nothing but an average of the original data with multinomially distributed random weights which are chosen independently of the original sample. Keeping this structural result in mind, one may immediately ask for natural extensions. One of them would be to replace the vector $\mathbf{M}_n$'s by another vector with exchangeable components, independent of the (X_i) sequence. Since a vector of n iid components is exchangeable, this would be the most trivial choice, but these weights would in general not satisfy conditions (2.24)–(2.26) which ensure that the weights cannot become too wild and that theoretical properties such as bootstrap consistency hold.

This idea can be considered as another resampling scheme extending the bootstrap in various directions. The resulting procedure is referred to as *weighted* or *general* or sometimes as *wild* bootstrap. It can be described as follows.

- Let $X, X_1, X_2, \ldots$ be iid with distribution F.
- Let $\mathbf{M}_n = (M_{n,1}, \ldots, M_{n,n})$ be an exchangeable random vector, independent of (X_i), satisfying (2.24)–(2.26). Given a realization $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$, the bootstrap sample mean is given by
$$\overline{X}_n^* = \frac{1}{n}\sum_{i=1}^{n} x_i\, M_{n,i}\,.$$
- The expectation of $\overline{X}_n^*$ and its variance can be calculated explicitly. Here and in what follows we write P^*, E^*, var^* for the probability measure generated by $\mathbf{M}_n$ given $x_1, \ldots, x_n$. Then, by (2.24)–(2.25) and exchangeability:
$$E^*(\overline{X}_n^*) = \frac{1}{n}\sum_{i=1}^{n} x_i E(M_{n,i}) = EM_{n,1}\,\overline{x}_n = \overline{x}_n\,,$$
$$\mathrm{var}^*(\overline{X}_n^*) = E^*\left(\frac{1}{n}\sum_{i=1}^{n} x_i M_{n,i} - \overline{x}_n\right)^2$$
$$= \frac{1}{n^2}\,\mathrm{var}(M_{n,1})\sum_{i=1}^{n} x_i^2 + \frac{1}{n^2}\,\mathrm{cov}(M_{n,1}, M_{n,2})\sum_{1\leq i\neq j\leq n} x_i x_j\,. \tag{2.27}$$
- The distribution of $\overline{X}_n^*$, its moments, quantiles, etc., can be approximated by resampling, i.e., draw iid copies $\mathbf{M}_n(1), \ldots, \mathbf{M}_n(B)$ of $\mathbf{M}_n$, calculate
$$\overline{X}_n^*(i) = \frac{1}{n}\sum_{k=1}^{n} x_k\, M_{n,k}(i)\,, \quad i = 1, \ldots, B\,.$$

Then for example, for P-a.e. ω,

$$P^*(\overline{X}_n^* \le x) = \lim_{B\to\infty} \frac{1}{B} \sum_{i=1}^{B} I_{(-\infty,x]}(\overline{X}_n^*(i)) \quad P^* - \text{a.s.}$$

$$\text{var}^*(\overline{X}_n^*) = \lim_{B\to\infty} \frac{1}{B-1} \sum_{i=1}^{B} \left[\overline{X}_n^*(i) - E^*(\overline{X}_n^*(i))\right]^2 \quad P^* - \text{a.s.} \tag{2.28}$$

Notice that for the naive bootstrap the variance in (2.27) reduces to a version of the sample variance

$$n \,\text{var}^*(\overline{X}_n^*) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 = s_n^2 \,.$$

If one wants an improvement on Efron's bootstrap, a variance smaller than s_n^2 would be an indication of this better behavior. We check this for the Bayesian bootstrap.

Example 2.6 (Bayesian bootstrap)
Since iid random variables are exchangeable but condition (2.25) does not hold, a natural alternative is the following: let $Y_1, \ldots, Y_n$ be iid positive random variables and define the weights

$$n^{-1} M_{n,i} = \frac{Y_i}{Y_1 + \cdots + Y_n} \,, \quad i = 1, \ldots, n \,,$$

which are exchangeable for every n. If one wants to determine the variance of the bootstrap mean one needs to determine $\text{var}(M_{n,1})$ and $\text{cov}(M_{n,1}, M_{n,2})$, according to formula (2.27). This is in general difficult and then one relies on simulations, see (2.28). If the Y_i's are iid standard exponential, then the joint distribution of the $n^{-1}M_{n,k}$'s is known: it is the distribution of the spacings of $n-1$ uniformly on $(0,1)$ distributed random variables. Then $\mathbf{M}_n$ has a Dirichlet distribution. This led Rubin to the introduction of his *Bayesian bootstrap* and he also gives Bayesian arguments for the superiority of his method.

From an operational point of view the two methods are not different. For the Efron bootstrap,

$$\text{var}(M_{n,1}^{\text{E}}) = 1 - n^{-1} \quad \text{and} \quad \text{cov}(M_{n,1}^{\text{E}}, M_{n,2}^{\text{E}}) = -n^{-1} \,.$$

Rubin mentions that for his bootstrap

$$\text{var}(M_{n,1}) = (n-1)/(n+1) \quad \text{and} \quad \text{cov}(M_{n,1}, M_{n,2}) = -(n+1)^{-1} \,.$$

By (2.27), the bootstrap variances of the two methods are not too so much different, and it seems one does not gain much by the new method. □

Comments

Clearly, the wild bootstrap idea depends on the particular form of the sample mean. For the latter and for more general asymptotic linear functionals of the sample an asymptotic theory was developed in Mason and Newton [40] and in Mammen [39]. The idea of weighted bootstrap already appeared in work of Rubin [51] who used Dirichlet distributed weights for his *Bayesian bootstrap*, i.e., the distribution of the spacings of $n-1$ iid $U(0,1)$ random variables. For other choices of weights see the references in [40, 39]. The fact that the bootstrap sample mean is actually nothing but a randomly weighted mean of the sample was also observed by Künsch [37] in the context of the moving blocks bootstrap for dependent data. Later his student Bühlmann [13] exploited this idea to choose suitable random weights (which are not exchangeable but adjusted to the correlation structure of the underlying data) for modifying this method such that the variance of the estimators for the bootstrap of dependent data would be reduced.

2.4.2 Bootstrap curve estimation

The bootstrap idea has the potential to be applied to statistical problems with an underlying iid structure. We touch on the two problems: regression and curve estimation, and indicate how the bootstrap idea can be used in this context. We refer to Hall's book [32], Chapter 4, for an extensive discussion, including the drawbacks and pitfalls of the method. We omit discussing any of them.

Consider the *linear regression problem*

$$\mathbf{Y}_i = \mathbf{x}_i \mathbf{c} + \mathbf{d} + \mathbf{T}_i\,, \quad i = 1, \ldots, n\,, \tag{2.29}$$

where $\mathbf{T}_i$ are iid errors in $\mathbb{R}^d$, $\mathbf{d} \in \mathbb{R}^d$ is an unknown constant, $\mathbf{c} \in \mathbb{R}^q$ is an unknown deterministic vector and $\mathbf{x}_i$ is a $d \times q$ matrix of design points. Least squares estimation of $\mathbf{c}$ and $\mathbf{d}$ yields $\widehat{\mathbf{c}}$ and $\widehat{\mathbf{d}}$. This allows one to calculate the *residuals*

$$\widehat{\mathbf{T}}_i = \mathbf{Y}_i - \mathbf{x}_i \widehat{\mathbf{c}} - \widehat{\mathbf{d}}\,, \quad i = 1, \ldots, n\,.$$

Although not independent (via the least squares estimators they depend on all observations), one may hope that they are "almost independent" if the fit is reasonable. A bootstrap method can now be based on the naive bootstrap of the residuals by drawing iid random variables $\mathbf{T}_i^*$ from the empirical distribution function of the residuals. The bootstrap versions of $\mathbf{Y}_i$ are then obtained by mimicing the structure (2.29):

$$\mathbf{Y}_i^* = \mathbf{x}_i \widehat{\mathbf{c}} + \widehat{\mathbf{d}} + \mathbf{T}_i^*\,.$$

Thus one exploits the iid structure of the regression error in order to get a bootstrap method. This is similar to the bootstrap for ARMA processes as explained in Section 2.3. Based on the latter bootstrap versions one can derive bootstrap least squares estimators for $\mathbf{c}$ and $\mathbf{d}$.

Hidden iid structures can also be used for estimating the *density* f underlying iid data $X_1, \ldots, X_n$. A classical estimator of $f(x)$ is given by

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\,.$$

Since the non-negative function K is called a *kernel* $\widehat{f}(x)$ is referred to as a *kernel density estimator*. In order to get some intuition you can choose K as the density of the standard normal distribution. The parameter $h > 0$ is the bandwidth and has to be chosen such that $h = h_n \to 0$ at an appropriate rate. This ensures that the estimator becomes asymptotically unbiased and the condition $nh_n \to \infty$ ensures consistency of the estimator. For details we refer to Silverman [53] and Wand and Jones [60] as standard references on density estimation.

From the construction of $\widehat{f}(x)$ a naive bootstrap version would be given by

$$\widehat{f}^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right)\,,$$

where X_i^* are iid with the empirical distribution function of the data.

None of the naive methods mentioned works properly. If you want to apply them you need to correct the naive bootstrap. Read Hall [32].

2.5 Concluding remarks

Above we considered a technique which is considered as one of the breakthroughs of modern statistics. Although its basic idea is very simple — replace the distribution underlying the data by its empirical distribution function — the new dimension of Efron's bootstrap consists of the resampling idea, i.e., by drawing a large number of pseudo samples from the empirical distribution function and plugging these pseudo samples in the function of interest. This enables one to approximate the distribution of functions of the data and their characteristics (moments, quantiles, etc.) by a simple drawing mechanism, in principle, arbitrarily accurately.

The bootstrap idea was successful because its invention coincided with the introduction of new and powerful computers in the 1980s. It would have failed in the 70s and earlier. But already in the 1950s the resampling idea was somewhat popular. Efron always refers to a predecessor of his resampling idea, Quenouille's [45] *jackknife*, which was a driving force for the invention of the bootstrap. The potential of the jackknife for statistics was early on realized by Tukey [57] who also coined the name.

The *jackknife* is easily explained:

- Assume you have data $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$ from an iid sample with distribution F_θ and let $\widehat{\theta}_n = \widehat{\theta}(x_1, \ldots, x_n)$ be an estimator based on the data.
- Generate jackknife samples by leaving out the ith point in the sample

$$\mathbf{x}_i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\,, \quad i = 1, \ldots, n\,.$$

- Calculate n estimators of θ by plugging in the jackknife samples

$$\widehat{\theta}(1) = \widehat{\theta}_{n-1}(\mathbf{x}_1)\,, \ldots, \widehat{\theta}(n) = \widehat{\theta}_{n-1}(\mathbf{x}_n)\,.$$

- The jackknife estimators of bias and variance are then given by

$$\begin{aligned} \widehat{\text{bias}}_{\text{jackknife}} &= (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}(i) - \widehat{\theta}_n\right), \\ \widehat{\text{var}}_{\text{jackknife}} &= \frac{n-1}{n}\sum_{i=1}^{n}\left[\widehat{\theta}(i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}(i)\right]^2, \end{aligned}$$

The (unusual) choices of bias and variance estimators are motivated as follows. First, choose $\widehat{\theta}_n = \overline{x}_n$. Then the corresponding jackknife variance is given by $\widehat{\text{var}}_{\text{jackknife}} = (n(n-1))^{-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$, i.e., it yields the (usual) unbiased estimator of the variance of the sample mean. Second, take the (biased) estimator of the variance $\widehat{\theta}_n = n^{-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$. Then the corresponding jackknife bias is given by $\widehat{\text{bias}}_{\text{jackknife}} = -(n(n-1))^{-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$, yielding $\widetilde{\theta}_n = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$, the unbiased estimate of the variance. See Efron and Tibshirani [24], Chapter 11.

One of the obvious drawbacks of the method is that we get exactly as many jackknife samples as the sample size; it is not really a method one needs the computer for. For small n this may not give one enough confidence in the values $\widehat{\text{bias}}_{\text{jackknife}}$ and $\widehat{\text{var}}_{\text{jackknife}}$. Efron asked the question: "Which is better, the bootstrap or jackknife?" His answer was clearly very much in favor of the bootstrap, without giving too many theoretical details. His simulations for sample mean and sample correlation in the case of an iid standard normal sample of size $n = 10$ indicates that the bootstrap does better both as regards bias and variance. Intuitively, the bootstrap should perform better because we use

more information in the sample. It is difficult to compare the two methods directly in a fair way, in particular for small sample sizes. Various authors have studied the asymptotic performance of standard statistics for the bootstrap and the jackknife. A quick search in Mathematical Reviews gives hundreds of sources, including various books. Efron and Tibshirani [24] and Hall [32] discuss both methods and try to compare them.

After its introduction the bootstrap gained huge popularity among applied statisticians and practitioners which success story has not finished yet. As outlined by Efron, the bootstrap is intuitively appealing and one can calculate distributional characteristics of a large variety of statistics without too much sophistication and theory. This makes its success plausible.

The bootstrap poses plenty of difficult probabilistic questions. Since one deals with a complicated continuity problem — the original data in the statistic are replaced by computer generated pseudo samples with the empirical distribution function of the data — it is a priori not clear whether the bootstrap works, and we have seen plenty of pitfalls above. Therefore the bootstrap has become a domain of very abstract probability theory since the middle of the 1980s. As usual in statistical history, this leads to a multitude of theoretical papers which are not understood any more by applied statisticians. In particular, the link of the bootstrap with empirical process theory, represented for example by Dudley [20] or Giné and Zinn [26, 27], was fertile for a better theoretical understanding of the bootstrap, but also led to an overkill of the theory. It seems that the probabilistic interest in the bootstrap has calmed down over the last few years, whereas statisticians still use it as a tool, often if nothing else works and often without the existing probabilistic tools.

The bootstrap ideology is not uncontested among mathematical statisticians. Some of the drawbacks we have discussed:

- The superiority of the bootstrap as compared to the CLT is proved for large sample sizes, involving asymptotic expansions with random remainder term.
- There exist various pitfalls of the naive bootstrap. Depending on the statistic at hand one has to correct for bias, studentize, subsample,...
- The naive bootstrap fails for dependent data.

As regards the first item, it is certainly important to know that the bootstrap works in a given situation, i.e., that it does not do something completely unreasonable. In this sense the asymptotic theory is just an indication of this fact. Whether or not the bootstrap does better in some sense is perhaps an interesting theoretical question, but it does not really go to the point. It is obvious that the bootstrap uses much more of the information in the sample than any other method.

As regards the second point, a single method cannot solve all problems. This would be a miracle. As we have seen, knowledge about the reasons of pitfalls helps one to correct the estimators. One of the "pitfalls" were non-linear statistics, including U-statistics, and one had to correct the bias in a sophisticated form to make the bootstrap work. This example shows the real superiority of the bootstrap method. Indeed, non-linear statistics often have non-normal limit distribution which are difficult to calculate. Degenerate U-statistics, for example, have weighted χ^2 distributions as limits which can be expressed in terms of multiple stochastic integrals. In this case the bootstrap really gives a simple alternative to too much sophistication.

As regards the third item, the moving blocks bootstrap and other alternative bootstraps are an answer to this problem. Since the limit theory for the sample mean and more complicated statistics of dependent data is very sophisticated and often not applicable, for example the limiting variance in the CLT for the sample mean is not explicitly calculable in general, the modified bootstraps give confidence bands, variance and bias estimates.

As should have become transparent from the above discussion, the bootstrap is a useful *ad hoc* method which gives one answers to various problems. If one is not careful one gets a wrong answer. Taking into account all pros and cons, various positive facts can be booked on the side of the bootstrap. In addition to the fact that practitioners can hardly be convinced to abandon the method, the bootstrap has delivered a useful methodology for using the information contained in a sample beyond calculating the usual sample mean and sample variance. Even when no theory is available, the resampling idea can be used in an ad hoc sense to get cheap information about the distribution underlying complicated statistical functionals. The simplicity of the idea makes it applicable to more complicated objects such as multivariate data or spatial patterns with or without spatio-temporal dependence.

Comments

Everybody interested in the bootstrap should start with Efron's and Tibshirani's book [24] which is accessible without sophisticated knowledge. Davison and Hinkley [16] is another book aiming at the practitioner and applied statistician. The reader interested in the theory of the bootstrap can find plenty of references in the above sections. As a matter of fact, most of the references require prior knowledge of the methods of asymptotic statistics and weak convergence methods in probability theory. A recent reference to the bootstrap via the abstract theory of empirical processes can be found in van der Vaart and Wellner [59].

References

[1] Arcones, M.A. and Giné, E. (1989) The bootstrap of the mean with arbitrary bootstrap sample size. *Ann. Inst. Henri Poincaré* **25**, 457–481.

[2] Arcones, M.A. and Giné, E. (1992) On the bootstrap of U- and V-statistics. *Ann. Statist.* **20**, 655–674.

[3] Athreya, K.B. (1983) Strong law for the bootstrap. *Statistics Probab. Letters* **1**, 147–150.

[4] Athreya, K.B. (1987) Bootstrap for the mean in the infinite variance case. *Ann. Statist.* **15**, 724–731.

[5] Athreya, K.B. and Fuh, C.D. (1992) Bootstrapping Markov chains: countable case. *J. Statist. Plan. Inf.* **33**, 311–331.

[6] Bentkus, V., Götze, F. and Zwet, W. R. van (1997) An Edgeworth expansion for symmetric statistics. *Ann. Statist.* **25** 851–896.

[7] Bhattacharya, R.N. and Rao, R.R. (1976) *Normal Approximation and Asymptotic Expansions.* Wiley, New York.

[8] Bickel, P. and Freedman, D. (1981) Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–1217.

[9] Bickel, P.J., Götze, F. and Zwet, W.R. van (1997) Resampling fewer than n observations. *Statistica Sinica* **7**, 1–31.

[10] Billingsley, P. (1968) *Convergence of Probability Measures.* Wiley, New York.

[11] Bretagnolle, J. (1983) Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré* **19**, 281–296.

[12] Brockwell, P. and Davis, R.A. (1991) *Time Series: Theory and Methods,* 2nd edition. Springer, New York.

[13] Bühlmann, P. (1994) Blockwise bootstrapped empirical process for stationary sequences. *Ann. Statist.* **22**, 995–1012.

[14] Carlstein, E. (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171–1179.

[15] Datta, S. and McCormick, W.P. (1995) Some continuous Edgeworth expansion for Markov chains with application to bootstrap. *J. Multivar. Anal.* **52** (1995), 83–106.

[16] Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

[17] Deheuvels, P., Mason, D.M. and Shorack, G.R. (1993) Some results on the influence of extremes on the bootstrap. *Ann. Inst. Henri Poincaré* **29**, 83–103.

[18] Dehling, H. and Mikosch, T. (1994) Random quadratic forms and the bootstrap for U-statistics. *J. Multivariate Analysis* **51**, 392–413.

[19] Doukhan, P. (1994) *Mixing, Properties and Examples.* Lecture Notes in Statistics **85**. Springer, New York.

[20] Dudley, R.M. (1978) Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899–929.

[21] Dynkin, E.B. and Mandelbaum, A. (1983) Symmetric statistics, Poisson point processes, and multiple Wiener integrals. *Ann. Statist.* **11**, 739–745.

[22] Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

[23] Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans.* CRMS-NSF Regional Conference Series in Applied Mathematics. SIAM Philadelphia.

[24] Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap.* Chapman & Hall, New York.

[25] Embrechts, P., Klüppelberg, C. and Mikosch, T. *Modelling Extremal Events for Insurance and Finance.* Springer, Heidelberg.

[26] Giné, E. and Zinn, J. (1989) Necessary conditions for the bootstrap of the mean. *Ann. Statist.* **17**, 684–691.

[27] Giné, E. and Zinn, J. (1990) Bootstrappping general empirical measures. *Ann. Probab.* **18**, 651–869.

[28] Götze, F. and Hipp, C. (1983) Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrsch. verw. Gebiete* **64** 211–239.

[29] Guillou, A. (2000) Bootstrap confidence intervals for the Pareto index. *Commun. Statist.: Theory and Methods* **29**, 211–226.

[30] Gut, A. (1988) *Stopped Random Walk.* Springer. New York.

[31] Hall, P. (1990) Asymptotic properties of the bootstrap for heavy-tailed distributions. *Ann. Probab.* **18**, 1342–1360.

[32] Hall, P. (1992) *The Bootstrap and Edgeworth Expansion.* Springer Series in Statistics. Springer-Verlag, New York.

[33] Heimann, G. and Kreiss, J.-P. (1996) Bootstrapping general first order autoregression. *Statist. Probab. Lett.* **30**, 87–98.

[34] Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293–325.

[35] Ibragimov, I.A. and Linnik, Yu.V. (1971) *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen.

[36] Kreiss, J.-P. and Franke, J. (1992) Bootstrapping stationary autoregressive moving-average models. *J. Time Ser. Anal.* **13**, 297–317.

[37] Künsch, H.R. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217–1241.

[38] Lee, A.J. (1990) *U-Statistics: Theory and Practice.* Marcel Dekker, New York.

[39] Mammen, E. (1992) *When Does Bootstrap Work?* Asymptotic Results and Simulations. Lecture Notes in Statistics 77. Springer, New York.

[40] Mason, D.M. and Newton, M.A. (1992) A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.* **20**, 1611–1624.

[41] O'Brien, G.L. (1980) A limit theorem for sample maxima and heavy branches in Galton-Watson trees. *J. Appl. Probab.* **17**, 539–545.

[42] Paparoditis, E. and Politis, D.N. (2002) The local bootstrap for Markov processes. J. Statist. Plan. Inf. **108**, 301-328.

[43] Petrov, V.V. (1975) *Sums of Independent Random Variables.* Springer, Berlin.

[44] Petrov, V.V. (1995) *Limit Theorems of Probability Theory.* Oxford University Press, Oxford.

[45] Quenouille, M. (1949) Approximate tests of correlation in time series. *J. Royal Statist. Soc. B* **11**, 18–44.

[46] Rachev, S.T. (1991) *Probability Metrics and the Stability of Stochastic Models.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, Chichester.

[47] Radulović, D. (1996) The bootstrap for empirical processes based on stationary observations. *Stoch. Proc. Appl.* **65**, 259–279.

[48] Radulović, D. (2001) Renewal type bootstrap for Markov chains. Technical Report. Princeton University, Department of Mathematics.

[49] Radulović, D. (2002) On the bootstrap and empirical processes for dependent sequences. In: Dehling, H.G., Mikosch, T. and Sørensen, M. (Eds.) (2002) *Empirical Process Techniques for Dependent Data,* pp. 345–364. Birkhauser, Boston.

[50] Resnick, S.I. (1987) *Extreme Values, Regular Variation, and Point Processes.* Applied Probability. A Series of the Applied Probability Trust, 4. Springer, New York.

[51] Rubin, D. (1981) The Bayesian bootstrap. *Ann. Statist.* **9**, 130–134.

[52] Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

[53] Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[54] Singh, K. (1981) On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187–1195.

[55] Spitzer, F. (1976) *Principles of Random Walks.* Second edition. Graduate Texts in Mathematics, Vol. 34. Springer, New York.

[56] Swanepoel (1986) A note in proving that the (modified) bootstrap works. *Commun. Statist.: Theory and Methods* **15**, 3193–3203.

[57] Tukey, J.W. (1958) Bias and confidence in not quite large samples. *Ann. Math. Statist.* **29**, 614.

[58] Vaart, A.W. van der (1998) *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (UK).

[59] Vaart, A.W. van der and Wellner, J.A. *Weak Convergence and Empirical Processes.* Springer, New York.

[60] Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing.* Chapman and Hall, London.

[61] Zolotarev, V.M. (1997) *Modern Theory of Summation of Random Variables.* Modern Probability and Statistics. VSP, Utrecht.

3 Numerical solution to stochastic differential equations

In this section we want to consider approximations to the solution of Itô stochastic differential equations (SDE). The latter have gained a lot of popularity over the last view years since SDEs are used to describe the evolution of prices and interest rates in mathematical finance. Moreover, solutions to SDEs describe prices of financial derivatives such as options and futures. Only in a very few cases one is able to give explicit formulae — the celebrated Black–Scholes formula is one of the few cases when this is possible and definitely contributed to its popularity. In the majority of cases one is dependent on simulations for approximating sample paths and distributions of the objects of interest. In this section we will learn about some simple approximation methods. We start with Brownian motion which is the backbone of the whole theory of Itô integration. We will learn how to simulate Brownian sample paths. Then we will recall the notion of stochastic integral, just in order to understand what the differences and similarities with deterministic integrals are. Mimicing the definition of these integrals, we will find simple approximation techniques for solutions to SDEs. We also briefly discuss extensions to stochastic integrals driven by Lévy jump processes or more exotic processes such as fractional Brownian motion.

3.1 Brownian motion

Brownian motion is a fundamental process in probability theory and statistics. Its role in stochastic process theory is very much comparable to that of the normal distribution in the theory of random variables and random vectors. It appears directly or in hidden form in many applications, for example in the asymptotic theory of goodness of fit tests based on empirical process theory. Although Brownian motion had been used by Bachelier [3] in 1900 to give the first mathematical theory of finance or by Einstein [15] in 1905 to describe the movement of a particle in a suspended liquid, it was only Norbert Wiener [34] in 1923 who gained enough mathematical understanding of this stochastic process, i.e., he gave the first mathematically correct construction of the object "Brownian motion".

Recall that a stochastic process B on $[0,\infty)$ is said to be *standard Brownian motion* if it satisfies the following conditions:

- It starts at zero: $B_0 = 0$ a.s.
- It has independent and stationary increments: for any partition $0 = t_0 < t_1 < \cdots < t_n$, the increments

  $$\Delta_i B = B_{t_i} - B_{t_i - 1}\,, \quad i = 1, \ldots, n\,,$$

  are independent and their distribution only depends on the difference of the instants of time $\Delta_i = t_i - t_{i-1}$:

  $$\Delta_i B \stackrel{d}{=} B_{\Delta_i}\,.$$

- For every t, B_t is $N(0,t)$ distributed.
- Almost every sample path of B is continuous.

We also use the notion of (standard) Brownian motion if we consider the restriction of the process to an interval. We usually skip the "standard" and refer to Brownian motion.

The above definition of Brownian motion is not optimal in the sense that one can skip or modify some of the conditions. For example, if one assumes that B has independent and stationary increments with normal distributions, then there always exists a version of B which has a.s. continuous sample paths. Alternatively, if the process B has independent and stationary increments and a.s. continuous sample paths, its marginal distributions must be normal. This follows from the general theory of *Lévy processes*, i.e., processes with independent and stationary increments.

The stationary and independent increments force Brownian sample paths to behave in a very unusual way: they are continuous, but non-differentiable a.e. This forces the sample path to change its direction at every instant of time in an unpredictable way. Since Brownian motion and financial processes have this property in common, it might be one explanation for the popularity of the process. Another argument for Brownian motion is that it has nice theoretical properties, and mathematical tractability, not so much the desire of having a realistic model, is often responsible for its use in various applied areas.

We mention another important property of Brownian motion: it is self-similar with exponent 0.5. This means that for any $0 \le t_1 < \cdots < t_n$ and $c > 0$,

$$(B_{ct_1}, \ldots, B_{ct_n}) \stackrel{d}{=} c^{0.5}(B_{t_1}, \ldots, B_{t_n})\,.$$

Self-similarity is a property which Brownian motion shares with many other processes, for example fractional Brownian motion or stable Lévy motion. However, the exponent 0.5 on the right-hand side has then usually to be replaced by other positive numbers. This property has an obvious practical application. According to Kolmogorov's consistency theorem the distribution of a stochastic process is determined via its finite-dimensional distributions. Therefore, in order to simulate a Brownian path on $[0, c]$, say, it suffices to generate one on $[0, 1]$, then stretch the time axis by the factor c and rescale the path with $c^{0.5}$. Then we have a path of Brownian motion on $[0, c]$.

Since the finite-dimensional distributions of Brownian motion are Gaussian, it is a Gaussian process. It has mean function identical to zero and therefore its finite-dimensional distributions are completely determined through the covariance function

$$\operatorname{cov}(B_t, B_s) = \min(t, s)\,, \quad t, s \ge 0\,. \tag{3.1}$$

Thus, an alternative definition of Brownian motion would be as follows: it is a mean zero Gaussian stochastic process on $[0, \infty)$ with a.s. continuous sample paths and covariance function (3.1). We mention that the form of the covariance function (3.1) implies that there exists a version of the stochastic process which has a.s. continuous sample paths. This follows from a classical result of Kolmogorov.

In what follows, we want to learn about some methods for the approximation of Brownian sample paths.

3.1.1 Almost sure representations of Brownian motion

Almost sure representations of Brownian motion were among the first attempts to get a constructive description of Brownian sample paths. Such representations are given as infinite series of stochastic processes, i.e., they are the additive ingredients to the process and their superposition shows nicely how a Brownian sample path is built up.

The following representation is referred to as *Lévy–Ciesielski representation*; see Ciesielski [10]. Let (Z_i) be iid standard normal random variables and (ϕ_k) be a complete orthonormal system with respect to the L_2 inner product on $[0, 1]$. Define the integrated orthonormal functions

$$\psi_k(t) = \int_0^t \phi_k(s)\, ds\,, \quad t \in [0, 1]\,.$$

The latter are in general not orthogonal, an exception being the trigonometric functions. Then the infinite series

$$B_t = \sum_{k=0}^{\infty} \psi_k(t)\, Z_k\,, \quad t \in [0,1]\,, \tag{3.2}$$

converges uniformly on $[0,1]$ for a.e. realization of the iid sequence (Z_k), and the limit represents Brownian motion. A proof of this fact is beyond the scope of this course. However, some properties of Brownian motion can be read off: $B_0 = 0$ and B has continuous paths since the uniform limit of continuous functions is continuous. Moreover, for fixed t, B_t is Gaussian since it is a superposition of independent normal random variables. Further properties of Brownian motion can be seen if one considers special cases.

Example 3.1 (Fourier decomposition of Brownian motion, Paley–Wiener representation)
One of the first representations of type (3.2) is due to Paley and Wiener, two of the most influential mathematicians of the first half of the 20th century. They proved the existence of complex-valued Brownian motion through an infinite series of type (3.2) with the system of complex trigonometric functions $\phi_k(t) = (2\pi)^{-1/2}\mathrm{e}^{\,ikt}$ on $[0,2\pi]$. Then, by projecting, one obtains a representation of standard Brownian motion of type (3.2). This leads to the Paley–Wiener representation of Brownian motion as a special case of (3.2):

$$B_t = Z_0\,\frac{t}{\sqrt{2\pi}} + \frac{2}{\sqrt{\pi}} \sum_{n=1}^{\infty} Z_n\,\frac{\sin(nt/2)}{n}\,, \quad t \in [0,2\pi]\,. \tag{3.3}$$

Here $Z_0, Z_1, \ldots,$ are iid standard normal. Formally calculating the variance of the right-hand side for fixed t fixed we obtain

$$\mathrm{var}(B_t) = \frac{t^2}{2\pi} + \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin^2(nt/2)}{n^2}\,. \tag{3.4}$$

The right-hand side is finite and therefore we conclude that the infinite series (3.3) converges a.s. for every fixed t, and a characteristic function argument shows that the limit is Gaussian. In order to show that the convergence is uniform one has to work harder. We omit details and refer to Hida [19].

Using that $\sin^2(x) = 0.5(1-\cos(2x))$ and

$$\sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2} = \frac{\pi^2}{6} - \frac{\pi x}{2} + \frac{x^2}{4}\,, \quad 0 \le x \le 2\pi\,, \tag{3.5}$$

(see Gradshteyn and Ryzhik [17], p. 39, FI III 547), it is not difficult to see that the right-hand side in (3.4) equals t. Thus the variance is the one of Brownian motion at time t.

Formally calculating the covariance for $s < t$, say, and using $\sin(x)\sin(y) = 0.5[\cos(x-y) - \cos(x+y)]$, we have

$$\begin{aligned} \mathrm{cov}(B_s, B_t) &= \frac{ts}{2\pi} + \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin(nt/2)\,\sin(ns/2)}{n^2} \\ &= \frac{ts}{2\pi} + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{\cos(n(t-s)/2) - \cos(n(t+s)/2)}{n^2}\,. \end{aligned}$$

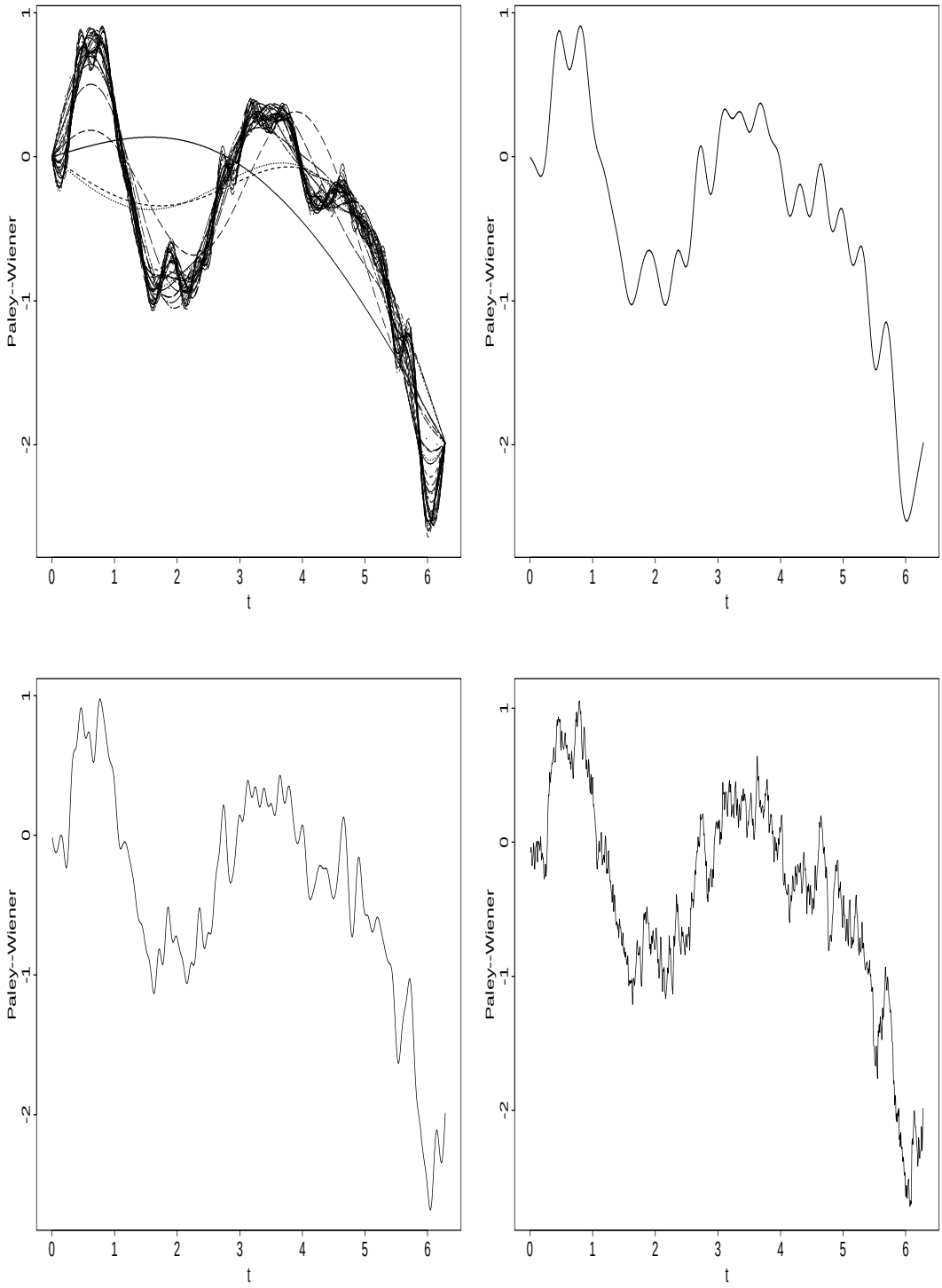

Figure 3.2 *Simulation of one Brownian sample path from the discretization* (3.6) *of the Paley–Wiener representation with* $N = 1,000$. Top left: *all paths for* $M = 2, \ldots, 40$. Top right: *the path only for* $M = 40$. Bottom left: $M = 100$. Bottom right: $M = 800$.

Now one can use (3.5) to conclude that $\operatorname{cov}(B_s, B_t) = \min(s,t)$, i.e., it is the covariance of Brownian motion. Notice that the above calculations were *formal* since we would have to justify the interchange of expectation and infinite series which is not obvious since the series (3.3) does not converge absolutely. Since we know that each of the infinite series for B_s and B_t converges a.s. and represents a Gaussian random variable, thus has all power moments finite, it is not difficult to show that

$$\sum_{n=1}^{M} Z_n \frac{\sin(nt/2)}{n} \sum_{k=1}^{N} Z_k \frac{\sin(ks/2)}{k}, \quad M, N \geq 1,$$

is uniformly integrable and therefore the above calculations are justified. Alternatively, use the L^2 convergence of the series (3.3) and the continuity of the inner product in in L^2.

The decomposition (3.3) was clearly born out of the idea to get a Fourier series representation of B which process has a.s. continuous paths. Since the resulting process is stochastic, the Fourier coefficients are random.

For practical purposes one has to truncate the infinite series (3.3) at a finite point M, say :

$$B_t^{(M)} = Z_0 \frac{t}{\sqrt{2\pi}} + \frac{2}{\sqrt{\pi}} \sum_{n=1}^{M} Z_n \frac{\sin(nt/2)}{n}.$$

Moreover, one can calculate this continuous function only at discrete instants of time. It would be natural to choose

$$t = t_i = \frac{i}{N} 2\pi, \quad i = 0, \ldots, N. \tag{3.6}$$

In order to get an impression of the error one encounters one would like to estimate the order of magnitude of the remainder term

$$B_t - B_t^{(M)} = \frac{2}{\sqrt{\pi}} \sum_{n=M+1}^{\infty} Z_n \frac{\sin(nt/2)}{n}.$$

It is a hard task to say something precise about a uniform estimate of the remainder term, i.e., about

$$\sup_{0 \leq t \leq 2\pi} |B_t - B_t^{(M)}|.$$

But it is also not impossible to solve this problem if one goes into the literature on suprema of Gaussian processes or Gaussian infinite series with values in the space of continuous functions. We refer to Adler [1] for the first approach and Ledoux and Talagrand [22] for the second one. These approaches would give one exponential estimates for the tail of the distribution of $\sup_{0 \leq t \leq 2\pi} |B_t - B_t^{(M)}|$, containing some unknown constants. Alternatively, one can prove a.s. approximation results for the remainder $B_t - B_t^{(M)}$ in some function space. This was done in Hall [18] for the uniform norm of $B - B^{(M)}$ and in Mikosch [25] for L^p and more exotic norms. A typical result says for the uniform norm that

$$\limsup_{M \to \infty} \sqrt{\frac{M}{\log M}} \sup_{0 \leq t \leq 1} |B_t - B_t^{(M)}| \leq \text{const} < \infty \quad \text{a.s.}$$

This is also not "too precise" since the rate depends on the realization of the Z_t's (or on ω).

A simple minded but perhaps more intuitive and useful approach is to focus on pointwise convergence. A measure of the quality of convergence for fixed t is

$$\mathrm{var}(B_t - B_t^{(M)}) = \frac{4}{\pi} \sum_{n=M+1}^{\infty} \frac{\sin^2(nt/2)}{n^2}\,.$$

Since $B_t - B_t^{(M)}$ is mean zero Gaussian this variance gives as an impression of the magnitude of the remainder term $B_t - B_t^{(M)}$ for fixed t. Typically, we would calculate the variance at the right end point of the interval of interest. We could choose M so large that the error $\mathrm{var}(B_t - B_t^{(M)})$ falls below a chosen small threshold $10^{-6}, 10^{-7}, \ldots$, say. □

Example 3.3 (The Lévy representation)
The Paley–Wiener representation is just one out of infinitely many possible series representations of Brownian motion. Another well-known such representation is due to Lévy, one of the founders of modern probability theory. In the *Lévy representation*, the sine functions are replaced by certain polygonal functions (the Schauder functions).

To be precise, first define the *Haar functions* H_n on $[0,1]$ as follows:

$$\begin{aligned}
H_1(t) &= 1\,,\\
H_{2^{m+1}}(t) &= \begin{cases} 2^{m/2}\,, & \text{if} \quad t \in \left[1 - \dfrac{2}{2^{m+1}}, 1 - \dfrac{1}{2^{m+1}}\right)\,,\\ -2^{m/2}\,, & \text{if} \quad t \in \left[1 - \dfrac{1}{2^{m+1}}, 1\right]\,,\\ 0\,, & \text{elsewhere}\,, \end{cases}\\
H_{2^m+k}(t) &= \begin{cases} 2^{m/2}\,, & \text{if} \quad t \in \left[\dfrac{k-1}{2^m}, \dfrac{k-1}{2^m} + \dfrac{1}{2^{m+1}}\right)\,,\\ -2^{m/2}\,, & \text{if} \quad t \in \left[\dfrac{2k-1}{2^{m+1}}, \dfrac{k}{2^m}\right)\,,\\ 0\,, & \text{elsewhere}\,, \end{cases}\\
& \qquad k = 1, \ldots, 2^m - 1\,; \quad m = 0, 1, \ldots\,.
\end{aligned}$$

From these functions define the system of the *Schauder functions* on $[0,1]$ by integrating the Haar functions:

$$\widetilde{H}_n(t) = \int_0^t H_n(s)\, ds\,, \quad n = 1, 2, \ldots\,.$$

Figures 3.4 and 3.5 show the graphs of H_n and $\widetilde{H}_n$ for the first n.

A series representation for a Brownian sample path on $[0,1]$ is then given by

$$B_t = \sum_{n=1}^{\infty} Z_n\, \widetilde{H}_n(t)\,, \quad t \in [0,1]\,, \tag{3.7}$$

where the convergence of this series is uniform for $t \in [0,1]$ and the Z_n's are realizations of an iid $N(0,1)$ sequence (Z_n). As for simulations of Brownian motion via sine functions, one has to choose a truncation point M of the infinite series (3.7). In Figure 3.6 we show how a Brownian sample path is approximated by the superposition of the first M terms in the series representation (3.7).

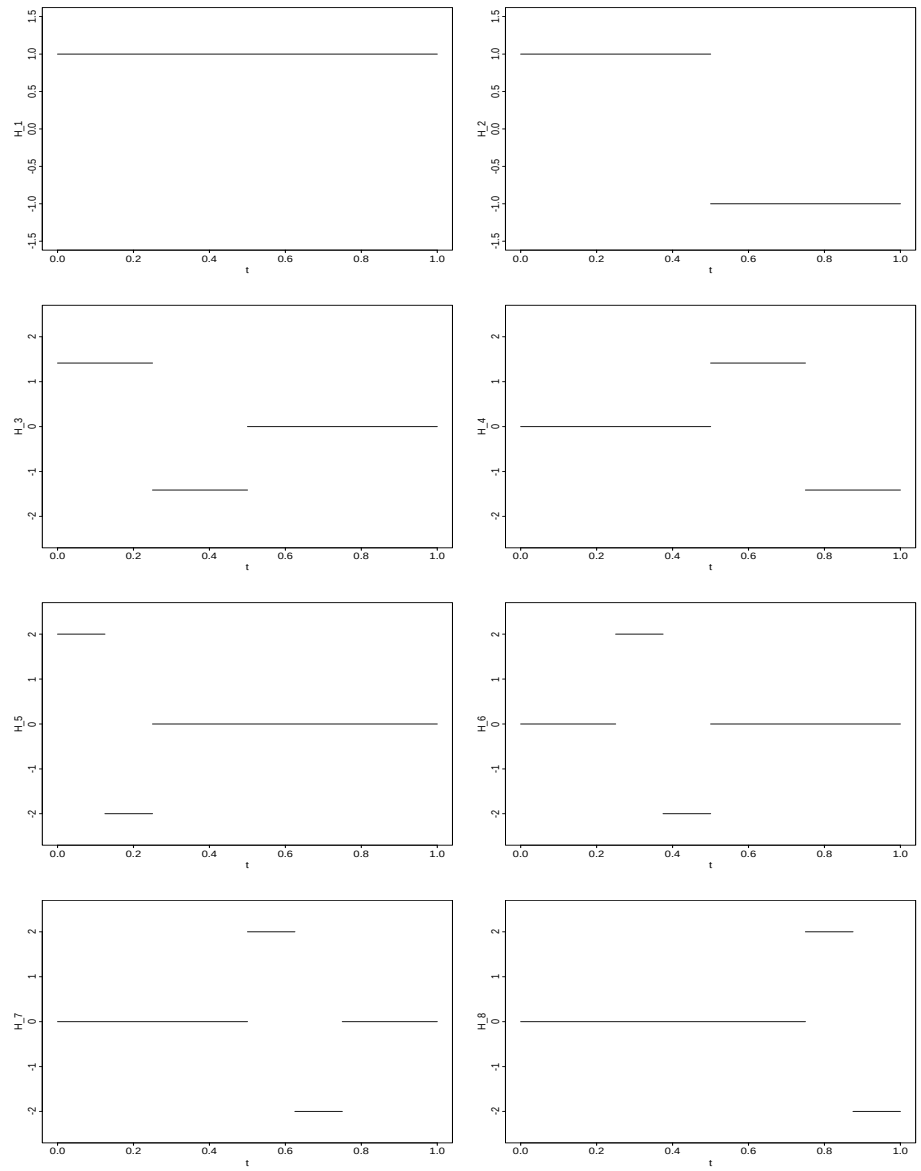

Figure 3.4 *The Haar functions* $H_1, \ldots, H_8$.

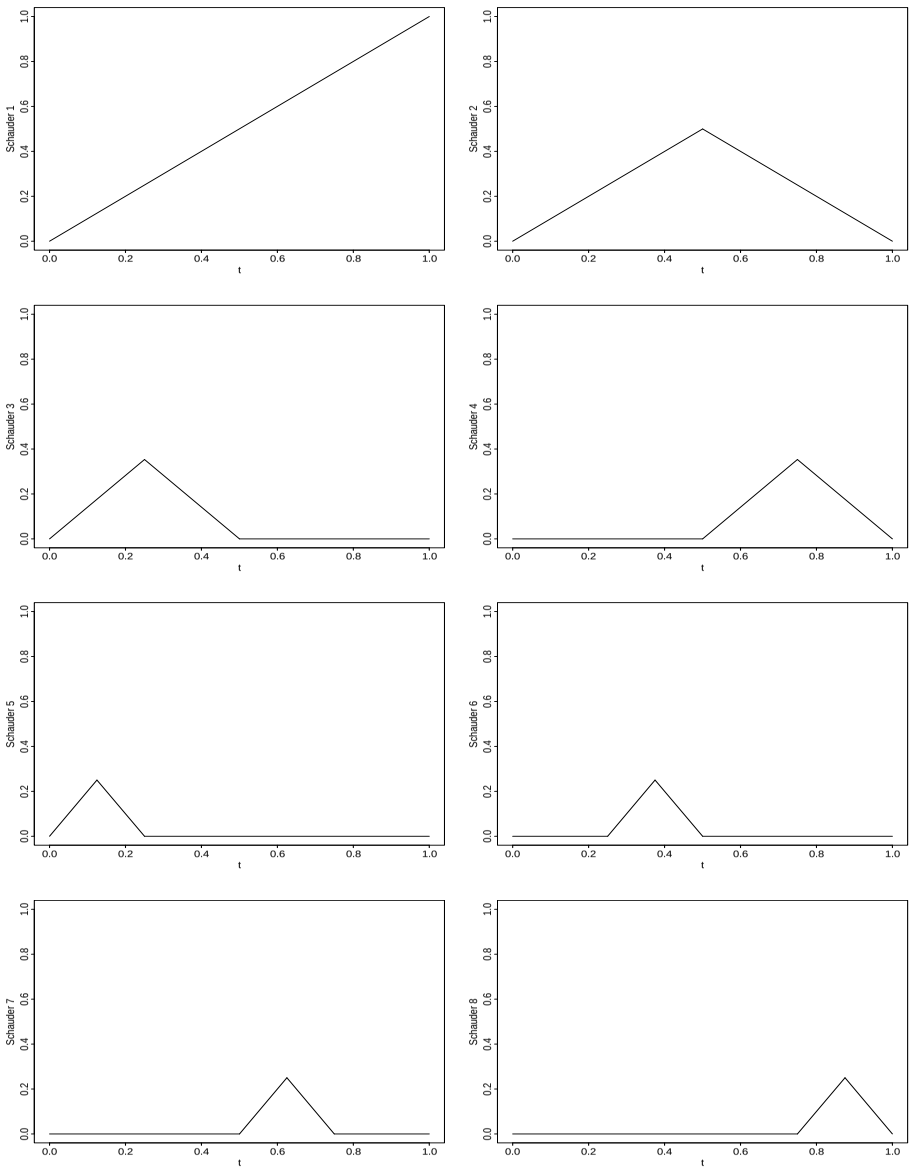

Figure 3.5 *The Schauder functions* $\widetilde{H}_1, \ldots, \widetilde{H}_8$.

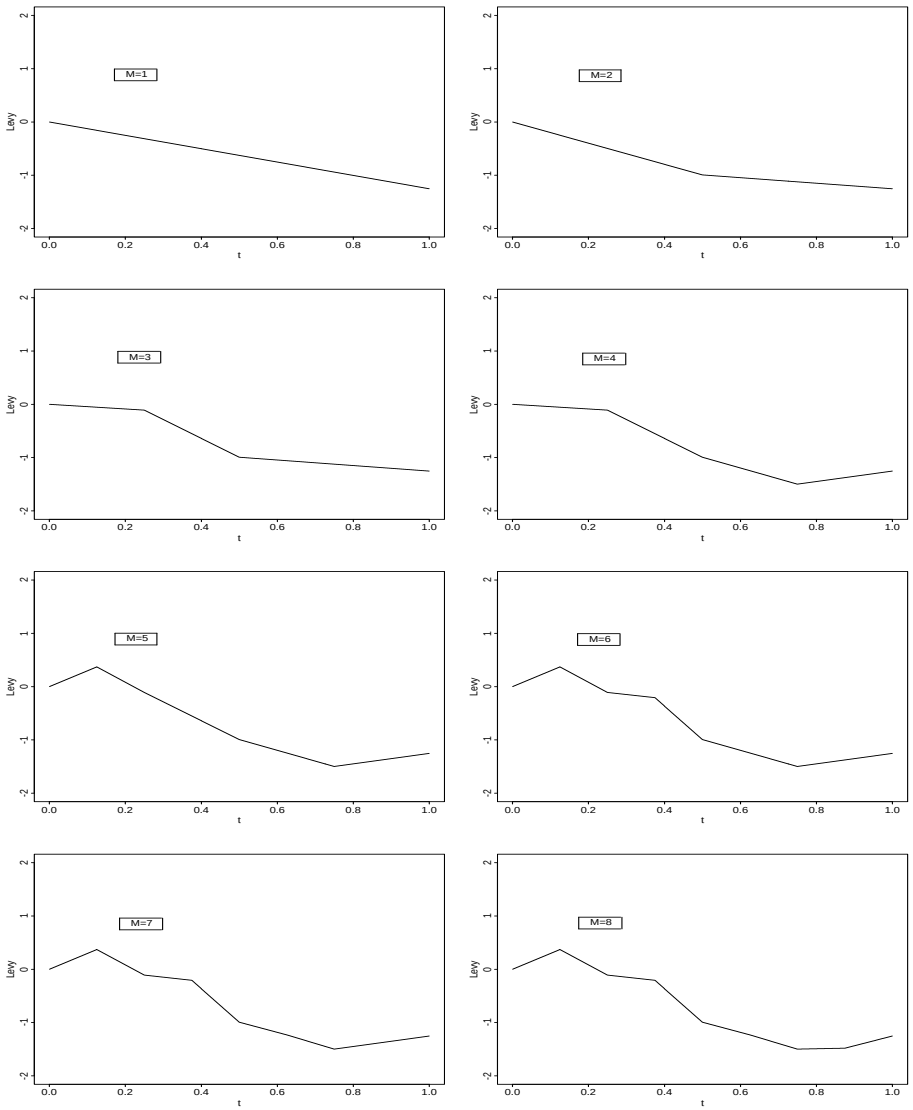

Figure 3.6 *The first steps in the construction of one Brownian sample path from the Lévy representation* (3.7) *via* M *Schauder functions,* $M = 1, \ldots, 8$.

In contrast to Figure 3.2, the polygonal shape of the Schauder functions already anticipates the irregular behavior of a Brownian path (its non-differentiability) for relatively small M.

The Lévy representation is computationally easier if one restricts the calculation to the points $t_{k,m} = k2^{-m}$ and increases M in powers of 2. Notice that summation at these instants of time stops when m is increased by 1, i.e., after a finite number of steps one has the final value of the infinite series at the points $t_{k,m}$. Thus one does not encounter an error at these points.

As for the Paley–Wiener representation, one can give tail estimates for the distribution of the supremum of the remainder term and derive a.s. bounds for it. We refer to the same references as above. □

3.1.2 Distributional approximations

For practical purposes it most often suffices to have a distributional approximation to Brownian motion. This means that we have stochastic processes ξ_n on $[0,1]$, say, satisfying $\xi_n \xrightarrow{d} B$. The latter relation is easily written down but it would take us quite some time to explain in detail what *convergence in distribution of stochastic processes* actually means. First of all, the values of a stochastic process on $[0,1]$ are functions. Thus the convergence would very much depend on the function space where ξ_n assumes its values. For our purposes it would suffice to choose ξ_n with values in the space $\mathbb{C}[0,1]$ of continuous functions on $[0,1]$ or in the Skorokhod space $\mathbb{D}[0,1]$ of càdlàg functions (the function has limits from the left in $(0,1]$ and is continuous from the right in $[0,1)$). The latter space is quite natural for many purposes in probability theory. For example, any distribution function with support on $[0,1]$ is an element of $\mathbb{D}[0,1]$. Then convergence in distribution $\xi_n \xrightarrow{d} B$ means that $P(\xi_n \in A) \to P(B \in A)$ for Borel sets A (i.e., an element of the σ-field generated from the open/closed sets in the function space equipped with an appropriate metric or topology) with $P(B \in \partial(A)) = 0$. The distribution of B on $\mathbb{C}[0,1]$ is called *Wiener measure*, and it was indeed Norbert Wiener in 1923 who proved the existence of this measure, as a main step in order to prove the existence of Brownian motion, see the reference on p. 77. It is quite obvious that we would need a course on weak convergence of probability measures.

The distributional convergence $\xi_n \xrightarrow{d} \xi$ in some function space requires two conditions and is equivalent to this convergence in $\mathbb{C}[0,1]$ and $\mathbb{D}[0,1]$:

1. Convergence of the finite-dimensional distributions: for any choice of $0 \le t_1 < \cdots < t_m \le 1$,

   $$(\xi_n(t_1), \ldots, \xi_n(t_m)) \xrightarrow{d} (\xi_{t_1}, \ldots, \xi_{t_m}) \,. \tag{3.8}$$

   (We write $\xi_n \xrightarrow{fidi} \xi$.)

2. Tightness of (ξ_n): probability mass should not disappear from compact sets in the function space as $n \to \infty$.

Fidi-convergence is something we are familiar with: it is just convergence in distribution of random vectors. It ensures that the processes ξ_n at any finite set of instants of time converge in distribution to the right limits. Tightness is something we might not be familiar with. It means, roughly speaking, that the process should not become too wild between fixed instants of time. For the processes we will consider this does not present a problem. Notice that the distribution of a stochastic process is determined through its finite-dimensional distributions, but for distributional convergence of stochastic processes the finite-dimensional distributions are not sufficient.

In order to explain what can happen when the sequence of processes is *not tight* we consider a simple example.

Example 3.7 (A counterexample to tightness)
One of the reasons why convergence in distribution $\xi_n \xrightarrow{d} \xi$ for stochastic processes $\xi, \xi_1, \xi_2, \ldots$ is an important notion is that $h(\xi_n) \xrightarrow{d} h(\xi)$ for continous functions h, see Section 1.1.6.
Let X_i be iid random variables with a common Fréchet distribution, i.e.,

$$P(X_1 \le x) = \mathrm{e}^{-x^{-\alpha}}, \quad x > 0,$$

for some $\alpha > 0$. It is easy to see that

$$n^{-1/\alpha} \max_{i=1,\ldots,n} X_i \stackrel{d}{=} X_1 . \tag{3.9}$$

Define the stochastic process ξ_n with paths in $\mathbb{C}[0,1]$ as follows:

$$\xi_n(t) = \begin{cases} 0 & t = 0, \\ n^{-1/\alpha} X_i & t = i/n, i = 1, \ldots, n, \\ \text{linearly interpolated} & \text{elsewhere}. \end{cases}$$

Then $\xi_n(t) \xrightarrow{P} 0$. Indeed, assume that $t \in n^{-1}[i-1, i]$, then for $x > 0$,

$$P(\xi_n(t) > x) \le P(n^{-1/\alpha} \max(X_{i-1}, X_i) > x) = P(n^{-1/\alpha} \max(X_1, X_2) > x) \to 0 .$$

Thus, since the distribution of a process is determined through its finite-dimensional distributions, the limiting process of the ξ_n's, if it existed, would be $\xi \equiv 0$. The distributional convergence $\xi_n \xrightarrow{d} \xi$ would imply that

$$\begin{aligned} h(\xi_n) &= \sup_{0 \le t \le 1} \xi_n(t) = n^{-1/\alpha} \max_{i=1\ldots,n} X_i \\ \xrightarrow{d}\ h(\xi) &= \sup_{0 \le t \le 1} \xi(t) = 0, \end{aligned}$$

since the supremum functional h is a continuous mapping from $\mathbb{C}[0,1]$ to $\mathbb{R}$. On the other hand, we know that this is wrong by virtue of (3.9). The reason for this failure is non-tightness of (ξ_n) in $\mathbb{C}[0,1]$: although the process converges to a limit at any finite set of instants of time, this fact does not control its behavior between these instants of time. Due to the independence of the X_i's, the process in between these instants of time is getting wilder and wilder; these increasing oscillations make that the continuous processes ξ_n do not converge in distribution to a continuous limit. □

The main reason for the badly behaved processes of the previous examples is the lack of dependence in the processes ξ_n. This kind of erratic behavior does not occur for the following process. Let $X, X_1, X_2, \ldots$, be iid non-degenerate random variables and

$$S_0 = 0, \quad S_n = X_1 + \cdots + X_n, \quad n \ge 1,$$

be the *random walk process* constructed from it. Clearly, (S_n) does not converge in any sense. The CLT tells us that

$$\frac{1}{\sqrt{n}\sigma} (S_n - n\, EX) \xrightarrow{d} N(0,1)$$

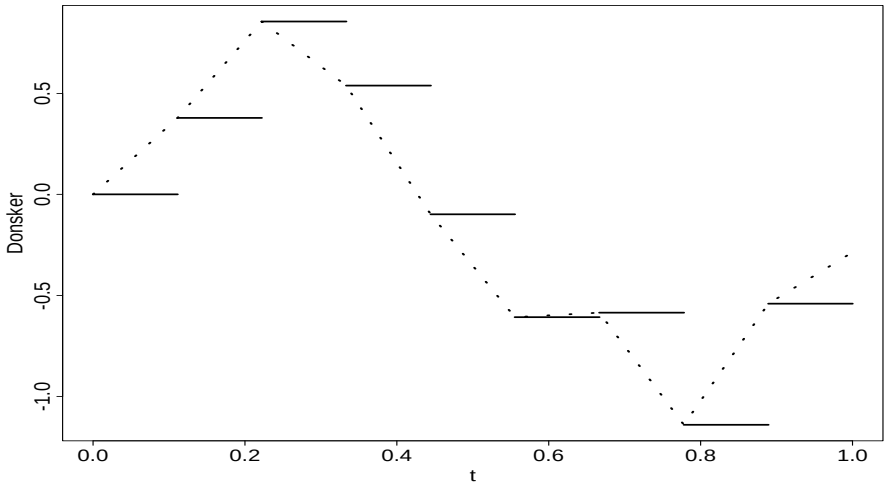

Figure 3.8 *One sample path of the processes* S_9 (*dotted line*) *and* $\widetilde{S}_9$ (*solid line*) *for the same sequence of realizations* $X_1(\omega), \ldots, X_9(\omega)$.

provided that $\sigma^2 = \mathrm{var}(X) < \infty$. Without loss of generality assume $EX = 0$ and $\sigma^2 = 1$. We construct the following continuous processes on $[0,1]$ from (S_n)

$$S_n(t) = \begin{cases} (\sqrt{n})^{-1} S_k & t = k/n\,, \\ \text{linearly interpolated} & \text{elsewhere.} \end{cases}$$

Define the process on $[0,1]$

$$\widetilde{S}_n(t) = (\sqrt{n})^{-1} S_{[nt]}\,,$$

where $[\cdot]$ denotes integer part. In order words $\widetilde{S}_n(n^{-1}k) = S_n(n^{-1}k)$ for $k = 0, \ldots, n$ and it is constant on $n^{-1}[k-1, k)$. Notice that the paths of $\widetilde{S}_n$ are càdlàg, thus in $\mathbb{D}[0,1]$. So is the process $S_n - \widetilde{S}_n$. Moreover,

$$\sup_{0 \le t \le 1} |S_n(t) - \widetilde{S}_n(t)| \le (\sqrt{n})^{-1} \max_{i=1,\ldots,n} |X_i|\,. \tag{3.10}$$

It is not difficult to see that the right-hand side converges to zero in probability since $\mathrm{var}(X) < \infty$. Hence, if one is interested in the limit properties of S_n in a distributional sense, it does not matter whether one considers S_n or $\widetilde{S}_n$, although the membership in $\mathbb{C}[0,1]$ for S_n and in $\mathbb{D}[0,1]$ for $\widetilde{S}_n$ makes *formally* quite a difference.

In order to study fidi-convergence it suffices by (3.10) to consider the finite-dimensional distributions of $\widetilde{S}_n$. Then for $0 = t_0 = t_1 < \cdots < t_m$,

$$\begin{aligned} &\left[\widetilde{S}_n(t_1), \ldots, \widetilde{S}_n(t_m)\right] \\ &= [\widetilde{S}_n(t_1), \widetilde{S}_n(t_1) + (\widetilde{S}_n(t_2) - \widetilde{S}_n(t_1)), \ldots, \\ &\quad \widetilde{S}_n(t_1) + (\widetilde{S}_n(t_2) - \widetilde{S}_n(t_1)) + \cdots + (\widetilde{S}_n(t_m) - \widetilde{S}_n(t_{m-1}))]\,. \end{aligned}$$

The ordinary CLT yields that

$$\widetilde{S}_n(t_i) - \widetilde{S}_n(t_{i-1}) \xrightarrow{d} N(0, t_i - t_{i-1}) \stackrel{d}{=} B_{t_i} - B_{t_{i-1}}\,.$$

Since the increments $\widetilde{S}_n(t_i) - \widetilde{S}_n(t_{i-1})$, $i = 1, \ldots, m$, are independent, a characteristic function argument shows that they converge jointly in distribution as $n \to \infty$ to the corresponding independent increments of Brownian motion: $B_{t_i} - B_{t_{i-1}}$, $i = 1, \ldots, m$. Hence by a continuous mapping theorem argument, both $S_n \stackrel{fidi}{\to} B$ and $\widetilde{S}_n \stackrel{fidi}{\to} B$ hold. It is possible to show that $S_n \stackrel{d}{\to} B$ in $\mathbb{C}[0,1]$ and $\widetilde{S}_n \stackrel{d}{\to} B$ in $\mathbb{D}[0,1]$, by proving tightness in these spaces.

Thus we have learned that the partial sum processes S_n and $\widetilde{S}_n$ converge in distribution to standard Brownian motion on $[0,1]$. These constructions are certainly the cheapest way of constructing an approximation to Brownian motion, and for practical purposes one calculates realizations of S_n or $\widetilde{S}_n$ for large n. It is no problem to choose $n = 10^4$, 10^5 or $n = 10^6$; on a modern computer we get a realization of these processes within microseconds.

Notice that the convergence results are independent of the distribution of the underlying X_i's. All we need is a finite second moment for their distribution. For this reason, the relations $S_n \stackrel{d}{\to} B$ and $\widetilde{S}_n \stackrel{d}{\to} B$ are often referred to as *invariance principles*. They are named after the first person who gave proofs for these results in 1951, M. Donsker [11]. The question as to which distribution of X gives the best approximation is clearly answered: it is the standard normal distribution. In this case, the distribution of $(S_n(n^{-1}i))_{i=0,\ldots,n}$ coincides with the distribution of $(\widetilde{S}_n(n^{-1}i))_{i=0,\ldots,n}$ and the distribution of $(B_{n^{-1}i})_{i=0,\ldots,n}$. If one has to simulate a huge number of Brownian sample paths, computer time can be an issue. Then one often takes a Bernoulli distribution for X or a symmetric three point distribution around zero. The theoretical results do not change but then we may expect that the convergence rates are not the best. This follows from an application of the Berry–Esséen inequality which in the above mentioned cases gives the pessimistic best rate of $1/\sqrt{n}$ in the CLT for $S_n(1) = \widetilde{S}_n(1)$ which is a special case of the invariance principle.

The *Donsker invariance principle* is also often referred to as *functional central limit theorem*. It is clearly motivated by the similarity to the CLT and the fact that for any continuous functional h on $\mathbb{C}[0,1]$ or $\widetilde{h}$ on $\mathbb{D}[0,1]$ the continuous mapping theorem (cf. Section 1.1.6) implies that $h(S_n) \stackrel{d}{\to} h(B)$ and $\widetilde{h}(\widetilde{S}_n) \stackrel{d}{\to} \widetilde{h}(B)$. In particular, it implies that

$$\begin{aligned} \sup_{0\le t\le 1} S_n(t) &= \frac{1}{\sqrt{n}} \max_{i=1,\ldots,n} S_i \stackrel{d}{\to} \sup_{0\le t\le 1} B_t\,, \\ \int_0^1 (\widetilde{S}_n(t))^2\,dt &= \frac{1}{n^2}\sum_{i=1}^n S_i^2 \stackrel{d}{\to} \int_0^1 B_t^2\,dt\,, \end{aligned}$$

and many other interesting functionals. Thus, by first simulating independent repetitions of the processes S_n or $\widetilde{S}_n$ for large n, one can approximate the distribution of functionals $h(B)$ or $\widehat{h}(B)$ arbitrarily close.

The continuous mapping theorem remains valid for a.s. continuous functions h, i.e., if $\xi_n \stackrel{d}{\to} \xi$ and the set D_h of discontinuity points of h has P_ξ probability 0, then $h(\xi_n) \stackrel{d}{\to} h(\xi)$. Notice that this statement is reasonable since h is defined on the set of values of ξ and therefore it makes sense to require $P_\xi(D_h) = P(\xi \in D_h) = 0$. This extension of the continuous mapping theorem might look exotic at a first glance but it is extremely useful.

For example, assume as before $S_n \stackrel{d}{\to} B$ and let $A \subset \mathbb{C}[0,1]$ be a Borel set. Then the indicator function I_A is a.s. continuous provided $P_B(\partial(A)) = P(B \in \partial(A)) = 0$. Indeed, the only points where the step function I_A can be discontinuous are the points of the boundary and this set has Wiener measure zero. Sets A of interest can be, for example,

$$A = \{f \in \mathbb{C}[0,1] : f(t) < y(t) \text{ for some } t \le 1\} \quad \text{and} \quad y(t) = (\log x(t) - ct)/\sigma\,,$$

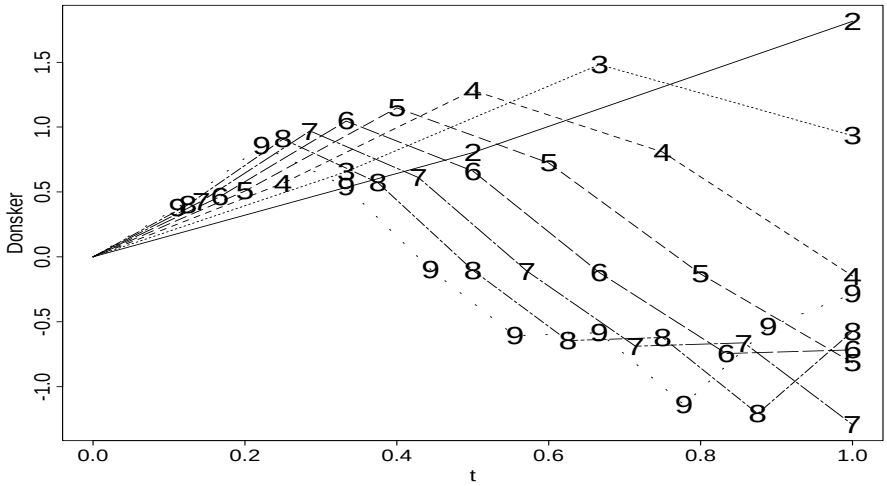

Figure 3.9 *Sample paths of the process S_n for one sequence of standard normal realizations $X_1(\omega), \ldots, X_9(\omega)$ and $n = 2, \ldots, 9$.*

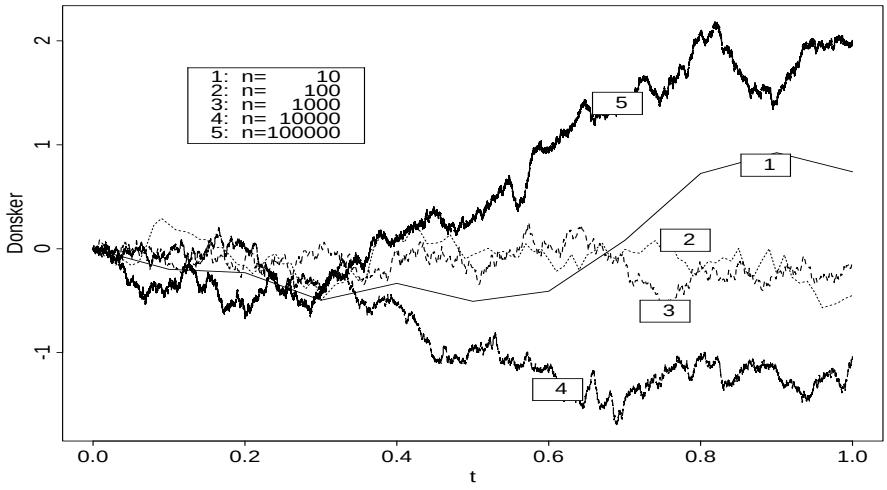

Figure 3.10 *Sample paths of the process S_n for different n and the same sequence of standard normal realizations $X_1(\omega), \ldots, X_{100,000}(\omega)$.*

for some curve $x(t)$ and positive σ, c. Thus, writing $P_t = \exp\{\sigma B_t + ct\}$ for geometric Brownian motion, which might be considered as a model for a speculative price,

$$P(B \in A) = P(P_t < x(t) \text{ for some } t \leq 1)$$

which describes the risk of the price P_t falling below a given deterministic threshold. The boundary

$$\{B \in \partial(A)\} = \{P_t \geq x(t) \text{ for all } t \in [0,1] \text{ and } P_t = x(t) \text{ for some } t \in [0,1]\}$$

is in general not easily shown to have Wiener measure zero.

The convergence rate in the functional CLT is a difficult problem. The Berry–Esséen inequality gives us a hint as to how large the error can be. In general, we would need a result of type

$$\sup_A |P(B \in A) - P(S_n \in A)| \leq c_n \,,$$

for some rate function $c_n \to 0$, where the supremum is taken over all Borel sets A of $\mathbb{C}[0,1]$ with $P(B \in \partial(A)) = 0$. This is quite an ambitious task. What has been shown are decay rates for the difference

$$\sup_{x \in \mathbb{R}} |P(h(B) \leq x) - P(h(S_n) \leq x)| \,,$$

where h is a smooth functional on $\mathbb{C}[0,1]$. The latter supremum is taken over a much smaller class of Borel sets in $\mathbb{C}[0,1]$ than above. With a few exceptions, the rate function c_n which was obtained in these results decays to zero slower than $1/\sqrt{n}$; see for example Borovkov and Sahanenko [8].

Comments

The Donsker invariance principle is treated in the classical book of Billingsley [6] both in $\mathbb{C}[0,1]$ and in $\mathbb{D}[0,1]$. The latter book is an excellent introduction to the weak convergence theory of stochastic processes. Further extensions of the theory can be found in Pollard [28].

3.1.3 Some extensions

Multivariate Brownian motion. For applications in finance one is often interested in modeling several sources of randomness which are dependent. For example, foreign exchange rates are heavily dependent and so are share prices of different stock which are linked through some economic factors (oil price, dependence on the US market, etc.)

A simple model for dependence is to consider correlated multivariate Brownian motion:

$$\mathbf{B}_t^{(d)} = (B_t^{(1)}, \ldots, B_t^{(d)})' \,,$$

where each of the $B^{(i)}$'s is a one-dimensional Brownian motion and $\mathbf{B}$ has indendent stationary Gaussian increments, but the components at a fixed instant of time t are dependent with a correlation matrix which does not depend on t. The simplest case appears when $\mathbf{B}$ consists of independent Brownian motions $B^{(i)}$.

Example 3.11 (Two-dimensional correlated Brownian motion)
Consider two independent standard Brownian motions $W^{(1)}$ and $W^{(2)}$. Then it is not difficult to see that

$$B^{(1)} = \frac{aW^{(1)} + bW^{(2)}}{\sqrt{a^2+b^2}} \quad \text{and} \quad B^{(2)} = \frac{aW^{(1)} - bW^{(2)}}{\sqrt{a^2+b^2}} \tag{3.11}$$

for any a, b with $a^2 + b^2 > 0$ are Brownian motions. As linear combinations of $W^{(1)}$ and $W^{(2)}$ they have independent stationary increments and a.s. continuous sample paths. For fixed t they have correlation structure

$$\operatorname{corr}(B_t^{(1)}, B_t^{(2)}) = \frac{E[(aW_t^{(1)} + bW_t^{(2)})(aW_t^{(1)} - bW_t^{(2)})]}{(a^2 + b^2)t} = \frac{a^2 - b^2}{a^2 + b^2} = \rho\,.$$

Thus two-dimensional correlated Brownian motion can be generated from two independent Brownian motions by linearly combining them. The correlation parameter $\rho \in [-1, 1]$ governs the dependence between the two components. It is clear that (3.11) can be used to simulate two-dimensional correlated Brownian motion from independent simulations of one-dimensional Brownian motions. □

The previous example can be extended to the multivariate case. Let

$$\mathbf{W} = (W^{(1)}, \ldots, W^{(d)})'$$

be a vector of independent Brownian motions and Σ be a covariance matrix which we assume to be invertible. The latter matrix can be written as

$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)\, R \operatorname{diag}(\sigma_1, \ldots, \sigma_d)\,,$$

where R is the correlation matrix of a vector whose components have unit variance. Thus the σ_i's are the standard deviations of those components. We will assume for simplicity that $\sigma_i = 1$ for all i, i.e., $\Sigma = R$. The latter maxtrix can be written as a product of two $d \times d$ matrices: $\Sigma = AA'$, see p. 57. Write

$$\mathbf{B} = A\,\mathbf{W}\,. \tag{3.12}$$

By construction as a linear transformation of $\mathbf{W}$, this process has independent stationary a.s. continuous increments. Every component process is a one-dimensional Brownian motion (linear combinations of Brownian motions are Brownian motion's) and the correlation structure for fixed t does not depend on t. To see this we calculate the covariance matrix for fixed t:

$$\operatorname{cov}(\mathbf{B}_t) = \operatorname{cov}(A\,\mathbf{W}_t) = A \operatorname{cov}(\mathbf{W}_t) A' = A\,[tI_d]\,A' = t\,\Sigma\,.$$

Then

$$\operatorname{corr}(B_t^{(i)}, B_t^{(j)}) = \frac{t\,\sigma_{ij}}{t\sqrt{\sigma_{ii}\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \sigma_{ij}\,,$$

i.e., $\Sigma = R$ is the correlation matrix of $\mathbf{B}_t$.

The above construction shows that it is not difficult to simulate multivariate Brownian motion with correlated components. For the dependence structure we only need to know the correlations between the components. The above construction is also easily extended to the case of different variances in the components, i.e., when standard Brownian motion B is replaced by σB for some positive σ. For this reason the modeling of prices of financial assets by multivariate geometric Brownian motion is quite attractive and exploited in financial practice.

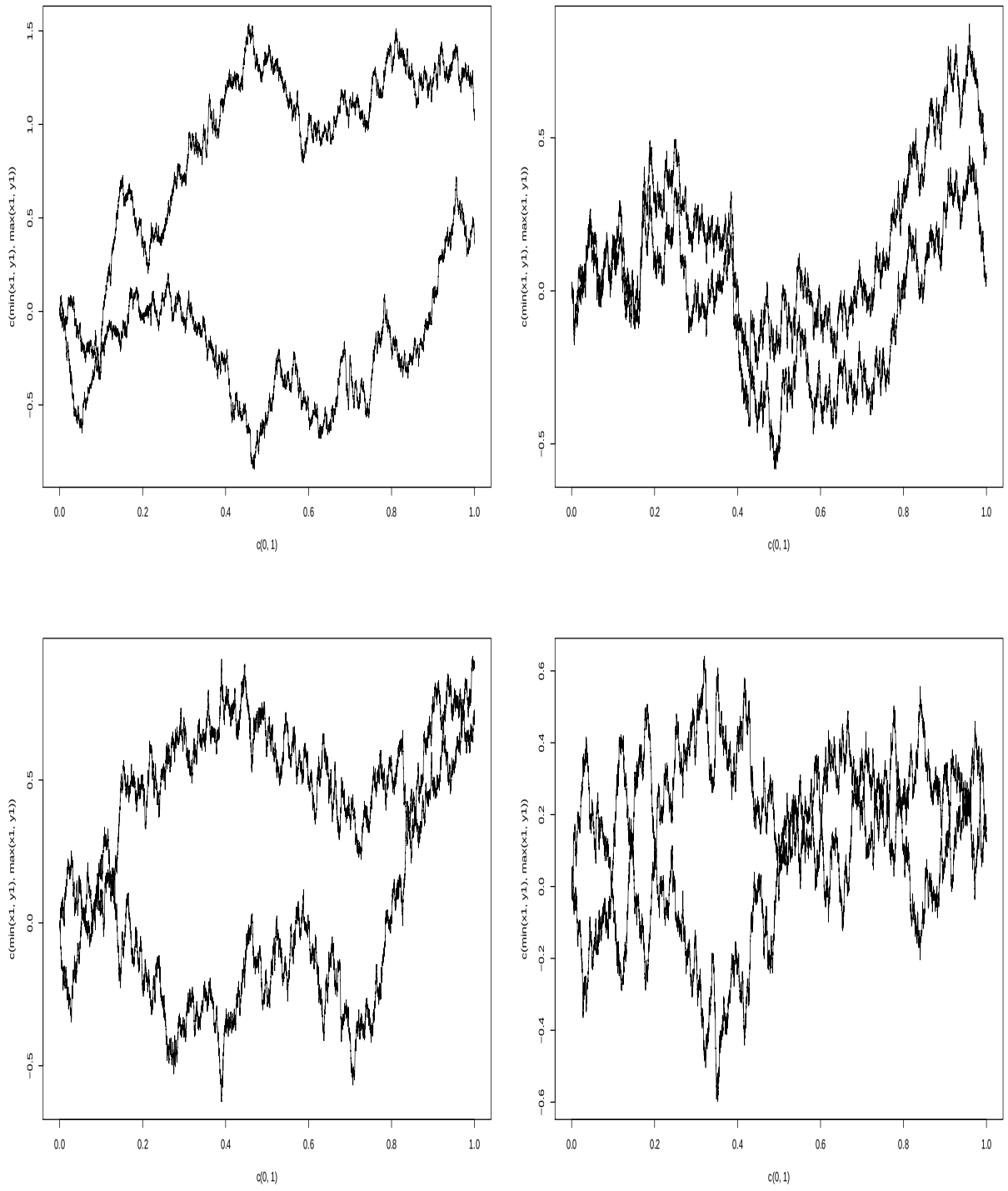

Figure 3.12 *Simulation of correlated Brownian sample path with* $\rho = 0.2$ (top left), $\rho = 0.9$ (top right) *and* $\rho = -0.6$ (bottom left), $\rho = -0.8$ (bottom right).

Gaussian processes

Recall that a process is said to be *Gaussian* if its finite-dimensional distributions are multivariate Gaussian. Brownian motion is a simple example of this class of functions. However, the class of Gaussian stochastic processes is much wider and offers plenty of opportunity for modeling real-life phenomena such as dependence, burstiness, fractality, self-similarity,.... A mean zero Gaussian processes is determined only through its covariance function. This makes them attractive "simple" models.

Example 3.13 (Fractional Brownian motion)
This process has become quite popular over the last few years due to Benoit Mandelbrot who coined the fancy name for it. However, fractional Brownian motion was already considered by Kolmogorov in the 1940s in order to model turbulence. Let $H \in (0, 1]$. A mean zero Gaussian process B_H on $[0, \infty)$ is said to be (*standard*) *H-fractional Brownian motion* if $B_H(0) = 0$ a.s. and if it has covariance structure

$$\operatorname{cov}(B_H(t), B_H(s)) = 0.5 \left[s^{2H} + t^{2H} - |t - s|^{2H}\right], \quad t, s \in [0, \infty).$$

Notice that $H = 0.5$ corresponds to standard Brownian motion. The case $H = 1$ corresponds to $B_H(t) = tZ$ for some $N(0, 1)$ random variable Z and is not of particular interest, and therefore excluded from our considerations. For $H \in (0, 1)$, the sample paths of this process can be shown to be non-differentiable. Moreover, they are self-similar with exponent H:

$$(B_{ct_1}, \ldots, B_{ct_n}) \stackrel{d}{=} c^H (B_{t_1}, \ldots, B_{t_n}), \quad c > 0.$$

This is easily seen from calculating the covariances on both sides of these Gaussian vectors which turn out to be the same.

We observed the self-similarity property with $H = 0.5$ as a useful one when we considered Brownian motion. Indeed, this property allows one to simulate fractional Brownian motion on any interval $[0, c]$ from a simulation on $[0, 1]$ simply by rescaling time and space.

The popularity of fractional Brownian motion for $H \in (0.5, 1)$ is based on one particular property which it does not share with fractional Brownian motion for $H \in (0, 0.5]$: *long-range dependence* (LRD) or *long memory* of its increments. This means the following. First, it is not difficult to see that fractional Brownian motion has stationary increments. If one considers the stationary time series

$$X_t = B_H(t) - B_H(t - 1), \quad t = 1, 2, \ldots,$$

its autocovariance function is given by

$$\gamma_X(h) = \operatorname{cov}(X_1, X_{1+h}) = 0.5 \left[|h + 1|^{2H} - 2h^{2H} + |h - 1|^{2H}\right].$$

One says that γ_X decays *hyperbolically* for $H \in (0.5, 1)$. In particular, it decays extremely slowly such that

$$\sum_{h=1}^{\infty} |\gamma_X(h)| = \infty,$$

and the latter property is often taken as defining property of LRD. It refers to the fact that there is extremely long memory in the time series; a value of B_H at time t has influence on the values at time $t + h$ for a very long time horizon. This unusual phenomenon has been observed in areas

such a climatology and meteorology, and it is also claimed to happen in finance but it seems that nobody has taken advantage of it to make money with this information.... We refer to Beran [5] for more information on LRD and how to detect it statistically.

The LRD present in the increments makes also one issue quite clear: it must be very difficult to simulate fractional Brownian motion by standards means. As for Brownian motion, one can give approximations to LRD fractional Brownian motion via functional CLTs, clearly not based on sums of independent random variables but on sums of moving averages of iid random variables; see for example Samorodnitsky and Taqqu [31] for some hints how to do that. However, via simulations of moving averages one cannot generate dependence over a very long time horizon, basically ranging into the infinite past; one has to "truncate the dependence", or in other words, by means of CLTs one can never get dependence reaching in the past for a very long time. In what follows, we will indicate how one can overcome this problem by a method which generates fractional Brownian motion at finitely many instants of time with the exact Gaussian distribution. We mention that one can find many methods for simulating fractional Brownian motion in the literature, for example by browsing on the Internet. Most of the methods one can find there are dubious and do not achieve what they are made for.

Extensions of fractional Brownian motion with two-dimensional index set, so-called fractional Brownian sheets (by fixing one of the components in the index one gets a one-dimensional fractional Brownian motion) are used in computer animations, for instance for military purposes in order to generate "landscape" on a computer screen. In contrast to Brownian motion, fractional Brownian motion for $H \in (0.5, 1)$ has much smoother sample paths (although being non-differentiable) and therefore a simulation of "smoother" fractional Brownian sheets gives one the impression of a "real" landscape.

A recent source where different methods for simulating fractional Brownian motion can be found is Doukhan et al. [12]. The authors also give a critical discussion and comparison of their methods used. It is perhaps the only reference of this kind. The theory of fractional Brownian motion and more general self-similar processes is nicely explained in Samorodnitsky and Taqqu [31]. □

In what follows, we want to indicate how one can simulate a Gaussian process ξ on an interval, say $[0, 1]$, at finitely many instants of time $t_1 = 0 < t_2 < \cdots < t_n = 1$. This means we are only interested in simulating a vector $(\xi_{t_1}, \ldots, \xi_{t_n})$ with the exact Gaussian distribution prescribed by the finite-dimensional distributions of the process. We assume that the process has mean function zero. Then the finite-dimensional distributions are described only by the covariance function of the process. Thus we are done if we can simulate a mean zero Gaussian vector

$$\mathbf{X} = (X_1, \ldots, X_n)'$$

with a given covariance matrix

$$\Sigma = (\mathrm{cov}(X_i, X_j))_{i,j=1,\ldots,n} .$$

One of the standard methods to solve this problem is to decompose the covariance matrix as

$$\Sigma = AA' ,$$

for a matrix $A = (a_{ij})_{i,j=1,\ldots,n}$. Then take a vector

$$\mathbf{Z} = (Z_1, \ldots, Z_n)'$$

of iid $N(0, 1)$ random variables. The vector

$$\mathbf{X} = A\,\mathbf{Z} ,$$

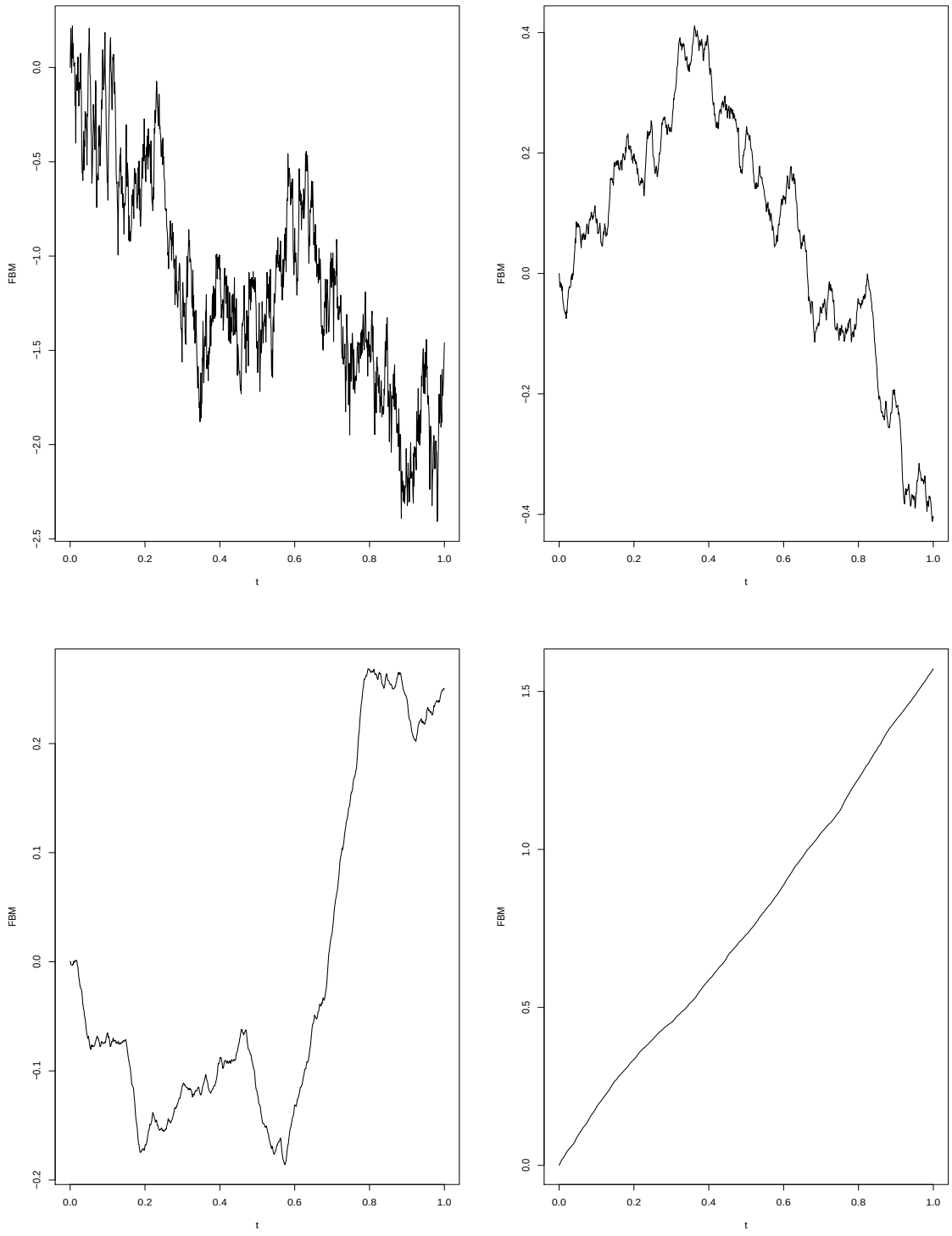

Figure 3.14 *Simulation of fractional Brownian sample paths $B_H(t)$ on $[0,1]$ at 1000 equidistant instants of time.* Top left: $H = 0.3$. Top right: $H = 0.7$. Bottom left: $H = 0.9$ Bottom right: $H = 0.99$.

then has the right distribution $N(\mathbf{0}, \mathbf{A}\mathbf{A}') = \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$.

The matrix A can have different forms. We want to discuss two of them.

The Cholesky decomposition. This is certainly one of the most popular methods which can be found in the literature (for example Golub and van Loan [16]) and is aso implemented in standard packages, e.g. in **R**. It is based on the assumption that A is a lower triangular matrix. Such a matrix is also convenient from a computational point of view because it reduces the calculation of $A\mathbf{Z}$. We focus on Σ positive definite; the semidifinite case (i.e., when Σ is singular) is also treated in [16]. We have to solve the matrix equation

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & & 0 \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ 0 & a_{22} & \cdots & a_{n2} \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} = \Sigma\,.$$

Straightforward calculation gives

$$\begin{aligned} a_{11}^2 &= \sigma_{11} \\ a_{21}a_{11} &= \sigma_{21} \\ &\vdots \\ a_{n1}a_{11} &= \sigma_{n1} \\ a_{21}^2 + a_{22}^2 &= \sigma_{22} \\ &\vdots \\ a_{n1}^2 + \cdots + a_{nn}^2 &= \sigma_{nn}\,, \end{aligned}$$

or in compact form,

$$\sigma_{ij} = \sum_{k=1}^{j} a_{ik}a_{jk}\,, \quad j \le i\,.$$

In particular,

$$\begin{aligned} a_{ij} &= \left(\sigma_{ij} - \sum_{k=1}^{j-1} a_{ik}a_{jk}\right) \Big/ a_{jj}\,, \quad j < i\,, \\ a_{ii} &= \sqrt{\sigma_{ii} - \sum_{k=1}^{i-1} a_{ik}^2}\,. \end{aligned}$$

Positive definiteness of Σ ensures that a_{ii} is always positive, so that one does not have to divide by a zero a_{ii}.

Diagonalization. Some basic theory on matrices tells us that we can get the decomposition $\Sigma = AA'$ by diagonalizaton. Indeed, since Σ is positive semidefinite and symmetric, there exist orthogonal matrices O, i.e., $OO' = I$, such that $\Sigma = O\Lambda O'$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ contains the eigenvalues λ_i of Σ. Thus one can choose

$$A = O\,\Lambda^{1/2} = O\,\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})\,,$$

leading to

$$AA' = O\,\Lambda O' = \Sigma\,.$$

In contrast to the Cholesky decomposition, the matrix A does not have a particular structure, providing a computational advantage in evaluating $A\mathbf{Z}$. Standard packages contain procedures for calculating O and Λ. In passing we mention that we can also choose $A = \Sigma^{1/2}$, i.e.,

$$\Sigma^{1/2} = O\Lambda^{1/2}O'\,. \tag{3.13}$$

Comments

There exist efficient sophisticated numerical methods for the described diagonalization and Cholesky algorithms which are contained in any good software package. Books like Golub and van Loan [16] discuss the problem of stability of these procedures. The Cholesky decomposition and the diagonalization method are computer intensive and should be avoided if possible. For example, the simulation of Brownian motion at a finite time grid via these methods would not be reasonable. The advantage of these methods is that the distributions of these simulations are exact due to the Gaussianity of the underlying distributions.

3.2 The Riemann–Stieltjes integral

Before we go to the stochastic (Itô) integral we make a short excursion to some deterministic integrals whose definition is close to the one for the Itô integral. Predecessors of the Itô integral, in historical order, are the Riemann–Stieltjes integral and the Lebesgue–Stieltjes integral. The latter integral (we restrict ourselves to the interval $[0,1]$)

$$\int f\,dg = \int_{[0,1]} f(x)\,dg(x)\,, \tag{3.14}$$

is nothing but the integral we have encountered in a course on measure and integral. Indeed, g is a distribution function in the sense that g is càdlàg, non-decreasing and finite in $(0,1)$, thus the values

$$g(a,b] = g(b) - g(a)$$

define a Radon or Lebesgue–Stieltjes measure g on $[0,1]$, and (3.14) is nothing but the integral of f with respect to this measure. We know that the expectation of any random variable can be written in this form.

The definition of the Lebesgue–Stieltjes integral goes through different steps (simple functions, non-negative integrands, general integrands) and therefore its construction is not very illuminating. However, it is the limit of integrals $\int f_n dg$ for simple (step) functions

$$f_n = \sum_{i=1}^{k_n} f_i^{(n)}\, I_{A_i^{(n)}}\,,$$

where $A_i^{(n)}$, $i = 1,\dots,k_n$, is some finite disjoint partition τ_n of the interval $[0,1]$. Then

$$\int f_n\,dg = \sum_{i=1}^{k_n} f_i^{(n)}\, g(A_i^{(n)})\,.$$

In some cases the partitions can be chosen as $A_i^{(n)} = (t_{i-1}^{(n)}, t_i^{(n)}]$, where

$$\mathrm{mesh}(\tau_n) = \max_{i=1,\ldots,k_n} (t_i^{(n)} - t_{i-1}^{(n)}) =: \max_{i=1,\ldots,k_n} \Delta_i^{(n)} \to 0\,.$$

Now define

$$f_i^{(n)} = f(\xi_i^{(n)}) \quad \text{for any } \xi_i^{(n)} \in (t_{i-1}, t_i].$$

If the limit $\lim_{n\to\infty} \int f_n dg$ exists for any choice of $(\xi_i^{(n)})$ and partitions (τ_n) with $\mathrm{mesh}(\tau_n) \to 0$ and the limiting value is the same for these choices, then the limiting value is called the *Riemann–Stieltjes integral* and denoted by $\int_0^1 f dg$. (We often suppress the limits.) The Riemann–Stieltjes integral is the integral which is used most often in probability theory and statistics. Although the Lebesgue–Stieltjes integral has become the standard — thanks to Kolmogorov's axioms we all know about measure theory and integrals with respect to a measure and therefore we believe we need and understand the Lebesgue–Stieltjes integral — when it comes to calculating expectations for concrete distributions or random variables, we usually calculate Riemann–Stieltjes integral.

The relation between Riemann–Stieltjes integral and Lebesgue–Stieltjes integral is well understood, but not too well known. The exceptional case is $g(x) = x$ (i.e., g defines Lebesgue measure) and f is a bounded function on $[0, 1]$. Then f is Riemann integrable if and only if it has discontinuities of Lebesgue measure zero (see Billinsgley [7], Problem 17.1). The case of the general Riemann–Stieltjes integral is more complicated. It is sometimes given as an exercise in measure theory to show that the Riemann–Stieltjes integral $\int f dg$ exists and coincides with the Lebesgue–Stieltjes integral if f is countinuous and g is a distribution function on $[0, 1]$. The fine structure of the Riemann–Stieltjes integral was studied by L.C. Young [35] in a seminal paper from 1936 and then almost forgotten for 60 years. His basic result says the following.

First we define the *p-variation* of a function $f : [0, 1] \to \mathbb{R}$. It is the number

$$\sup_\tau \sum_i |\Delta_i f|^p\,,$$

where the supremum is taken over all finite partitions τ of the interval $[0, 1]$. If the latter number is finite, *f is said to be of finite or bounded p-variation.* Notice that $p = 1$ corresponds to the well known bounded variation case. L.C. Young's theorem says that the Riemann–Stieltjes integral $\int f dg$ exists if

- f and g do not have discontinuities at the same point $t \in [0, 1]$,
- f is of bounded p-variation and g is of bounded q-variation for $p^{-1} + q^{-1} > 1$.

He also showed that these conditions cannot be relaxed. For a recent treatment and some extensions of integrals of Riemann–Stieltjes integral-type we refer to the recent book of Dudley and Norvaiša [14].

On obvious advantage of a Riemann–Stieltjes integral is that the integral $\int f dg$ can always be approximated by Riemann–Stieltjes sums, just by exploiting the definition of the Riemann–Stieltjes integral:

$$\int f\,dg = \lim_{n\to\infty} \sum_{i=1}^{n} f(\xi_i^{(n)})\,\Delta_i^{(n)} g\,,$$

where

$$\Delta_i^{(n)} g = g(t_i^{(n)}) - g(t_{i-1}^{(n)})$$

and the limit is taken along any sequence of partitions τ_n of the interval $[0,1]$ with $\text{mesh}(\tau_n) \to 0$. A natural choice of partitions would be $t_i^{(n)} = i/n$ for $i = 0, \ldots, n$.

The Riemann–Stieltjes integral is a purely deterministic object. It can be used to define a stochastic integral if both f and g are interpreted as sample paths of stochastic processes. The main problem is then to determine whether these paths have finite p-variation for some $p > 0$. This has fortunately been done in the literature for major classes of processes. We give some examples; references to particular classes of processes can be found in Mikosch and Norvaiša [26].

Example 3.15 (p-variation of Brownian sample paths)
The sample paths of Brownian motion B on $[0,1]$ have finite p-variation for $p > 2$ with probability 1 but they have infinite 2-variation with probability 1. This is a property it shares with many elements of the class of semimartingales. Notice that 2-variation is *not* the quadratic variation of Brownian motion used in Itô calculus, which is finite.

A particular consequence is that we can integrate any differentiable function f with bounded derivative f' with respect to a Brownian sample path. Indeed, if f is differentiable,

$$\Delta_i f = f(t_i) - f(t_{i-1}) = f'(\xi_i)\,\Delta_i = f'(\xi_i)\,(t_i - t_{i-1})\,,$$

for some $\xi_i \in (t_{i-1}, t_i)$, and then it is easy to see that f has bounded q-variation with $q = 1$. Since Brownian motion has bounded p-variation with $p > 2$, $p^{-1} + q^{-1} > 1$, and therefore $\int f(t)dB_t$ is defined in a pathwise sense as Riemann–Stieltjes integral. In turn, any integral of this kind can be approximated by Riemann–Stieltjes sums, i.e., we do not need a theory of numerical approximation in that case.

However, integrals of type $\int_0^1 B_s^k dB_s$ for some integer $k \geq 1$ cannot be interpreted as Riemann–Stieltjes integrals; see also the discussion in Section 3.3. In that case the condition $p^{-1} + q^{-1} = 2p^{-1} > 1$ cannot be satisfied because Brownian motion and its powers have infinite 2-variation. This is one reason why the Itô integral is needed.

In the literature, and sometimes even in textbooks, one can find the remarks that Riemann–Stieltjes integrals with respect to Brownian motion cannot be defined *because Brownian sample paths have unbounded variation.* The above discussion shows that this is indeed incorrect, and one can define a stochastic integral for certain classes of integrands without using Itô calculus. As a matter of fact, something different is correct: if the Riemann–Stieltjes integral $\int f dg$ exists for *all* continuous functions f on $[0,1]$, then g is necessarily of bounded variation; see Theorem 52 in Protter [29]. Thus, the Riemann–Stieltjes integral $\int f(t)dB_t$ is not defined for all continuous functions since Brownian motion has unbounded variation. The Itô integral ensures that $\int f(t)dB_t$ exists for all *deterministic* continuous integrands. This can be considered as an advantage of the Itô integral.

□

Example 3.16 (Fractional Brownian motion B_H for $H \in (0.5, 1)$)
In Example 3.13 we learned about fractional Brownian motion B_H which like Brownian motion $B = B_{0.5}$ is a Gaussian process with stationary increments. The process B_H has smoother sample paths for $H \in (0.5, 1)$ due to the very strong dependence (long range dependence) of its increments. This does not mean that they are differentiable, but they are close to having bounded variation. To be more precise, they have bounded p-variation with $p > H^{-1}$. Thus the Riemann–Stieltjes integrals $\int B_H(t)dB_t$, $\int B(t)dB_H(t)$, $\int B_H(t)dB_H(t)$ are all well defined with probability 1 and, thus, their Riemann–Stieltjes sum can be considered as numerical approximation to these integrals.

A nice property of the Riemann–Stieltjes integral is that integral equations which are based on Riemann–Stieltjes integrals can be shown to have solutions by using the rules of "classical calculus".

For example, the linear stochastic differential equation

$$X_t = X_0 + c \int_0^t X_s \, ds + \sigma \int_0^t X_s \, dB_H(s) \,, \quad 0 \le t \le 1 \,, \tag{3.15}$$

has pathwise solution $X_t = X_0 \mathrm{e}^{\, ct + \sigma B_H(t)}$ and

$$X_t = X_0 - c \int_0^t X_s \, ds + \sigma \int_0^t dB_H(s) \,, \quad 0 \le t \le 1 \,,$$

has pathwise solution $X_t = \mathrm{e}^{\,-ct}(X_0 + \sigma \int_0^t e^{cs} \, dB_H(s))$; see [26]. □

Example 3.17 (Lévy processes)
A Lévy process is a process with stationary independent increments and càdlàg sample paths. We know the homogeneous Poisson process and Brownian motion as two representatives. This class is however much wider and quite flexible when it comes to modeling tails and jumps. As a matter of fact, every Lévy process X on $[0, 1]$ is a linear combination of two independent Lévy processes, one of them being Brownian motion with a.s. continuous sample paths, and the other one is a pure jump process. The latter processes have gained particular popularity over the last few years since they are considered as more realistic models for applications in finance, i.e., the jump character of the sample paths is closer to real-life financial processes than Brownian motion with its continuous and nowhere differentiable paths. The interested reader is referred to Sato's book [32] which gives a gentle introduction and an up-to date overview of Lévy process theory. For applications of Lévy processes, see Barndorff-Nielsen et al. [4]. For a long time it was considered a difficult task to simulate the paths of a particular Lévy process different from Brownian motion and the Poisson process. For an overview of some recently developed methods (series representations) for the simulation of Lévy paths, see Rosiński in [4] and Asmussen and Rosiński [2].

Many Lévy jump processes L which have gained some popularity (normal inverse Gaussian process, hyperbolic Lévy motion, generalized inverse Gaussian process, infinite variance stable process) have bounded p-variation with $p < 2$. This again allows one to determine the Riemann–Stieltjes integral $\int X_t dL_t$ or $\int L_t dX_t$ in the Riemann–Stieltjes sense provided the q-variation of X is such that $p^{-1} + q^{-1} > 1$. Riemann–Stieltjes sums are then numerical approximations to these integrals. □

3.3 A prime example of an Itô integral

After all, we might be tempted to define the Itô integral in a way similar to the Riemann–Stieltjes integral, as a pathwise integral, but our instinct tells us that something must go wrong. Indeed, we know that for a Brownian motion B

$$\int_0^1 B_t \, dB_t = \frac{1}{2} \, \left(B_1^2 - 1 \right) \,.$$

This would contradict the rules of Riemann–Stieltjes integration which would give us $B_1^2/2$. Moreover, we know that the rule $p^{-1} + q^{-1} > 1$ for the variation of integrand and integrator are not satisfied for Brownian motion.

Recall that the Itô integral $\int_0^1 B_t dB_t$ is the limit of the Riemann–Stieltjes sums

$$\sum_{i=1}^n B_{t_{i-1}} \, \Delta_i B = \frac{1}{2} B_1^2 - \frac{1}{2} \sum_{i=1}^n (\Delta_i B)^2 = \frac{1}{2} B_1^2 - \frac{1}{2} Q_n(1) \tag{3.16}$$

along any sequence of partitions $\tau_n : 0 = t_0 < \cdots < t_n = 1$ with $\text{mesh}(\tau_n) \to 0$. The convergence is, however, not for a.e. sample path but has to be taken in an L^2-sense. Indeed, since $EQ_n(1) = 1$, one can easily verify that $E(Q_n(1) - 1)^2 \to 0$, in particular

$$\sum_{i=1}^{n} B_{t_{i-1}} \, \Delta_i B \stackrel{L^2}{\to} \frac{1}{2} \left(B_1^2 - 1 \right) .$$

Thus the Itô integral is somewhat like a washing machine avoiding bad sample path behavior: although we do not have convergence along one sample path, we have convergence in L^2, i.e., in an averaging sense over all sample paths.

This example also highlights another crucial difference to the Riemann–Stieltjes integral: it is essential that the Riemann–Stieltjes sums (3.16) are defined in this particular form, i.e., one does not have the freedom to define the Itô integral as the limit of the Riemann–Stieltjes sums

$$\sum_{i=1}^{n} B_{\xi_i} \, \Delta_i B \tag{3.17}$$

for any choice of $\xi_i \in [t_{i-1}, t_i]$. Indeed, if one chooses $\xi_i = t_{i-1} + p\Delta_i$ for any fixed $p \in [0, 1]$, one can see that the Riemann–Stieltjes sums (3.17) have the limit

$$\frac{1}{2} B_1^2 + \left(p - \frac{1}{2} \right)$$

in L^2, i.e., we have infinitely many possibilities to define a stochastic integral $\int B_s dB_s$. The second best known integral corresponds to the case $p = 0.5$ which is called *Stratonovich integral* and denoted by $\int_0^1 B_s \circ dB_s$. It has value $0.5B_1^2$, and the rules for this *Stratonovich calculus* are very much the same as for classical calculus. Therefore it gained some popularity. Clearly, Itô calculus gained its major popularity due to the fact that the Itô integral process $(\int_0^t B_s dB_s \, , 0 \le t \le 1)$ is a martingale with respect to the Brownian filtration. All the other integrals do not have this property.

If one has an explicit formula for the Itô integral in terms of Brownian motion it is not difficult to simulate it by first simulating a Brownian sample path and then plugging it into the formula for the integral. For example, the Itô integral $\int B_s dB_s$ would allow for such an approach. The latter integral also gives us an idea how we could get a numerical approximation *without going through all steps of the definition of an Itô integral.* First of all, notice that the Riemann–Stieltjes sums (3.16) converge in L^2 to $\int B_s dB_s$, hence in probability, and so we can find a subsequence (n_k) along which this convergence holds a.s., i.e., pathwise. Unfortunately, it does not tell us how we can choose the partitions and what (n_k) is. Dudley [13] improved upon this result by showing that the Riemann–Stieltjes sums (3.16) converge a.s. to the Itô integral provided the partitions satisfy $\text{mesh}(\tau_n) = o(1/\log n)$, and he showed that this fact does not remain valid if $\text{mesh}(\tau_n) = O(1/\log n)$; see Dudley and Norvaiša [14] for more details. This example shows us that there is some hope that we can define the Itô integral in a pathwise sense if we do not consider any sequence of partitions (τ_n) but if we choose some particular partition. This is the basic idea of numerical solution to stochastic differential equations and to approximations to the Itô integral.

3.4 Strong solutions to stochastic differential equations

3.4.1 Uniqueness and existence of the solution

In what follows, we consider the stochastic differential equation

$$X_t = X_0 + \int_0^t a(s, X_s)\, ds + \int_0^t b(s, X_s)\, dB_s\,, \quad 0 \le t \le 1\,, \tag{3.18}$$

where the first integral is interpreted as a Riemann integral and the second one as an Itô integral. The coefficient functions a and b are specified below. They must satisfy some measurability and integrability conditions which we take for granted without mentioning them. It is assumed that you know how the stochastic integral is defined and what the meaning of a stochastic differential equation is.

A *strong solution* to (3.18) is a stochastic process X which is adapted to the natural Brownian filtration $\mathcal{F}_t = \sigma(B_s, s \le t)$, X is a function of the underlying sample path and the coefficient functions a and b. Thus, if you were to change one Brownian path by another one, the solution would be given by exactly the same functional relationship.

The following conditions are standard ones for the existence and uniqueness of strong solutions; see e.g. Chung and Williams [9], Theorem 10.5, or Kloeden and Platen [20], Theorem 4.5.3. for proofs.

Theorem 3.18 (Uniqueness and existence of a strong solution)
Assume that a and b satisfy

- *the* Lipschitz condition

$$|a(t,x) - b(t,x)| \le K\,|x-y|\,, \quad |b(t,x) - b(t,y)| \le K\,|x-y|\,, \quad t \in [0,1]\,, t \in \mathbb{R}\,,$$

  for some constant K,

- *and the* linear growth condition*:*

$$|a(t,x)| \le K\,(1+|x|)\,, \quad |b(t,x)| \le K\,(1+|x|)\,, \quad t \in [0,1]\,, x \in \mathbb{R}\,,$$

- *X_0 is constant.*

Then the stochastic differential equation (3.18) *has a unique strong solution X on* $[0,1]$ *with*

$$\sup_{0 \le t \le 1} EX_t^2 < \infty\,.$$

Example 3.19 (Linear stochastic differential equation)
Prime examples of stochastic differential equations for which the assumptions of Theorem 3.18 are satisfied are the *linear stochastic differential equations*. As indicated by the name, the functions a and b are then linear functions:

$$a(t,x) = a_1(t)\,x + a_2(t) \quad \text{and} \quad b(t,x) = b_1(t)\,x + b_2(t)\,,$$

where a_i and b_j are continuous. Then it is easily seen that the above linear growth and Lipschitz conditions are satisfied.

For many applications, in particular in finance, one chooses linear stochastic differential equations, often with constant a_i and b_j. The advantage is that one has explicit solutions for these

stochastic differential equations. For example, the Black-Scholes model is based on the belief that speculative prices P_t can be modeled by

$$dP_t = c\,P_t\,dt + \sigma\,P_t\,dB_t\,, \tag{3.19}$$

where c and σ are given constants. The unique solution is geometric Brownian motion

$$P_t = P_0\,\mathrm{e}^{\,(c-0.5\sigma^2)\,t+\sigma\,B_t}\,.$$

At this point it might be worthwhile looking back at Example 3.16. The equation (3.19) with B replaced by fractional Brownian motion B_H for some $H > 0.5$ (see (3.15)) has solution $P_t = P_0\,\mathrm{e}^{\,c\,t+\sigma\,B_H(t)}$. The crucial difference in the exponent is due to the chain rule of Itô calculus – the Itô formula.

Another linear equation is used for modeling interest rates r_t in the Vasicek model given by

$$dr_t = c\,(\mu - r_t)\,dt + \sigma\,dB_t\,, \tag{3.20}$$

where c, μ and σ are positive constants. The rationale of this model is that r_t fluctuates around the value μ. If it deviates from μ it will immediately be driven back to μ; one says *it reverts to the mean*. The spead at which this happens is adjusted by the parameter c. The solution to (3.20) is given by

$$r_t = r_0\,\mathrm{e}^{\,-ct} + \mu\,(1 - \mathrm{e}^{\,-ct}) + \sigma\,\mathrm{e}^{\,-ct}\int_0^t \mathrm{e}^{\,cs}\,dB_s\,.$$

The mean reversion is seen from

$$Er_t = r_0\,\mathrm{e}^{\,-ct} + \mu\,(1 - \mathrm{e}^{\,-ct}) \quad \text{and} \quad \mathrm{var}(r_t) = \frac{\sigma^2}{2c}\,(1 - \mathrm{e}^{\,-2ct})\,.$$

□

3.4.2 Strong numerical solution – the Euler scheme

Under the assumptions of Theorem 3.18 one can also show the existence of a strong numerical solution to the stochastic differential equation (3.18). First we want to explain what this means. A natural way of mimicing the stochastic differential equation

$$dX_t = a(t, X_t)\,dt + b(t, X_t)\,dB_t\,, \quad 0 \le t \le 1\,, \tag{3.21}$$

where B is standard Brownian motion and a, b are given coefficient functions, is to replace the differentials dt and dB_t by differences on a discrete partition

$$\tau_n : t_0 = 0 < t_1 < \cdots < t_n = 1\,.$$

The points t_i depend on n which fact we supress in the notation for ease of presentation. We write

$$\Delta_i = t_i - t_{i-1} \quad \text{and} \quad \Delta_i f = f(t_i) - f(t_{i-1})$$

for any function or stochastic process f defined on $[0, 1]$. We assume that we know the *initial value* X_0 which we always assume constant.

The *Euler approximation scheme* to the solution of the stochastic differential equation (3.21), provided the latter exists, is given by the iterative scheme:

$$\begin{aligned} X_0^{(n)} &= X_0 \quad \text{and, for } i = 1, \dots, n, \\ X_{t_i}^{(n)} &= X_{t_{i-1}}^{(n)} + a(t_{i-1}, X_{t_{i-1}}^{(n)})\, \Delta_i + b(t_{i-1}, X_{t_{i-1}}^{(n)})\, \Delta_i B\,. \end{aligned}$$

In this way, the Euler approximation $X^{(n)}$ is determined at the points t_i of the partition. In order to define a continuous time process it is common to linearly interpolate elsewhere. This means we write for $t \in [t_{i-1}, t_i]$,

$$\begin{aligned} X_t^{(n)} &= X_{t_{i-1}}^{(n)} + (t - t_{i-1})\, a(t_{i-1}, X_{t_{i-1}}^{(n)}) + b(t_{i-1}, X_{t_{i-1}}^{(n)})\, (B_t - B_{t_{i-1}}) \\ &= X_{t_{i-1}}^{(n)} + \int_{t_{i-1}}^{t} a(t_{i-1}, X_{t_{i-1}}^{(n)})\, ds + \int_{t_{i-1}}^{t} b(t_{i-1}, X_{t_{i-1}}^{(n)})\, dB_s\,, \quad t \in [t_{i-1}, t_i]\,. \end{aligned} \tag{3.22}$$

From this construction and since we assume X_0 constant it is clear that $(X_t^{(n)})$ is adapted to the natural Brownian filtration.

We write

$$\delta_n = \text{mesh}(\tau_n) = \max_{i=1,\dots,n} \Delta_i\,.$$

It is usual to express the quality of the axppromixation in terms of δ_n. For practical purposes one often chooses the *equidistant Euler scheme*:

$$\tau_n : \quad t_i = n^{-1}\, i\,, \quad i = 0, 1, \dots, n\,.$$

This implies in turn that

$$\begin{aligned} X_0^{(n)} &= X_0 \quad \text{and, for } i = 1, \dots, n \\ X_{i/n}^{(n)} &= X_{(i-1)/n}^{(n)} + a((i-1)/n, X_{(i-1)/n}^{(n)})/n + b((i-1)/n, X_{(i-1)/n}^{(n)})\, \Delta_i B\,, \\ \text{mesh}(\tau_n) &= n^{-1}\,. \end{aligned}$$

The Euler approximation $X^{(n)}$ is said to *converge strongly* to the solution X of the stochastic differential equation (3.21) if

$$\lim_{n\to\infty} E|X_1 - X_1^{(n)}| = 0\,. \tag{3.23}$$

This is quite an arbitrary definition since it focuses on convergence at the right endpoint of the interval of interest. A more sensitive definition would be to require

$$\lim_{n\to\infty} E(\sup_{0\le t\le 1} |X_t - X_t^{(n)}|) = 0\,. \tag{3.24}$$

The definition (3.23), taken from the standard textbook in this field by Kloeden and Platen [20], is certainly dictated by convenience; the choice of the absolute mean difference in (3.23) is another issue one may question.

Relations such as (3.23) and (3.24) refer to a pathwise comparison of X and $X^{(n)}$ which processes are both supposed to be driven by the same Brownian motion. In this sense, one considers a *strong*

or a.s. approximation to X, and therefore $X^{(n)}$ is usually referred to as a *strong numerical solution.* Notice, however, that in real life we never know the underlying Brownian sample paths, and all we can do is to compare X and $X^{(n)}$ for a simulated path of Brownian motion.

A quantity for judging the quality of a strong solution is given by

$$e_s(\delta_n) = E|X_1 - X_1^{(n)}|\,.$$

For example, if

$$e_s(\delta_n) \le c\,\delta_n^{\gamma}\,, \quad \text{for sufficiently large } n,$$

for some positive γ and c, then the numerical solution $X^{(n)}$ is said to converge strongly to X *with order* γ.

Theorem 3.20 (Kloeden and Platen [20], Theorem 10.2.2) *Suppose that the assumptions of Theorem 3.18 are satisfied and that, in addition,*

$$|a(s,x) - a(t,x)| + |b(s,x) - b(t,x)| \le K\,(1+|x|)\,\sqrt{|t-s|}\,, \quad s,t \in [0,1]\,, \quad x,y \in \mathbb{R}\,. \tag{3.25}$$

for some $K > 0$. Then, for some constant $c > 0$, for the Euler scheme,

$$E(\sup_{0\le t\le 1} |X_t - X_t^{(n)}|) \le c\,\delta_n^{0.5}\,,$$

In particular, the Euler approximation scheme is strongly convergent with order $\gamma = 0.5$.

Notice that (3.25) is satisfied for a linear stochastic differential equation with coefficients $a(t,x) = a_1 x + b_2$ and $b(t,x) = b_1 x + b_2$.

It can be shown that the order $\gamma = 0.5$ cannot be improved in general.

Proof. In the proof we focus on the *autonomous case*, i.e.,

$$a(t,x) = a(x) \quad \text{and} \quad b(t,x) = b(x)\,,$$

and only show the result for the right endpoint, i.e.,

$$E|X_1 - X_1^{(n)}| \le c\,\delta_n^{0.5}\,. \tag{3.26}$$

The general case can be found in [20], proof of Theorem 10.2.2. In what follows, we write c for any positive constant whose value is not of interest.

We write for $t \in [t_{i-1}, t_i)$,

$$\widetilde{X}_t^{(n)} = X_{t_{i-1}}^{(n)}$$

Notice that for $t \in [t_{i-1}, t_i)$,

$$\begin{aligned}
X_t - \widetilde{X}_t^{(n)} &= \int_{t_{i-1}}^t a(X_s)\,ds + \int_{t_{i-1}}^t b(X_s)\,dB_s + X_{t_{i-1}} - X_{t_{i-1}}^{(n)} \\
&= \int_{t_{i-1}}^t a(X_s)\,ds + \int_{t_{i-1}}^t b(X_s)\,dB_s \\
&\quad + \sum_{j=1}^{i-1} \int_{t_{j-1}}^{t_j} [a(X_s) - a(\widetilde{X}_s^{(n)})]\,ds + \sum_{j=1}^{i-1} \int_{t_{j-1}}^{t_j} [b(X_s) - b(\widetilde{X}_s^{(n)})]\,dB_s \\
&= \left[\int_{t_{i-1}}^t a(X_s)\,ds + \int_0^{t_{i-1}} [a(X_s) - a(\widetilde{X}_s^{(n)})]\,ds \right] \\
&\quad + \left[\int_{t_{i-1}}^t b(X_s)\,dB_s + \int_0^{t_{i-1}} [b(X_s) - b(\widetilde{X}_s^{(n)})]\,dB_s \right] \\
&= A(t) + B(t)\,.
\end{aligned}$$

We see that for $0 \le s \le 1$,

$$z(s) = \sup_{0\le t\le s} E|X_t - \widetilde{X}_t^{(n)}|^2 \le c \sup_{0\le t\le s} E[A(t)]^2 + c \sup_{0\le t\le s} E[B(t)]^2\,.$$

Let $i_t = \max\{i : t_i \le t\}$. Since $(B(t), 0 \le t \le 1)$ is an Itô integral, its variance can be calculated from the isometry property of the Itô integral and estimated in conjunction with the linear growth and the Lipschitz conditions on b:

$$\begin{aligned}
\sup_{0\le t\le s} E[B(t)]^2 &= \sup_{0\le t\le s} E\left[\int_{t_{i_t}}^t [b(X_r)]^2\,dr + \int_0^{t_{i_t}} [b(X_r) - b(\widetilde{X}_r^{(n)})]^2\,dr \right] \\
&\le c \sup_{0\le t\le s} E\left[\int_{t_{i_t}}^t (1 + X_r^2)\,dr + \int_0^{t_{i_t}} [X_r - \widetilde{X}_r^{(n)}]^2\,dr \right] \\
&\le c \left[\delta_n + \sup_{0\le t\le s} \int_0^{t_{i_t}} E|X_r - \widetilde{X}_r^{(n)}|^2\,dr \right] \\
&\le c \left[\delta_n + \int_0^s z(t)\,dt \right]\,.
\end{aligned}$$

Here we also used that $EX_t^2 < \infty$ uniformly for $t \in [0,1]$. A similar calculation applies to the process A:

$$\begin{aligned}
\sup_{0\le t\le s} E[A(t)]^2 &\le c \sup_{0\le t\le s} E\left[\int_{t_{i_t}}^t a(X_r)\,dr \right]^2 + c \sup_{0\le t\le s} E\left[\int_0^{t_{i_t}} [a(X_r) - a(\widetilde{X}_r^{(n)})]\,dr \right]^2 \\
&\le c \sup_{0\le t\le s} E\left[\int_{t_{i_t}}^t (1 + |X_r|)\,dr \right]^2 + c \sup_{0\le t\le s} E\left[\int_0^{t_{i_t}} |X_r - \widetilde{X}_r^{(n)}|\,dr \right]^2
\end{aligned}$$

$$\leq \quad c \sup_{0\leq t\leq s} \delta_n \int_{t_{i_t}}^{t} (1 + EX_r^2)\, dr + \int_0^s z(t)\, dt$$

$$\leq \quad c \left[\delta_n^2 + \int_0^s z(t)\, dt \right] .$$

where we used Lyapunov's inequality and the uniform boundedness of the second moments of the process X in the last step. Combining the bounds for A and B, we obtain

$$z(s) \leq c \left[\delta_n + \int_0^s z(r)\, dr \right] .$$

Iterating the latter inequality shows that there exists a constant C such that

$$z(s) \leq C\, \delta_n \, .$$

The latter fact is very often used in the theory of stochastic differential equations and referred to as *Gronwall's inequality,* see Remark 3.21 below. An application of Lyapunov's inequality shows that

$$E|X_1 - X_1^{(n)}| \leq \sqrt{z(1)} = O(\sqrt{\delta_n}) \, .$$

This concludes the proof of (3.26). □

Remark 3.21 Gronwall's inequality can be formulated as follows. If there exist real-valued functions f, g on $[0, 1]$ such that

$$0 \leq f(t) \leq g(t) + c \int_0^t f(s)\, ds \, , \quad t \in [0, 1] \, ,$$

for some constant $c > 0$, then

$$f(t) \leq g(t) + c \int_0^t \mathrm{e}^{\,c\,(t-s)}\, g(s)\, ds \, , \quad t \in [0, 1] \, . \tag{3.27}$$

The proof is elementary and follows by plugging in the inequality in the right-hand side iteratively. After evaluating the multiple integrals on the right-hand side, one receives a Taylor expansion for $\mathrm{e}^{\,c\,(t-s)}$ under the integral. Then letting the expansion converge, one may conclude that (3.27) holds.

Comments

Until recently, the problem of numerical solution to stochastic differential equations was not considered as a relevant issue. Only since stochastic differential equations have become bread and butter of financial mathematics and its real-life applications, the need for simulating sample paths of solutions to stochastic differential equations has attracted a lot of attention. The standard textbook in this context is Kloeden and Platen [20] which for a long time was also the book which was accessible to a wider range of practitioners. The companion book by Kloeden et al. [21] is written on a rather elementary level, aiming more at the computational skills than on the understanding of the underlying concepts. The idea of numerical solution to deterministic differential equations is a well established field in numerical mathematics. The idea of using numerical schemes to solve stochastic differential equations is due to Talay [33] who dealt with this topic in his PhD thesis.

Most of the numerical solution results in [20] are proved for multivariate stochastic differential equation.

Little is known if one considers stochastic differential equations driven by non-Brownian processes. The Euler scheme for stochastic differential equations driven by Lévy processes was considered by Protter and Talay [30]. As mentioned in Section 3.2, the consistency of the Euler scheme is often a simple consequence of the fact that the stochastic integrals coincide with the Riemann–Stieltjes integral. Then the Riemann–Stieltjes sums can be taken as an Euler approximation. However, this approach does not provide additional information about the quality of the approximation.

3.4.3 Improvement on the Euler scheme - Taylor-Itô expansion and Milstein scheme

The order of the Euler scheme can be substantially improved by some kind of Taylor expansion. Since the chain rule of Itô calculus, the Itô formula, is used to obtain it, one also refers to a *Taylor-Itô* expansion. We explain the latter for the autonomous stochastic differential equation

$$X_t = X_0 + \int_0^t a(X_s)\,ds + \int_0^t b(X_s)\,dB_s\,, \quad 0 \le t \le 1\,. \tag{3.28}$$

For the coefficient functions a and b we assume that all conditions for the existence of a unique strong solution are satisified; see Theorem 3.18. As before, consider a partition

$$\tau_n : \quad 0 = t_0 < \cdots < t_n = 1\,.$$

Then

$$X_{t_i} = X_{t_{i-1}} + \int_{t_{i-1}}^{t_i} a(X_s)\,ds + \int_{t_{i-1}}^{t_i} b(X_s)\,dB_s\,, \quad i = 1, \ldots, n\,. \tag{3.29}$$

The Euler approximation is based on a discretization of the integrals (3.29). To see this, first consider the approximations

$$\int_{t_{i-1}}^{t_i} a(X_s)\,ds = \int_{t_{i-1}}^{t_i} [a(X_s) - a(X_{t_{i-1}})]\,ds + a(X_{t_{i-1}})\Delta_i \approx a(X_{t_{i-1}})\,\Delta_i\,, \tag{3.30}$$

$$\int_{t_{i-1}}^{t_i} b(X_s)\,dB_s = \int_{t_{i-1}}^{t_i} [b(X_s) - b(X_{t_{i-1}})]\,dB_s + b(X_{t_{i-1}})\,\Delta_i B \approx b(X_{t_{i-1}})\,\Delta_i B\,, \tag{3.31}$$

and then replace X_{t_i} with $X_{t_i}^{(n)}$:

$$X_{t_i}^{(n)} = X_{t_{i-1}}^{(n)} + a(X_{t_{i-1}}^{(n)})\,\Delta_i + b(X_{t_{i-1}}^{(n)})\,\Delta B_i\,, \quad i = 1, \ldots, n\,.$$

For obvious reasons, some people call the approximations (3.30) and (3.31) and "Euler freeze". In the latter relations, the integrands $a(X_s) - a(X_{t_{i-1}})$ and $b(X_s) - b(X_{t_{i-1}})$ were replaced with zero. An improvement of the Euler scheme may be expected if we expand the integrals using Itô calculus:

$$\begin{aligned} X_{t_i} - X_{t_{i-1}} &= \int_{t_{i-1}}^{t_i} \left[a(X_{t_{i-1}}) + \int_{t_{i-1}}^{s} \left(aa' + \frac{1}{2} b^2 a'' \right) dy + \int_{t_{i-1}}^{s} ba'\,dB_y \right] ds\,, \\ &\quad + \int_{t_{i-1}}^{t_i} \left[b(X_{t_{i-1}}) + \int_{t_{i-1}}^{s} \left(ab' + \frac{1}{2} b^2 b'' \right) dy + \int_{t_{i-1}}^{s} bb'\,dB_y \right] dB_s \\ &= a(X_{t_{i-1}})\,\Delta_i + b(X_{t_{i-1}})\,\Delta_i B + R_i\,, \end{aligned} \tag{3.32}$$

with remainder

$$R_i = \int_{t_{i-1}}^{t_i} \left[\int_{t_{i-1}}^{s} bb' \, dB_y \right] dB_s + S_i \, .$$

Using the "Euler freeze", the first integral is approximated by

$$\begin{aligned}
& b(X_{t_{i-1}})\, b'(X_{t_{i-1}}) \int_{t_{i-1}}^{t_i} \left(\int_{t_{i-1}}^{s} dB_y \right) dB_s \\
&= b(X_{t_{i-1}})\, b'(X_{t_{i-1}}) \int_{t_{i-1}}^{t_i} (B_s - B_{t_{i-1}})\, dB_s \\
&= b(X_{t_{i-1}})\, b'(X_{t_{i-1}}) \left[\int_{t_{i-1}}^{t_i} B_s \, dB_s - B_{t_{i-1}} \Delta_i B \right] \\
&= b(X_{t_{i-1}})\, b'(X_{t_{i-1}}) \left[\frac{1}{2}[B_{t_i}^2 - t_i] - \frac{1}{2}[B_{t_{i-1}}^2 - t_{i-1}] - B_{t_{i-1}} \Delta_i B \right] \\
&= b(X_{t_{i-1}})\, b'(X_{t_{i-1}})\, \frac{1}{2}[(\Delta_i B)^2 - \Delta_i] \, .
\end{aligned}$$

Thus we get the following expansion

$$X_{t_i} - X_{t_{i-1}} = \left[a(X_{t_{i-1}})\, \Delta_i + b(X_{t_{i-1}})\, \Delta_i B \right] + \left[\frac{1}{2}\, b(X_{t_{i-1}})\, b'(X_{t_{i-1}})\, [(\Delta_i B)^2 - \Delta_i] \right] + S_i \, .$$

Recall that we received the Euler approximation from the first bracket by replacing X_{t_i} with $X_{t_i}^{(n)}$. Proceeding in the same way with the first and second brackets and neglecting the remainder S_i, we arrive at the *Milstein approximation* to the stochastic differential equation (3.28):

$$\begin{aligned}
X_0^{(n)} &= X_0 \, , \\
X_{t_i}^{(n)} &= X_{t_{i-1}}^{(n)} + \left[a(X_{t_{i-1}}^{(n)})\, \Delta_i + b(X_{t_{i-1}}^{(n)})\, \Delta_i B \right] + \left[\frac{1}{2}\, b(X_{t_{i-1}}^{(n)})\, b'(X_{t_{i-1}}^{(n)})\, [(\Delta_i B)^2 - \Delta_i] \right] \, .
\end{aligned}$$

Writing as before $\mathrm{mesh}(\tau_n) = \delta_n$, it can be shown that

$$e_s(\delta_n) = E|X_1 - X_1^{(n)}| \leq c\, \delta_n \, ,$$

for some constant c, provided the coefficient functions a and b and their first derivatives satisfy some Lipschitz and linear growth conditions (Theorem 10.3.5 in Kloeden and Platen [20].) Thus the Milstein approximation scheme is *strongly convergent with order* $\gamma = 1$. This is a substantial improvement upon the Euler scheme with $\gamma = 0.5$. The improvement in the approximation can be seen by eyeball inspection if we compare the paths of X with $X^{(n)}$ for the Euler and Milstein schemes. Clearly, one needs to know the solution X to the stochastic differential equation (3.28) in this case.

A glance at (3.32) also tells us how one can further improve upon the Milstein scheme: further expand the integrals in (3.32) by "freezing" and by using the Itô chain rule for the remainder term.

For example, a 1.5-order strongly converging Taylor-Itô scheme is given by

$$\begin{aligned}
X_0^{(n)} &= X_0\,, \\
X_{t_i}^{(n)} &= X_{t_{i-1}}^{(n)} + a(X_{t_{i-1}}^{(n)})\,\Delta_i + b(X_{t_{i-1}}^{(n)})\,\Delta_i B \quad \text{(Euler scheme)} \\
&\quad + \frac{1}{2}\,b(X_{t_{i-1}}^{(n)})\,b'(X_{t_{i-1}}^{(n)})\,[(\Delta_i B)^2 - \Delta_i] \quad \text{(Milstein scheme)} \\
&\quad + a'(X_{t_{i-1}}^{(n)})\,b(X_{t_{i-1}}^{(n)})\,\Delta_i^{(2)}B + \frac{1}{2}\left(a(X_{t_{i-1}}^{(n)})\,a'(X_{t_{i-1}}^{(n)}) + \frac{1}{2}\,b^2(X_{t_{i-1}}^{(n)})\,a''(X_{t_{i-1}}^{(n)})\right)\,\Delta_i^2 \\
&\quad + \left(a(X_{t_{i-1}}^{(n)})\,b'(X_{t_{i-1}}^{(n)}) + \frac{1}{2}\,b^2(X_{t_{i-1}}^{(n)})\,b''(X_{t_{i-1}}^{(n)})\right)\,[\Delta_i B\,\Delta_i - \Delta_i^{(2)}B] \\
&\quad + \frac{1}{2}\,b(X_{t_{i-1}}^{(n)})\,\left(b(X_{t_{i-1}}^{(n)})\,b''(X_{t_{i-1}}^{(n)}) + (b'(X_{t_{i-1}}^{(n)}))^2\right)\left[\frac{1}{3}\,(\Delta_i B)^2 - \Delta_i\right]\,\Delta_i B\,,
\end{aligned}$$

where

$$\Delta_i^{(2)}B = \int_{t_{i-1}}^{t_i}\int_{t_{i-1}}^{s} dB_y\,ds\,. \tag{3.33}$$

The length of the latter formula already indicates that higher order Taylor-Itô expansions can be quite tedious since one has to evaluate higher order stochastic integrals involving Brownian motion. For example, in order to evaluate $\Delta_i^{(2)}B$ one has to employ approximation schemes as well. However, the idea of this expansion is quite simple and relatively easy to realize on a computer.

Comments

Taylor-Itô expansions of higher order can be found in Kloeden and Platen [20]. The basic idea of Taylor-Itô expansions goes back to Talay [33] and Milstein [23, 24].

There exists evidence from simulations that the Milstein scheme outperforms the Euler scheme as regards accuracy with respect to different error criteria; see Müller [27]. This may not be a surprise in view of the fact that the Milstein scheme converges strongly with order 1 in contrast to 0.5 for the Euler scheme. However, in order to achieve the same accuracy for a given number n of partition points, the Euler scheme is much more costly as regards computer time than the Milstein scheme. Therefore the Euler scheme should be avoided. Simulation evidence also shows that strong approximation schemes of order higher than 1 are computer intensive; the gain in accuracy by higher order schemes is often not justified in view of the increase of computer time.

The Milstein scheme is not easily implemented for multivariate stochastic differential equations, i.e., when the driving process is a multivariate Brownian motion. Then the Taylor-Itô expansion possibly contains multiple Stratonovich integrals of the Brownian components for which no explicit formulae exists. Kloeden and Platen [20] suggest to approximate the multiple Stratonovich integrals by a Fourier series expansion which uses the techniques of the Paley-Wiener decomposition; see Example 3.1.

Higher order schemes simplify if certain derivatives of the coefficient functions a and b vanish. For example, if $b(x) \equiv \sigma$ for some constant σ, $b'(x) = 0$ and therefore the Milstein and the Euler scheme coincide. This applies to the Vasicek model; see (3.20). In this case, the Euler scheme has the strong order 1.

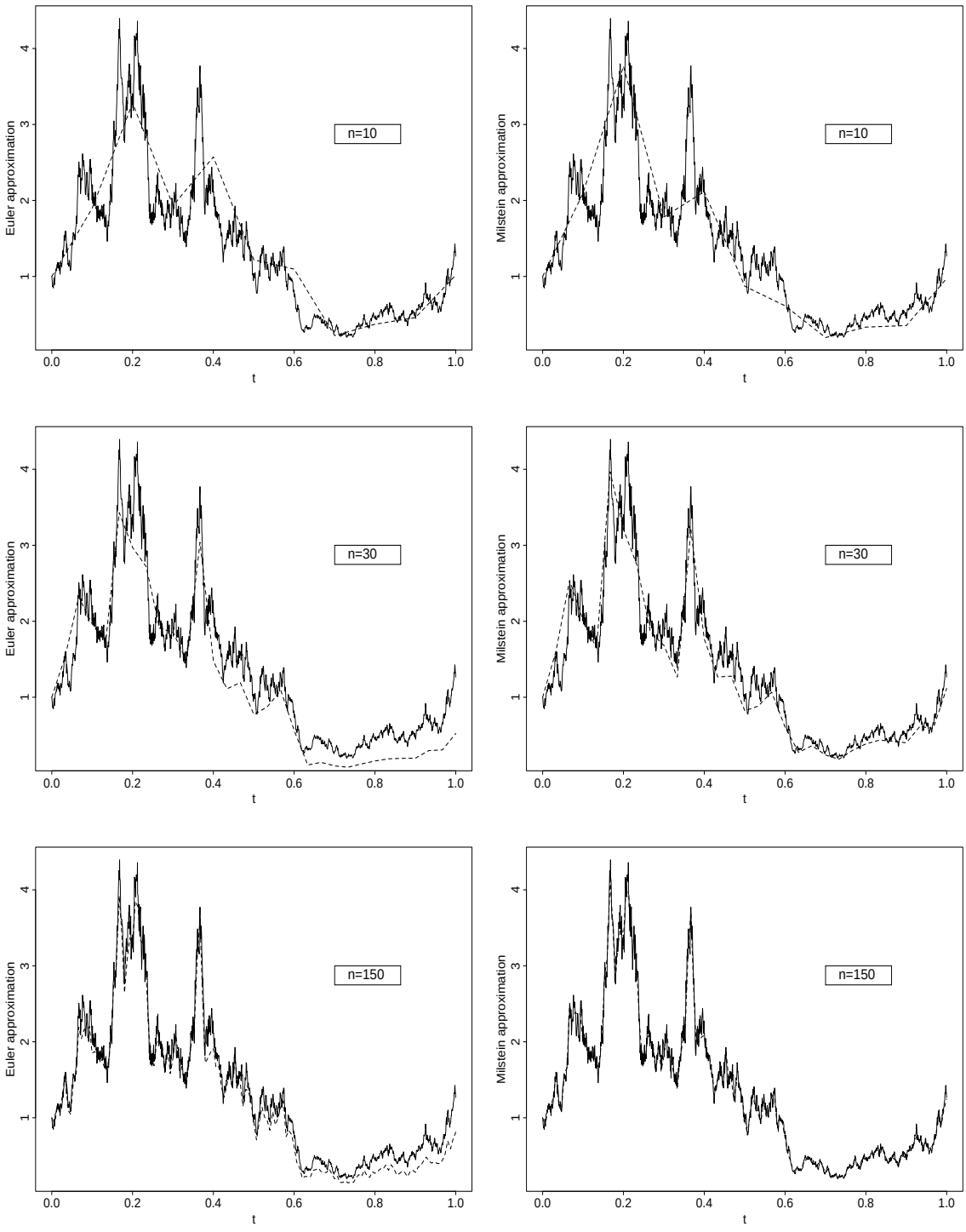

Figure 3.22 *A comparison of the equidistant Euler* (left column) *and Milstein* (right column) *schemes. In every figure a numerical solution* (*dashed lines*) *and the exact solution to the stochastic differential equation* $dX_t = 0.01X_t\,dt + 2X_t\,dB_t$, $t \in [0,1]$, *with* $X_0 = 1$ *are given. Notice the significant improvement of the approximation by the Milstein scheme for large* n, *in particular at the end of the interval.*

3.4.4 Weak approximations

For practical purposes it is often sufficient to have a distributional approximation to the solution of a stochastic differential equation. Indeed, since we never know the Brownian path underlying a stochastic differential equation, simulations are commonly used to get some impression of the shape of the sample paths of the solution (so-called "scenario"), or to determine certain moments (expectation, variance, skewness, kurtosis, etc.) at a fixed instant of time or probabilities and quantiles of the solution at a fixed instant of time. The latter can then be used to define pointwise confidence bands for the solution of the stochastic differential equation.

This concept corresponds to the notion of *weak solution* to a stochastic differential equation. Whereas *strong solution* means that we assume we have a prescribed Brownian sample path for which we obtain a unique solution, *weak solution* means that we are free to choose a Brownian motion for which we then find a solution to the stochastic differential equation.

Example 3.23 (Vasicek model)
In Example 3.19 we learned about the Vasicek interest model which is described by a stochastic differential equation with strong solution

$$r_t = r_0 \, \mathrm{e}^{\,-ct} + \mu \, (1 - \mathrm{e}^{\,-ct}) + \sigma \, \mathrm{e}^{\,-ct} \int_0^t \mathrm{e}^{\,cs} \, dB_s \, .$$

Thus, in order to evaluate the strong solution given by the latter relation, we would have to approximate the Itô integral $\int_0^t \mathrm{e}^{\,cs} \, dB_s$ through a strong numerical approximation for a given Brownian path B, i.e., a Riemann–Stieltjes sum. If one is only interested in a distributional approximation, one may observe that the process r has the same distribution as

$$\widetilde{r}_t = r_0 \, \mathrm{e}^{\,-ct} + \mu \, (1 - \mathrm{e}^{\,-ct}) + \sigma \, \mathrm{e}^{\,-ct} \, \frac{1}{\sqrt{2c}} \, W_{\mathrm{e}^{\,2ct}-1} \tag{3.34}$$

for any Brownian motion W. In order to simulate a weak solution to the underlying stochastic differential equation it thus suffices to simulate (3.34) which is a much easier task. □

A similar simplification of the problem can be observed when it comes to the evaluation of the double integral $\Delta_i^{(2)} B$ in (3.33). Indeed, observe that $\Delta_i^{(2)} B$ is Gaussian with mean zero and variance $\Delta_i^3/3$, and $(\Delta_i B, \Delta_i^{(2)} B)$ are jointly normal with $\mathrm{cov}(\Delta_i B, \Delta_i^{(2)} B) = \Delta_i^2/2$. This allows one to simulate the iid pairs $(\Delta_i B, \Delta_i^{(2)} B)$ in one of the standard ways mentioned on pp. 59–59.

As a matter of fact, various stochastic differential equations have *only* weak solutions. Examples can be found in Chung and Williams [9].

A *weak numerical solution* is any "weak approximation" in the sense that $X^{(n)}$ is close in distribution to X. A criterion to judge the goodness of the approximation to is often based on the quantity

$$e_w(\delta_n) = |Ef(X_1) - Ef(X_1^{(n)})| \, ,$$

where δ_n is again the mesh of the partition and f is chosen from a class of smooth functions, for example polynomials of functions of polynomial growth. The function f is said to be *q-polynomially bounded* if

$$|f(x)| \leq c \, (1 + |x|^q) \, .$$

The weak approximation schemes improve upon the order quite substantially. For example, the Euler scheme has weak order 1, i.e.,

$$e_w(\delta_n) \leq c\,\delta_n^{\gamma}\,,$$

with $\gamma = 1$, where f, a and b have to satisfy additional technical conditions (see Kloeden and Platen [20], Theorem 14.1.5). The Milstein scheme already yields the weak order $\gamma = 2$ under smoothness conditions on f.

Comments

The theory of weak and strong numerical solutions to stochastic differential equation driven by Brownian motion can be found in the monograph by Kloeden and Platen [20]. They treat most cases of practical interest, including the case of multivariate Itô stochastic differential equations. The case of stochastic differential equations driven by processes different from Brownian motion is largely a map with white gaps. In particular, no improvements upon the Euler scheme in the spirit of higher order expansions seem to be known.

References

[1] Adler, R.J. (1990) *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes.* Hayward, California: Institute of Mathematical Statistics.

[2] Asmussen, S. and Rosiński, J. (2001) Approximation of small jumps of Lévy processes with a view towards simulation. *J. Appl. Probab.* **38**, 482–493.

[3] Bachelier, L. (1900) Théorie de la spéculation. *Ann. Sci. École Norm. Sup.* **III–17**, 21–86. Translated in: Cootner, P.H. (Ed.) (1964) *The Random Character of Stock Market Prices,* pp. 17–78. MIT Press, Cambridge, Mass.

[4] Barndorff-Nielsen, O., Mikosch, T. and Resnick, S. (Eds.) (2001) *Lévy Processes. Theory and Applications.* Birkhauser, Boston.

[5] Beran, J. (1994) *Statistics for Long-Memory Processes.* CRC Press, Boca Raton.

[6] Billingsley, P. (1968) *Convergence of Probability Measures.* Wiley, New York.

[7] Billingsley, P. (1995) *Probability and Measure.* 3rd edition. Wiley, Now York.

[8] Borovkov, A.A. and Sahanenko, A.I. (1980) Estimates for the convergence rate in the invariance principle for Banach spaces. *Teor. Veroyatnost. i Primenen. (Th. Probab. Appl.)* **25**, 734–744.

[9] Chung, K.L. and Williams, R.J. (1990) *Introduction to Stochastic Integration.* Second edition. Birkhauser, Boston.

[10] Ciesielski, Z. (1965) *Lectures on Brownian Motion. Heat Conduction and Potential Theory.* Aarhus University Lecture Notes Series, Aarhus.

[11] Donsker, M. (1951) An invariance principle for certain probability limit theorems. *Memoirs Amer. Math. Soc.* **6**.

[12] Doukhan, P., Oppenheim, G. and Taqqu, M.S. (2003) *Long-Range Dependence: Theory and Applications.* Birkhauser, Boston.

[13] Dudley, R.M. (1973) Sample functions of the Gaussian process. *Ann. Probab.* **1**, 66–103.

[14] Dudley, R.M.and Norvaiša, R. (1999)*Differentiability of Six Operators on Nonsmooth Functions and p-Variation.* Lecture Notes in Mathematics, 1703. Springer-Verlag, Berlin.

[15] Einstein, A. (1905) On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat. *Ann. Phys.* **17**.

[16] Golub, G. and Loan, C.F. van (1996) *Matrix Computations.* Johns Hopkins University Press, Baltimore.

[17] Gradshteyn, I.S. and Ryzhik, I.M. (1980) *Table of Integrals, Series, and Products.* Academic Press, New York.

[18] Hall, P. (1982) The order of the approximation to a Wiener process by its Fourier series. *Math. Proc. Cambridge Phil. Soc.* **92**, 547–562.

[19] Hida, T. (1980) *Brownian Motion.* Springer, New York.

[20] Kloeden, P.E. and Platen, E. (1992) *Numerical Solution of Stochastic Differential Equations.* Springer, Berlin.

[21] Kloeden, P.E. and Platen, E. (1994) *Numerical Solution of SDE through Computer Experiments.* Springer, Berlin.

[22] Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces. Isoperimetry and Processes.* Springer-Verlag, Berlin.

[23] Milstein, G.N. (1974) Approximate integration of stochastic differential equations. *Th. Probab. Appl.* **19**, 557–562.

[24] Milstein, G.N. (1978) A method of second-order accuracy integration of stochastic differential equations. *Th. Probab. Appl.* **23**, 396–401.

[25] Mikosch, T. (1990) Almost sure behavior of tail series in functional spaces. *Analysis Math.* **16**, 123–133.

[26] Mikosch, T. and Norvaiša, R. (2000) Stochastic integral equations without probability. *Bernoulli* **6** (2000), 401–434.

[27] Müller, J. (2002) Approximation and simulation of processes and distributions. M.Sc. Thesis, Laboratory of Actuarial Mathematics, Copenhagen.

[28] Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer Series in Statistics. Springer-Verlag, New York.

[29] Protter, P. (1992)*Stochastic Integration and Differential Equations.* Springer, Berlin.

[30] Protter, P. and Talay, D. (1997) The Euler scheme for Lévy driven stochastic differential equations. *Ann. Probab.* **25**, 393–42.

[31] Samorodnitsky, G. and Taqqu, M.S. (1994) *Stable Non-Gaussian Random Processes.* Chapman and Hall, London.

[32] Sato, K. (2000) *Lévy Processes and Infinitely Divisible Distributions.* Cambridge University Press, Cambridge (UK).

[33] Talay, D. (1983) How to discretize a stochastic differential equation? Lecture Notes in Mathematics **972**. Springer, Berlin.

[34] Wiener, N. (1923) Differential space. *J. Math. Phys.* **2**, 131–174.

[35] Young, L.C. (1936) An inequality of Hölder type, connected with Stieltjes integration. *Acta Math.* **67**, 251–282.