Numerical Methods for Mathematical Finance

Peter Philip*

Lecture Notes

Originally Created for the Class of Spring Semester 2010 at LMU Munich,
Revised and Extended for the Class of Winter Semester 2011/2012

March 6, 2023

Contents

*E-Mail: philip@math.lmu.de

1 Motivation: Monte Carlo Methods

1.1 Numerical Integration in High Dimensions

Given a measure space $(\Omega, \mathcal{A}, P)$ (for example, a probability space (cf. Def. A.17 and Def. B.1)), $P(A)$ can be interpreted as a number that measures the size of the set $A \in \mathcal{A}$. The idea of Monte Carlo methods is to use this correspondence in reverse, i.e. to calculate the size of sets by interpreting the size as a probability.

For example, consider a set S with subset $T \subseteq S$, and assume we have a method for randomly and independently drawing elements $s \in S$ and for deciding if $s \in T$. Then we are performing what is known as a sequence of Bernoulli trials, since each instance of our sampling experiment can have precisely two possible outcomes, namely $s \in T$ and $s \notin T$. In the usual way, assigning $s \in T$ the value 1 and $s \notin T$ the value 0, one is employing the model space

$$\begin{aligned} \Omega_0 := \{0,1\}, \quad \mathcal{A}_0 := \mathcal{P}(\Omega_0), \quad A := \{1\}, \\ P_0(A) := p \in [0,1], \quad P_0(\{0\}) = 1-p, \end{aligned} \tag{1.1}$$

where $\mathcal{P}(\Omega_0)$ denotes the power set of Ω_0. Then one is actually simulating a sequence $(X_i)_{i\in\mathbb{N}}$ of random variables (cf. Def. B.2(a)) on the usual product probability space $(\Omega, \mathcal{A}, P)$,

$$\Omega := \Omega_0^{\mathbb{N}}, \quad \mathcal{A} := \mathcal{A}_0^{\mathbb{N}}, \quad P := P_0^{\mathbb{N}}, \tag{1.2}$$

where

$$\bigwedge_{i\in\mathbb{N}} \quad X_i : \Omega \longrightarrow \Omega_0, \quad X_i\big((\omega_k)_{k\in\mathbb{N}}\big) = \omega_i. \tag{1.3}$$

Since the X_i are independent and identically distributed (i.i.d.) with $P_{X_i} = P_0$ for each $i \in \mathbb{N}$ (cf. Def. B.2(b) and Def. B.9), we know $E(X_i) = p$, $\sigma^2(X_i) = p(1-p)$, and the strong law of large numbers Th. B.32 yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \lim_{n\to\infty} \frac{n\,p}{n} = p \quad P\text{-almost surely.} \tag{1.4}$$

In many practical situations, p is a reasonable measure for the size of the set T as a subset of S, and the above procedure provides a way of computing it approximately.

Going one step further, we can use a similar idea to compute integrals, for example of the form $\int_0^1 f(x)\,\mathrm{d}x$ with integrable $f : [0,1] \longrightarrow \mathbb{R}$: Given a probability space $(\Omega, \mathcal{A}, P)$ and a random variable $U : \Omega \longrightarrow [0,1]$ that is uniformly distributed (i.e. $P_U(A) = \lambda_1(A)$ for each $A \in \mathcal{B}^1 \cap [0,1]$), one can write

$$\alpha := \int_0^1 f(x)\,\mathrm{d}x = \int_0^1 f\,\mathrm{d}P_U = E(f \circ U). \tag{1.5}$$

Suppose, we have some method to independently and uniformly draw points from $[0,1]$ (i.e. a method to simulate a sequence $U_1, U_2, \ldots$ of i.i.d. copies of U), then, letting

$$\bigwedge_{n\in\mathbb{N}} \quad S_n := \frac{1}{n} \sum_{i=1}^{n} f(U_i), \tag{1.6}$$

in analogy with (1.4), the strong law of large numbers Th. B.32 provides

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E(f\circ U_i) = \lim_{n\to\infty} \frac{n\,\alpha}{n} = \alpha \quad P\text{-almost surely.} \tag{1.7}$$

If $f \in L^2[0,1]$ is nonconstant, then

$$\begin{aligned}\sigma_f^2 := V(f\circ U_i) &= E\big((f\circ U_i - E(f\circ U_i))^2\big) = \int_0^1 f^2(x)\,\mathrm{d}x \,-\alpha^2\\ &= \int_0^1 \big(f(x)-\alpha\big)^2\,\mathrm{d}x \,> 0.\end{aligned} \tag{1.8}$$

Thus, the Central Limit Th. B.36 applies, yielding

$$\lim_{n\to\infty} \left(\frac{1}{\sigma_f\sqrt{n}}\sum_{i=1}^{n}\big(f\circ U_i-\alpha\big)\right)(P) = N(0,1) \quad \text{in distribution,} \tag{1.9}$$

or, using Th. B.36(b),

$$\begin{aligned}&\lim_{n\to\infty} P\left\{\frac{\sigma_f\,a}{\sqrt{n}} \le \frac{1}{n}\sum_{i=1}^{n}\big(f\circ U_i-\alpha\big) < \frac{\sigma_f\,b}{\sqrt{n}}\right\}\\ &= \lim_{n\to\infty} P\left\{a \le \frac{1}{\sigma_f\sqrt{n}}\sum_{i=1}^{n}\big(f\circ U_i-\alpha\big) < b\right\} = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}\,\mathrm{d}x\end{aligned} \tag{1.10}$$

uniformly in a, b with $-\infty \le a < b \le \infty$. The expressions p_n under the limit sign on the left-hand side of (1.10) measure the probability that the absolute error when estimating α via Monte Carlo simulation using the S_n lies between $\frac{\sigma_f\,a}{\sqrt{n}}$ and $\frac{\sigma_f\,b}{\sqrt{n}}$. And one learns from (1.10) that the error converges to 0 with arbitrarily large probability $p < 1$. More precisely, given $p < 1$, there exist $a, b \in \mathbb{R}$ and $N \in \mathbb{N}$ such that $p_n > p$ for each $n > N$ ((1.10) does not give any information regarding the size of N). Thus, the absolute error converges to 0 with probability $> p$. However, a large constant σ_f can be a problem as the convergence is slow, namely $O(n^{-1/2})$. Recalling that the convergence rate of the standard composite trapezoidal rule

$$\alpha \approx \frac{f(0)+f(1)}{2n} + \frac{1}{n}\sum_{i=1}^{n-1} f\left(\frac{i}{n}\right) \tag{1.11}$$

is $O(n^{-2})$, at least for $f \in C^2[0,1]$ (see [Phi23, Th. 4.35]), it becomes clear, why Monte Carlo is usually not a competitive method for the approximation of 1-dimensional integrals.

However, the situation changes dramatically if $\int_0^1 f$ is replaced by $\int_{[0,1]^d} f$ with large $d \in \mathbb{N}$. The generalization of (1.11) to d dimensions comes with a convergence rate $O(n^{-2/d})$, and this kind of decay of the convergence rate with d is characteristic for all deterministic numerical integration methods. For the Monte Carlo method, everything

can still be carried out as described above, still resulting in the $O(n^{-1/2})$ convergence rate, in general, of course, with worse constants.

Thus, Monte Carlo methods become more and more attractive the higher the dimension d. The slow convergence rate $O(n^{-1/2})$ is a characteristic of Monte Carlo methods and can usually not be helped in situations, where the use of Monte Carlo methods is warranted.

A central goal in mathematical finance is the pricing of financial products known as derivatives. As it turns out, prices of such products can often be represented as expected values given as integrals in high-dimensional, sometimes even infinite-dimensional, spaces. Typically, the price of a derivative depends on quantities (e.g. stock prices, interest rates etc.) evolving according to so-called stochastic processes. Where we had to draw random points from $[0,1]$ or $[0,1]^d$ in the above example, calculating the derivative price using Monte Carlo usually means drawing randomly from a space of paths of a stochastic process.

1.2 Payoff of a European Call Option

The calculation of the payoff of a European call option involves the following quantities, which will be explained in more detail below:

$$\begin{aligned} t: &\quad \text{time}, && (1.12a)\\ T: &\quad \text{strike time}, && (1.12b)\\ S_t: &\quad \text{stock price at time } t, && (1.12c)\\ K: &\quad \text{strike price}, && (1.12d)\\ f^+ := \max\{0, f\}: &\quad \text{positive part of } f, && (1.12e)\\ r: &\quad \text{interest rate}, && (1.12f)\\ \sigma: &\quad \text{volatility}, && (1.12g)\\ C: &\quad \text{payoff}. && (1.12h) \end{aligned}$$

A European call option grants its holder the right to buy a given stock at some fixed strike time T for a fixed strike price K. Here, the present time is assumed to be $t = 0$. Thus, for $T = 0$, the payoff of the option is

$$C(T = 0) = (S_T - K)^+. \tag{1.13}$$

To obtain the (present) value $C(T)$ of the payoff for $T > 0$, it is discounted by the factor e^{-rT}, allowing for a continuously compounded interest rate. Moreover, the stock price S_t, $0 \le t \le T$, is not a constant quantity, but is usually modeled as a random variable evolving according to a stochastic process. In consequence, $C(T)$ is given by the expected value

$$C(T) = E\big(e^{-rT}(S_T - K)^+\big). \tag{1.14}$$

For the expression (1.14) to be meaningful, we have to provide the distribution of the random variable S_T. According to the risk-neutral Black-Scholes model, S_t is the solu-

tion to the stochastic differential equation (SDE)

$$\frac{\mathrm{d}S_t}{S_t} = r\,\mathrm{d}t\, + \sigma\,\mathrm{d}W_t\,, \tag{1.15}$$

where W denotes a standard Brownian motion, a given stochastic process with special properties. Here, (1.15) is the usual shorthand notation for

$$S_t = S_0 + \int_0^t rS_u\,\mathrm{d}u\, + \int_0^t \sigma S_u\,\mathrm{d}W_u\,, \tag{1.16}$$

where the last integral in (1.16) is a so-called Itô integral [Øk03, Sec. 3]. If you are not familiar with these notions, simply accept that a solution to (1.15) is given by

$$S_t = S_0 \exp\left(\left(r - \frac{\sigma^2}{2}\right)\, t + \sigma\, W_t\right), \tag{1.17}$$

where, for each $t \in [0,T]$, S_t and W_t are real-valued random variables. Moreover, W_t is actually $N(0,t)$-distributed. Thus, without changing the distribution, we can write $\sqrt{t}\,Z$, where Z is a generic $N(0,1)$-distributed random variable. This is customarily done, replacing (1.17) with

$$S_t = S_0 \exp\left(\left(r - \frac{\sigma^2}{2}\right)\, t + \sigma\,\sqrt{t}\,Z\right), \tag{1.18}$$

showing that the stock price S_t has a lognormal distribution (i.e. the logarithm of the stock price has a normal distribution). Since S_0 represents the current stock price, its value is assumed to be known, $S_0 \in \mathbb{R}^+$. As it turns out, in this still relatively simple situation, one can obtain an explicit formula for value of the payoff according to (1.14) (exercise):

$$C(T) = E\big(e^{-rT}(S_T - K)^+\big) = \mathrm{BS}(S_0, \sigma, T, r, K), \tag{1.19a}$$

where

$$\begin{aligned}
&\mathrm{BS} : \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}_0^+ \times \mathbb{R}^+ \longrightarrow \mathbb{R}^+,\\
&\mathrm{BS}(s,\sigma,T,r,K)\\
&:= s\,\Phi\left(\frac{\ln(s/K) + (r + \frac{1}{2}\sigma^2)\,T}{\sigma\sqrt{T}}\right) - e^{-rT}\,K\Phi\left(\frac{\ln(s/K) + (r - \frac{1}{2}\sigma^2)\,T}{\sigma\sqrt{T}}\right)
\end{aligned} \tag{1.19b}$$

and

$$\Phi : \mathbb{R} \longrightarrow]0,1[, \quad \Phi(x) := (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\xi^2/2}\,\mathrm{d}\xi \tag{1.20}$$

is the standard normal cumulative distribution function. The result (1.19) is known as the *Black-Scholes formula* for a (European) call option.

In practise, the availability of the explicit formula (1.19), makes using Monte Carlo to estimate the integral $E\big(e^{-rT}\big(S_T - K\big)^+\big)$ unnecessary (and even if there were no explicit formula at hand, according to the discussion in Sec. 1.1, one would rather apply some

deterministic numerical integration method to approximate this 1-dimensional integral). However, the basic structure of this problem is similar to more involved mathematical finance applications, where, in general, Monte Carlo is warranted for value calculations. Therefore, it should be instructive to go through the basic steps one had to take if one were to apply Monte Carlo in the above situation of a European call option. Moreover, simple situations, where a solution via an explicit formula is available, are always good to have in hand for testing one's numerical method.

Where we needed an i.i.d. sequence of random variables uniformly distributed on $[0,1]^d$ for the numerical integration example in Sec. 1.1, we now require the availability of an i.i.d. sequence of $N(0,1)$-distributed random variables $Z_1, Z_2, \ldots$. Given such a sequence, one applies the following algorithm:

$$\begin{aligned}
&\text{set} \quad \hat{C}_0 := 0 \\
&\text{for } i = 1, \ldots, n \\
&\qquad \text{generate } Z_i \in \mathbb{R} \\
&\qquad \text{compute} \quad S_i(T) := S_0 \exp\left(\left(r - \frac{1}{2}\,\sigma^2\right) T + \sigma\sqrt{T} Z_i\right) \in \mathbb{R}^+ \\
&\qquad \text{compute} \quad C_i := e^{-rT}\big(S_i(T) - K\big)^+ \in \mathbb{R}_0^+ \\
&\qquad \text{compute} \quad \hat{C}_i := \hat{C}_{i-1} + C_i/n \in \mathbb{R}_0^+ \\
&\text{return} \quad \hat{C}_n \in \mathbb{R}_0^+
\end{aligned} \tag{1.21}$$

As in previous examples, the strong law of large numbers Th. B.32 applies to yield

$$\lim_{n\to\infty} \hat{C}_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\left(e^{-rT}(S_T - K)^+\right) = E\left(e^{-rT}(S_T - K)^+\right) \quad \text{almost surely,} \tag{1.22}$$

and one can obtain error estimates along the lines of Sec. 1.1.

A feature of the most recent example, that is typical for this kind of mathematical finance application, is the successive computation of several random variables

$$Z_i \longrightarrow S_i(T) \longrightarrow C_i \longrightarrow \hat{C}_i,$$

where randomness (or pseudorandomness) is only applied in the generation of the first step (here, for the Z_i), whereas all successive steps are deterministic.

1.3 Asian Call Options, Path Dependence

In the previous example, the payoff of the derivative security, namely the European call option, depended on the (price of the) underlying asset S in such a way that one only needed to know S_0 and S_T, but no intermediates S_t with $0 < t < T$ – one did not need to know the *path* of S between 0 and T. However, in general, this path of the underlying asset is important for the determination of a derivative security.

For example, the payoff of so-called Asian call options depends on the average stock price

$$\overline{S} := \frac{1}{m} \sum_{j=1}^{m} S_{t_j}, \tag{1.23}$$

where the $0 = t_0 < t_1 < \cdots < t_m = T$ are a fixed finite sequence of agreed-upon dates. The value of the payoff of the Asian call option then is

$$C(T) = E\left(e^{-rT}\big(\overline{S} - K\big)^+\right). \tag{1.24}$$

This can be done following the same strategy as in the previous section, except one now has to generate an independent sequence of (discrete) paths $S_i(t_1), \ldots, S_i(t_m)$ over which to average at the end. To that end, one starts with an i.i.d. sequence of $N(0,1)$-distributed random variables Z_{ij}, where $j = 1, \ldots, m$ for each i. Assuming the same model for the evolution of S as in the previous section, we obtain the following modified version of the algorithm (1.21):

set $\hat{C}_0 := 0$
for $i = 1, \ldots, n$
  set $\overline{S}_{i0} := 0$
  set $S_i(0) := S_0 \in \mathbb{R}^+$
  for $j = 1, \ldots, m$
    generate $Z_{ij} \in \mathbb{R}$
    compute $S_i(t_j) := S_i(t_{j-1}) \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(t_j - t_{j-1}) + \sigma\sqrt{t_j - t_{j-1}}\, Z_{ij}\right)$
    compute $\overline{S}_{ij} := \overline{S}_{i,j-1} + S_i(t_j)/m \in \mathbb{R}^+$
  compute $C_i := e^{-rT}\big(\overline{S}_{im} - K\big)^+ \in \mathbb{R}_0^+$
  compute $\hat{C}_i := \hat{C}_{i-1} + C_i/n \in \mathbb{R}_0^+$
return $\hat{C}_n \in \mathbb{R}_0^+$

(1.25)

Another typical situation, just briefly mentioned at this stage, where one needs to sample paths via Monte Carlo, arises when (1.15) is replaced by a more complicated model, where an explicit solution is not available, for example, if the volatility σ is allowed to depend on the stock price:

$$dS_t = rS_t\,\mathrm{d}t\; + \sigma(S_t)\, S_t\,\mathrm{d}W_t\,. \tag{1.26}$$

As it turns out, the solution to (1.26) can be reasonably approximated by dividing $[0, T]$ into finitely many time steps, e.g. of the equidistant length $\Delta t := T/m$, $m \in \mathbb{N}$, replacing (1.26) by the discrete form

$$S_{t+\Delta t} = S_t + r\, S_t\, \Delta t + \sigma(S_t)\, S_t \sqrt{\Delta t}\, Z \tag{1.27}$$

with $N(0,1)$-distributed Z, and sampling discrete paths, resulting in an algorithm similar to (1.25).

2 Random Number Generation

2.1 Introduction

For the examples in Sec. 1 above, illustrating the use of Monte Carlo methods, we always assumed the availability of a method for generating an i.i.d. sequence of random variables, e.g. uniformly distributed on $[0,1]^d$ in Sec. 1.1 and $N(0,1)$-distributed in Sections 1.2 and 1.3.

The availability of such methods is essential for all Monte Carlo simulation. In the following, we will study how such methods can be obtained.

At the core lies the generation of so-called random numbers or pseudorandom numbers.

Definition 2.1. **(a)** *Random numbers* consist of a sequence $u_1, u_2, \ldots$ in some set S ($S \subseteq \mathbb{R}$ in most cases) such that u_i is in the range of U_i, where $U_1, U_2, \ldots$ is a sequence of S-valued i.i.d. random variables.

(b) *Uniform deviates* are random numbers for uniformly distributed U_i (in situations where this makes sense, S finite and $S = [0,1]$ being the most important examples). Thus, for uniform deviates, within the given range, any number is (ideally) just as likely to be generated as any other.

We will start by studying uniform deviates. We will later see how they can be transformed to provide random variables of different distributions.

2.2 Hardware Random Number Generators

Our main focus will be on generators for pseudorandom numbers, i.e. on the use of deterministic algorithms that generate sequences of numbers that *appear* to be random. However, in the current section, we briefly treat generators for genuine random numbers, known as *hardware random number generators*.

Such generators are always based on physical processes believed to be genuinely random, e.g. radioactive decay or the detection of photons in a double-slit experiment or of photons traveling through a semi-transparent mirror. In the case of radioactive decay, one can measure the lengths $t_1, t_2, \ldots$ of time intervals between two consecutive detections of emissions from a radioactive source. One can then obtain a bitstream by generating a 1 if $t_{n+1} > t_n$, a 0 if $t_{n+1} < t_n$, and no bit if $t_{n+1} = t_n$ (to avoid the introduction of non-randomness due to the resolution of the clock). From the bitstream, one generates the random numbers. According to quantum theory, the generated bits must be completely random, provided a *perfectly working detector*.

Other disadvantages of hardware random number generators include the problems of reproducibility and speed: Large applications can use 10^{12} random numbers. As, for hardware random number generators, there is no algorithm to reproduce the exact sequence, the program run can only be reproduced if the entire sequence is stored – a

nontrivial task if the sequence is long. Likewise, generating random numbers sufficiently fast can be an issue for hardware random number generators.

While hardware random number generators are occasionally used in practise (there exist such generators that provide random bits over the internet), good pseudorandom number generators should be sufficient for most applications.

2.3 Pseudorandom Number Generators

2.3.1 General Considerations

In pseudorandom number generators, the u_i of Def. 2.1 are obtained from a deterministic algorithm. In this case, there is no need to store the u_i for reproducibility, in contrast to the case of the hardware generators considered above. Speed can still be an issue for large applications and should be kept in mind when choosing suitable algorithms and implementations.

Pseudorandom number generators typically have the form

$$x_{i+1} = f(x_i), \quad u_{i+1} = g(x_{i+1}), \tag{2.1}$$

with deterministic functions f and g.

The obvious question is if pseudorandom numbers are at all reasonable to use as a substitute for genuine random numbers. Experience shows that pseudorandom numbers can be successfully employed provided they satisfy certain quality requirements, by which one usually means the numbers should not be detectable as nonrandom by a large range of statistical tests, see below.

Note that, in view of the finite range of numbers that can be represented on a given computer system, it is clear from (2.1) that the x_i (and, hence, the u_i) must become periodic sooner or later. In principle, it can happen that the sequence has an initial nonperiodic part, which could then even become constant in extreme cases. This is not desirable, and we restrict ourselves to periodic sequences:

Definition 2.2. *Pseudorandom numbers* consist of a *periodic* sequence in some set S, mimicking random numbers $u_1, u_2, \ldots$ as defined in Def. 2.1(a).

—

To be useful, the period of a pseudorandom number sequence should neither be too small nor too large, see (b) and (g) below.

The following advise is quoted from [PTVF07, pp. 341–342] (the item labels (a)–(g) are not present in [PTVF07]):

(a) Never use a generator principally based on a *linear congruential generator* (LCG) or a *multiplicative linear congruential generator* (MLCG) [see Sec. 2.3.3 below] ...

(b) Never use a generator with a period less than $\sim 2^{64} \approx 2 \cdot 10^{19}$, or any generator whose period is undisclosed.

(c) Never use a generator that warns against using its low-order bits [because they lack randomness]. That was good advice once, but it now indicates an obsolete algorithm (usually a LCG).

(d) Never use the built-in generators in the *C* and *C++* languages, especially `rand` and `srand`. These have no standard implementation and are often badly flawed.

... You may also want to watch for indications that a generator is overengineered and therefore wasteful of resources:

(e) Avoid generators that take more than (say) two dozen arithmetic or logical operations to generate a 64-bit integer or double precision floating result.

(f) Avoid using generators (over-)designed for serious cryptographic use.

(g) Avoid using generators with period $> 10^{100}$. You *really* will never need it, and, above some minimum bound, the period of a generator has little to do with its quality.

Since we have told you what to avoid from the past, we should immediately follow with the received wisdom of the present:

> An acceptable random generator must combine at least two (ideally, unrelated) methods. The methods combined should evolve independently and share no state. The combination should be by simple operations that do not produce results less random than their operands.

End of quote from [PTVF07].

In terms of statistical tests, it is recommended in [PTVF07] that each method combined to form a good random generator should *individually* pass the so-called DIEHARD tests [Mar03a].

2.3.2 64-Bit Xorshift

Recommended for use in a combined generator, see Sec. 2.3.5.

A 64-bit number can be considered as an element of $\mathbb{Z}_2^{64}$, $\mathbb{Z}_2 = \{0, 1\}$.

'Xor', usually written as 'XOR' refers to the logical operation of *exclusive or*, which turns out to be identical to addition on $\mathbb{Z}_2$:

$$x \text{ XOR } y = x + y \mod 2 \quad \text{for each } x, y \in \mathbb{Z}_2. \tag{2.2}$$

For elements of $\mathbb{Z}_2^n$, $n \in \mathbb{N}$, XOR is applied in the usual componentwise way:

$$x \text{ XOR } y = (x_1 \text{ XOR } y_1, \ldots, x_n \text{ XOR } y_n) \quad \text{for each } x, y \in \mathbb{Z}_2^n. \tag{2.3}$$

In xorshift generators, XOR is combined with the following shift operators:

Definition 2.3. Let $n \in \mathbb{N}$ and $k \in \{0, \ldots, n\}$. The right-shift operators R_k and the left-shift operators L_k are defined by

$$R_k : \mathbb{Z}_2^n \longrightarrow \mathbb{Z}_2^n, \quad R_k(x_1, \ldots, x_n) := (\underbrace{0, \ldots, 0}_{k \text{ times}}, x_1, \ldots, x_{n-k}), \tag{2.4a}$$

$$L_k : \mathbb{Z}_2^n \longrightarrow \mathbb{Z}_2^n, \quad L_k(x_1, \ldots, x_n) := (x_{k+1}, \ldots, x_n, \underbrace{0, \ldots, 0}_{k \text{ times}}), \tag{2.4b}$$

respectively.

Lemma 2.4. *Let $n \in \mathbb{N}$ and $k \in \{0, \ldots, n\}$. Representing $x \in \mathbb{Z}_2^n$ as column vectors, and letting M_k, M_{-k} be the $n \times n$ matrices having only 0-entries, except for 1-entries on the kth superdiagonal (respectively, subdiagonal), i.e.*

$$M_k = (m_{k,i,j}), \quad m_{k,i,j} := \begin{cases} 1 & \text{if } j = i + k, \\ 0 & \text{otherwise,} \end{cases} \quad M_{-k} := M_k^{\mathrm{t}}, \tag{2.5}$$

we obtain for the shift operators

$$M_k x = \left(L_k x^{\mathrm{t}}\right)^{\mathrm{t}} \quad \text{for each } x \in \mathbb{Z}_2^n, \tag{2.6a}$$

$$M_{-k} x = \left(R_k x^{\mathrm{t}}\right)^{\mathrm{t}} \quad \text{for each } x \in \mathbb{Z}_2^n. \tag{2.6b}$$

Proof. One computes

$$(M_k x)_i = \sum_{j=1}^{n} m_{k,i,j}\, x_j = \begin{cases} x_{i+k} & \text{for } i + k \leq n, \\ 0 & \text{otherwise,} \end{cases} = \left(L_k x^{\mathrm{t}}\right)_i^{\mathrm{t}} \tag{2.7a}$$

$$(M_{-k} x)_i = \sum_{j=1}^{n} m_{-k,i,j}\, x_j = \begin{cases} x_{i-k} & \text{for } 1 \leq i - k, \\ 0 & \text{otherwise,} \end{cases} = \left(R_k x^{\mathrm{t}}\right)_i^{\mathrm{t}} \tag{2.7b}$$

thereby establishing the case. ∎

Definition and Remark 2.5. Let $n \in \mathbb{N}$. By a *xorshift* we mean any operation that combines $x \in \mathbb{Z}_2^n$ with a shifted version of x via XOR, i.e. any mapping of the form

$$x \mapsto x \text{ XOR } L_k(x) \quad \text{or} \quad x \mapsto x \text{ XOR } R_k(x) \quad \text{for some } k \in \{0, \ldots, n\}. \tag{2.8}$$

For $n = 16, 32, 64$, we can write (2.8) in *C/C++* notation:

```
x ^= x << k;   or   x ^= x >> k;   for some k ∈ {0, ..., n}.
```
(2.9)

Returning, for the moment, to the situation of a general $n \in \mathbb{N}$, and using that XOR is just addition on $\mathbb{Z}_2$, as discussed above, together with Lem. 2.4, we obtain that a map $X : \mathbb{Z}_2^n \longrightarrow \mathbb{Z}_2^n$ is a xorshift if, and only if,

$$X = X_k := \mathrm{Id} + M_k \quad \text{for a suitable } k \in \{-n, \ldots, n\}. \tag{2.10}$$

In [Mar03b], Marsaglia described the following xorshift-based random number generators:

Definition 2.6. A map $A : \mathbb{Z}_2^{64} \longrightarrow \mathbb{Z}_2^{64}$ is called a 64-*bit xorshift random number generator (RNG)* if, and only if, there exists a triple $(k_1, k_2, k_3) \in \{-64, \ldots, 64\}^3$ such that either $k_1, k_3 > 0$, $k_2 < 0$, or $k_1, k_3 < 0$, $k_2 > 0$, and

$$A = X_{k_3} X_{k_2} X_{k_1}, \tag{2.11}$$

where each X_{k_i} is a xorshift according to (2.10). We will then also write $A(k_1, k_2, k_3)$ instead of A.

—

Thus, each 64-bit xorshift RNG is a composition of precisely 3 xorshifts. Recalling (2.9), for $k_1, k_2, k_3 > 0$, the update step $x \mapsto A(k_1, -k_2, k_3)(x)$ can be implemented in *C* or *C++* using

```
x ^= x << k1;
x ^= x >> k2;
x ^= x << k3;
```
(2.12a)

and $x \mapsto A(-k_1, k_2, -k_3)(x)$ using

```
x ^= x >> k1;
x ^= x << k2;
x ^= x >> k3;
```
(2.12b)

Not all 64-bit xorshift RNG are good generators. First, one is interested in finding triples (k_1, k_2, k_3) such that $A(k_1, k_2, k_3)$ has a full period of $2^{64} - 1$ (the missing value is 0, which is a fixed point of all 64-bit xorshift RNG and must be avoided). One can find such triples using the following Th. 2.10. We start with some preparations, first giving a precise definition of full period:

Definition 2.7. Let K be a finite field, $n \in \mathbb{N}$, and $A : K^n \longrightarrow K^n$ a linear map on the finite vector space K^n. Then A is said to have *full period* if, and only if,

$$O_A(x) := \left\{A^k(x) : k \in \mathbb{N}_0\right\} = K^n \setminus \{0\} \quad \text{for each } x \in K^n \setminus \{0\}. \tag{2.13}$$

The set $O_A(x)$ is sometimes called the *orbit* of x under A.

Lemma 2.8. *Let $n \in \mathbb{N}_0$. The number of polynomials of degree at most n over $\mathbb{Z}_2$ is 2^{n+1}: $\#(\mathbb{Z}_2)_n[x] = 2^{n+1}$, where $(\mathbb{Z}_2)_n[x]$ denotes the $\mathbb{Z}_2$-vector space of polynomials of degree at most n over $\mathbb{Z}_2$.*

Proof. The usual linear isomorphism

$$I : \mathbb{Z}_2^{n+1} \cong (\mathbb{Z}_2)_n[x], \quad I(a_0, \ldots, a_n) := \sum_{i=0}^{n} a_i x^i, \tag{2.14}$$

immediately yields $\#(\mathbb{Z}_2)_n[x] = \#\mathbb{Z}_2^{n+1} = 2^{n+1}$. ∎

Proposition 2.9. *Let $n \in \mathbb{N}$, let K^n be the n-dimensional vector space over the field K, and let $A : K^n \longrightarrow K^n$ be linear. Then, for each $k \in \mathbb{N}_0$, there exists a polynomial P_k of degree less than n over K (i.e. $P_k \in K_{n-1}[x]$) such that*

$$A^k = P_k(A). \tag{2.15}$$

Proof. Exercise. ■

Theorem 2.10 (Marsaglia, Tsay, 1985, see [Mar03b, Sec. 2.1])**.** *Let $n \in \mathbb{N}$. A linear map $A : \mathbb{Z}_2^n \longrightarrow \mathbb{Z}_2^n$ has full period if, and only if, A has order $2^n - 1$, i.e. if, and only if, $A^{2^n-1} = \mathrm{Id}$ and $A^k \neq \mathrm{Id}$ for each $k \in \{1, \ldots, 2^n - 2\}$.*

Proof. Note $\#(\mathbb{Z}_2^n \setminus \{0\}) = 2^n - 1$. First, suppose A has full period. If there were $1 \leq k < 2^n - 1$ and $v \in \mathbb{Z}_2^n \setminus \{0\}$ with $A^k v = v$, then $\#O_A(v) \leq k < 2^n - 1$, i.e. A could not have full period. In particular, $A^k \neq \mathrm{Id}$ for each $k \in \{1, \ldots, 2^n - 2\}$ and, since $\#(\mathbb{Z}_2^n \setminus \{0\}) = 2^n - 1$, $A^{2^n-1}v = v$ for each $v \in \mathbb{Z}_2^n \setminus \{0\}$, showing $A^{2^n-1} = \mathrm{Id}$.

Conversely, suppose A has order $2^n - 1$. Then, the finite sequence $A, A^2, \ldots, A^{2^n-1}$ consists of $2^n - 1$ distinct invertible maps. From Prop. 2.9 we know that each A^k, $1 \leq k < 2^n - 1$, can be represented by a polynomial $P_k \in (\mathbb{Z}_2)_{n-1}[x]$ via (2.15). Since the A^k are all distinct, so are the P_k. However, from Lem. 2.8, we know there are precisely $2^n - 1$ polynomials in $(\mathbb{Z}_2)_{n-1}[x] \setminus \{0\}$. Thus, we can conclude, in reverse, that, for each $P \in (\mathbb{Z}_2)_{n-1}[x] \setminus \{0\}$, there exists $k \in \{1, \ldots, 2^n - 1\}$ such that

$$P(A) = A^k. \tag{2.16}$$

Now suppose there were $l \in \{1, \ldots, 2^n - 2\}$ and $v \in \mathbb{Z}_2^n \setminus \{0\}$ such that $A^l v = v$. Then $A^l - \mathrm{Id}$ is not invertible. On the other hand, $P_l - 1 \in (\mathbb{Z}_2)_{n-1}[x] \setminus \{0\}$, i.e. there is $k \in \{1, \ldots, 2^n - 1\}$ such that

$$A^k = (P_l - 1)(A) = P_l(A) - \mathrm{Id} = A^l - \mathrm{Id}\,. \tag{2.17}$$

Since A^k is invertible, but $A^l - \mathrm{Id}$ is noninvertible, we have a contradiction that shows $A^l v \neq v$ for each $l \in \{1, \ldots, 2^n - 2\}$, $v \in \mathbb{Z}_2^n \setminus \{0\}$. This, in turn, proves that A has full period. ■

Proposition 2.11. *Let G be a group, $g \in G$. If $g^k = e$ for some $k \in \mathbb{N}$, where e denotes the neutral element of G, then the order of g, i.e.*

$$o(g) := \min\{n \in \mathbb{N} : g^n = e\}, \tag{2.18}$$

is a divisor of k.

Proof. Seeking a contradiction, assume $o(g) < k$ is not a divisor of k. Then

$$\exists_{m \in \mathbb{N}} \quad \exists_{r \in \{1, \ldots, o(g)-1\}} \quad k = m \cdot o(g) + r. \tag{2.19}$$

Thus,

$$e = g^k = g^{m \cdot o(g)} g^r = e g^r = g^r, \tag{2.20}$$

in contradiction to $o(g)$ being the smallest number with that property. ■

Theorem 2.12. *Setting $M := 2^{64} - 1$, a linear map $A : \mathbb{Z}_2^{64} \longrightarrow \mathbb{Z}_2^{64}$ (in particular, a map representing a 64-bit xorshift RNG) has full period if, and only if,*

$$A^M = \mathrm{Id}, \quad A^k \neq \mathrm{Id} \quad \textit{for each} \quad k \in \left\{ \frac{M}{6700417}, \frac{M}{65537}, \frac{M}{641}, \frac{M}{257}, \frac{M}{17}, \frac{M}{5}, \frac{M}{3} \right\}. \tag{2.21}$$

Proof. That A having full period implies (2.21) is immediate from Th. 2.10. For the converse, it is noted that

$$M = 3 \cdot 5 \cdot 17 \cdot 257 \cdot 641 \cdot 65537 \cdot 6700417 \tag{2.22}$$

is the prime factorization of $2^{64} - 1$. Thus, each divisor $\neq M$ of M must be a divisor of at least one of the numbers in the set in (2.21). Combining (2.21) with Prop. 2.11 proves that A has order M; then Th. 2.10 implies A has full period. ■

Using Th. 2.12, it is now actually possible to check, for all admissible $(k_1, k_2, k_3) \in \{-64, \ldots, 64\}^3$, if they satisfy (2.21) and, thus, produce a full period xorshift RNG – this obviously requires the use of suitable computer code, where one would use that one can readily compute the needed high powers of A by successive squaring (i.e. from A^2, A^4, ...) (exercise). Still, as it turns out, not all full period triples produce RNG of equal quality. Only some of them pass the DIEHARD tests [Mar03a].

Remark 2.13. Since the DIEHARD tests require a sequence of 32-bit numbers, for a sequence of 64-bit numbers to pass DIEHARD is supposed to mean that both its low and high bits pass DIEHARD. More precisely, the 64-bit sequence

$$(x^{(n)})_{n\in\mathbb{N}} = (x_1^{(n)}, \ldots, x_{64}^{(n)})_{n\in\mathbb{N}}$$

is defined to pass DIEHARD if, and only if, the sequences

$$(x_1^{(n)}, \ldots, x_{32}^{(n)})_{n\in\mathbb{N}} \quad \text{and} \quad (x_{33}^{(n)}, \ldots, x_{64}^{(n)})_{n\in\mathbb{N}}$$

both pass DIEHARD.

—

The following Table 1 shows three triples that, according to [PTVF07], yield 64-bit xorshift RNG that pass DIEHARD (the first 3 out of 9 such triples provided on page 347 of [PTVF07]).

We summarize the 64-bit xorshift RNG that are recommended for use in a combined RNG:

$$\begin{array}{ll} \text{state :} & x_n \in \mathbb{Z}_2^{64} \setminus \{0\} \quad \text{(unsigned 64-bit)}, \quad n \in \mathbb{N}, \\ \text{initialize :} & x_1 \in \mathbb{Z}_2^{64} \setminus \{0\}, \\ \text{update :} & x_{n+1} = X_{k_3} X_{k_2} X_{k_1} x_n \quad \text{according to (2.11)} \\ & \text{with } k_1, k_3 > 0,\ k_2 < 0, \text{ or } k_1, k_3 < 0,\ k_2 > 0 \text{ according to Tab. 1,} \\ \text{period :} & 2^{64} - 1. \end{array} \tag{2.23}$$

ID	$\lvert k_1\rvert$	$\lvert k_2\rvert$	$\lvert k_3\rvert$
A_1	21	35	4
A_2	20	41	5
A_3	17	31	8

Table 1: Triples that, according to [PTVF07], produce 64-bit xorshift RNG that pass DIEHARD. This holds for both forms $k_1, k_3 > 0$, $k_2 < 0$, and $k_1, k_3 < 0$, $k_2 > 0$.

Remark 2.14. A weakness of the 64-bit xorshift RNG lies in the fact that each bit of x_{n+1} depends on at most 8 bits of x_n: If $x, y \in \mathbb{Z}_2^{64}$ and $y = X_k x$ with X_k as in (2.10), $0 < k \leq 64$, then, using (2.7a),

$$\underset{i\in\{1,\ldots,64\}}{\forall} \quad y_i = \begin{cases} x_i + x_{i+k} & \text{for } i + k \leq n, \\ x_i & \text{for } i + k > n, \end{cases} \tag{2.24}$$

i.e. y_i depends on at most x_i and x_{i+k}. Analogously, one sees that y_i depends on at most x_i and x_{i+k} for $k < 0$. Thus, since a 64-bit xorshift RNG A combines precisely 3 xorshifts, each bit of $y = Ax$ depends on at most 8 bits of x.

2.3.3 Linear Congruential Generators

As mentioned before, by itself, a linear congruential generator (LCG) should *not* be used as an RNG. However, when using care, LCG can be useful in combined generators (see Sec. 2.3.5 below).

Definition 2.15. Given numbers $a, c \in \mathbb{N}_0$, $m \in \mathbb{N}$, each map of the form

$$C : \mathbb{Z}_m \longrightarrow \mathbb{Z}_m, \quad C(x) := (a\,x + c) \mod m, \tag{2.25}$$

is called a *linear congruential generator (LCG)* with *modulus* m, *multiplier* a, and *increment* c. If $c = 0$, then the LCG is also called *multiplicative linear congruential generator (MLCG)*. Analogous to Def. 2.7, an LCG is defined to have *full period* if, and only if,

$$O_C(x) := \left\{C^k(x) : k \in \mathbb{N}_0\right\} = \mathbb{Z}_m \quad \text{for each } x \in \mathbb{Z}_m. \tag{2.26}$$

Theorem 2.16. *Let $a, c \in \mathbb{N}_0$, $m \in \mathbb{N}$. Then an LCG according to* (2.25) *has full period if, and only if, the following three conditions are satisfied:*

(i) *c and m are relatively prime, i.e. 1 is their only common divisor.*

(ii) *Each prime number that is a divisor of m is also a divisor of $a - 1$.*

(iii) *If 4 is a divisor of m, then 4 is also a divisor of $a - 1$.*

Proof. See [Knu98, p. 17ff.]. ■

Note that the trivial case $a = c = 1$ shows that, even though desirable, a full period by itself does not ensure a useful LCG. Even if a, c, m are chosen more wisely, LCG have serious weaknesses:

Theorem 2.17. *Let $a, c \in \mathbb{N}_0$, $m \in \mathbb{N}$, and let C be the LCG according to (2.25). Given $x_0 \in \mathbb{Z}_m$, $d \in \mathbb{N}$, define sequences $(x_n)_{n\in\mathbb{N}_0} \in \mathbb{Z}_m$, $(r_n)_{n\in\mathbb{N}_0} \in [0,1]$, $(p_n)_{n\in\mathbb{N}_0} \in [0,1]^d$ by*

$$x_{n+1} := C(x_n), \quad r_n := x_n/m, \quad p_n := (r_n, \ldots, r_{n+d-1}). \tag{2.27}$$

Then there exist $k < (d!\, m)^{1/d}$ parallel $(d-1)$-dimensional hyperplanes $H_1, \ldots, H_k \subseteq \mathbb{R}^d$ such that

$$\{p_n : n \in \mathbb{N}_0\} \subseteq [0,1]^d \cap \bigcup_{i=1}^{k} H_i. \tag{2.28}$$

Proof. See [Mar68, Th. 1]. ■

So, while a true uniformly distributed random variable would tend to fill $[0,1]^d$ uniformly, the output of an LCG is always concentrated in relatively few discrete planes. If a, c, m are not chosen carefully, the number of planes can actually be much smaller than the bound $(d!\, m)^{1/d}$ of the theorem. A number-theoretical test, the so-called *spectral test* (see [Knu98, Sec. 3.3.4]) has been developed to characterize the density of planes of LCG output.

Another serious weaknesses of LCG is related to short periods of low bits if m is chosen as a power of 2:

Remark 2.18. It can be shown (exercise) that if the modulus m of an LCG is a power of 2, then the lowest bit has period ≤ 2, the 2 lowest bits have period ≤ 4, the k lowest bits have period $\leq 2^k$. If, on the other hand, m is not a power of 2, then efficient implementation of (2.25) tends to be challenging.

—

The following Table 2 shows three values for a, c, that, together with $m = 2^{64}$, are recommended in [PTVF07] for the use in combined RNG. According to [PTVF07], in each case, the LCG strongly passes the spectral test [Knu98, Sec. 3.3.4], the high 32 bits almost (but do not quite) pass the DIEHARD tests [Mar03a], whereas the low 32 bits are a complete disaster.

ID	a	c
C_1	3935559000370003845	2691343689449507681
C_2	3202034522624059733	4354685564936845319
C_3	2862933555777941757	7046029254386353087

Table 2: Values for a, c, that, together with $m = 2^{64}$, are recommended in [PTVF07] for the use in combined RNG.

We summarize the 64-bit LCG that are recommended for use in a combined RNG:

$$\begin{aligned} &\text{state :} && x_n \in \mathbb{Z}_2^{64} \quad \text{(unsigned 64-bit)}, \quad n \in \mathbb{N}, \\ &\text{initialize :} && x_1 \in \mathbb{Z}_2^{64}, \\ &\text{update :} && x_{n+1} = (a\, x_n + c) \mod 2^{64} \\ & && \text{with } a, c \text{ according to Tab. 2}, \\ &\text{period :} && 2^{64}. \end{aligned} \tag{2.29}$$

2.3.4 Multiply with Carry

Recommended for use in a combined generator, see Sec. 2.3.5.

We will only consider so-called lag-1 multiply with carry (MWC) RNG. For more general types of MWC RNG, see [CL97].

Definition 2.19. Given numbers $a, b \in \mathbb{N}$, each map of the form

$$B : \mathbb{N}_0 \times \mathbb{Z}_b \longrightarrow \mathbb{N}_0 \times \mathbb{Z}_b, \quad B(c, x) := (c', x'), \tag{2.30a}$$

where

$$x' + c'\, b = a\, x + c \tag{2.30b}$$

is called a *multiply with carry (MWC)* RNG with *multiplier* a and *base* b. The first component of the pairs in (2.30a) is referred to as the *carry* component of the MWC RNG.

Remark 2.20. (a) Note that, due to the requirement $0 \le x' < b$, the numbers c' and x' are uniquely determined by (2.30b). They can be computed as

$$x' = (a\, x + c) \mod b, \quad c' = \left\lfloor \frac{a\, x + c}{b} \right\rfloor, \tag{2.31}$$

where $\lfloor y \rfloor$ denotes the largest integer smaller than or equal to y.

(b) If $c < a$ in (2.30b), then $c' < a$ as well. Indeed,

$$c'\, b = a\, x + c - x' < a\,(x+1) - x' \le a\,(x+1) \le ab, \tag{2.32}$$

i.e. dividing by b proves $c' < a$.

Definition and Remark 2.21. Given an MWC RNG B as in (2.30), $(c, x) \in \mathbb{N}_0 \times \mathbb{Z}_b$ is called a *recurrent state* if, and only if, (c, x) has finite order, i.e. if, and only if, there exists $n \in \mathbb{N}$ such that $B^n(c, x) = (c, x)$. In that case the order of (c, x), which is the same as

$$\#O_B(c, x) = \#\{B^k(c, x) : k \in \mathbb{N}\}, \tag{2.33}$$

is also called the *period* of (c, x) under B, denoted $\pi_B(c, x)$.

As a consequence of Rem. 2.20(b) and $B(0, 0) = (0, 0)$, if $ab > 1$ and $c < a$, then the number $M := ab - 1$ is an upper bound for $\pi_B(c, x)$.

—

Here, we are mostly interested in MWC RNG with base $b = 2^{32}$ and carry $c \in \mathbb{Z}_b$. Representing c as the high bits and $x \in \mathbb{Z}_b$ as the low bits of a 64-bit number y, the update step $y = (c, x) \mapsto B(c, x)$ can be implemented in *C* or *C++* using

```
y = a*(y & 0xffffffff) + (y >> 32);
```
(2.34)

To see that (2.34) does correspond to (2.31), note that `0xffffffff` is the hexadecimal representation of $2^{32} - 1$, i.e. `y & 0xffffffff` $= x$, whereas `y >> 32` $= c$.

Theorem 2.22. *Given $a \in \mathbb{N}$, $b = 2^{32}$, let B be the MWC RNG according to* (2.30)*. Set $M := ab - 1$.*

(a) *If M is prime and $c < a$, then $\pi_B(c, x) < M$.*

(b) *If both M and $(M - 1)/2$ are prime, then there exist $(c, x) \in \mathbb{Z}_b \times \mathbb{Z}_b$ such that $\pi_B(c, x) = (M - 1)/2$.*

Proof. See [CL97]. ■

In consequence of Th. 2.22, one aims at finding a such that both $2^{32}\, a - 1$ and $(2^{32}\, a - 2)/2$ are prime. Several such a exist, where larger a yield larger periods. The following Table 3 shows two such values for a, where the period is close to 2^{64}. These values for a, together with $b = 2^{32}$, are recommended in [PTVF07] for the use in combined RNG. According to [PTVF07], in both cases, the resulting RNG passes the DIEHARD tests [Mar03a] (cf. Rem. 2.13), even though the high 32 bits miss some 8000 values (which a uniformly distributed random variable would obviously not do).

ID	a
B_1	4294957665
B_2	4294963023

Table 3: Values for a, that, together with $b = 2^{32}$, are recommended in [PTVF07] for the use in combined RNG.

We summarize the MWC RNG with base $b = 2^{32}$ that are recommended for use in a combined RNG. Here, as described above, we combine (c, x) into a 64-bit number y. The set Σ of possible states is the orbit of the initial value. It is always a strict subset of $\mathbb{Z}_2^{64}$, and, if the initial value is of the form stated below, then $\#\Sigma = (2^{32}\, a - 2)/2$.

$$
\begin{array}{ll}
\text{state :} & y_n = (c_n, x_n) \in \Sigma \subsetneq \mathbb{Z}_2^{64} \setminus \{0\} \quad \text{(unsigned 64-bit)}, \quad n \in \mathbb{N}, \\
\text{initialize :} & (0, x_1), \text{ where } x_1 \in \mathbb{Z}_2^{32} \setminus \{0\}, \\
\text{update :} & y_{n+1} = (c_{n+1}, x_{n+1}), \quad x_{n+1} = (a\, x_n + c_n) \mod b, \quad c_{n+1} = \lfloor \dfrac{a\, x_n + c_n}{b} \rfloor, \\
 & \text{with } a \text{ according to Tab. 3,} \\
\text{period :} & (2^{32}\, a - 2)/2 \text{ (a prime number).}
\end{array}
\tag{2.35}
$$

2.3.5 Combined Generators

In the previous sections, we described several types of RNG and recommended some of them for use in so-called combined generators. We will actually distinguish between *combined* and *composed* RNG:

Definition 2.23. Let M be some nonempty set.

(a) Given maps $A, B : M \longrightarrow M$, representing RNG,

$$B.A : M \longrightarrow M \times M, \quad x \mapsto \big(B(Ax), Ax\big) \tag{2.36}$$

is called a *composed* RNG (B is composed with A).

(b) Given maps $A_i : M \longrightarrow M$, $i = 1, \ldots, n$, $n \in \mathbb{N}$, representing RNG, and $f : M^n \longrightarrow M$,

$$\begin{aligned} &f.(A_1, \ldots, A_n) : M^n \longrightarrow M^{n+1}, \\ &\quad (x_1, \ldots, x_n) \mapsto \big(f(A_1x_1, \ldots, A_nx_n), A_1x_1, \ldots, A_nx_n\big). \end{aligned} \tag{2.37}$$

is called a *combined* RNG (combining $A_1, \ldots, A_n$).

Remark 2.24. (a) For a composed RNG as in Def. 2.23(a), one wants A to proceed independently of the output of B, i.e., when iterating, the first component of $B.A$ should *not* be fed back into A; one should rather use the second component, i.e. the output of A. For that reason, the *period* of $B.A$ is defined to be the same as the period of A – more precisely, for each $x \in M$ with finite order under A,

$$\pi_{B.A}(x) := \pi_A(x) := \#O_A(x). \tag{2.38}$$

(b) For a combined RNG as in Def. 2.23(b), one wants the A_i to proceed mutually independent, i.e., when iterating, A_ix_i is fed back into A_i rather than $f(A_1x_1, \ldots, A_nx_n)$. In particular, the *period* of $f.(A_1, \ldots, A_n)$ is defined to be the period of the map $A = (A_1, \ldots, A_n)$ – more precisely, for each $x \in M^n$ with finite order under A,

$$\pi_{f.(A_1,\ldots,A_n)}(x) := \pi_A(x) := \#O_A(x). \tag{2.39}$$

Remark 2.25. The following guidelines for forming combined generators are not mathematically precise, but should still be useful:

(a) The combined methods should evolve mutually independent.

(b) The combined methods should not rely on similar algorithms.

(c) The combination should be done such that the output of the combined method is not less random than any one input if the other inputs are kept fixed.

Example 2.26. To illustrate the condition in Rem. 2.25(c), consider 32- or 64-bit arithmetic: Multiplication is *not* suitable for combining generators, since, if one factor is a power of 2, then the low bits of the result will all be zero, no matter how random the second factor is. For 32- or 64-bit arithmetic, one should only combine generators using + or XOR.

—

By combining generators using + or XOR one can increase the size of the state space as well as period. The same can not be done using composed generators. However, the following example shows that composed generators can still be useful:

Example 2.27. Let A_1 be a 64-bit xorshift RNG with values k_1, k_2, k_3 according to the corresponding entry of Table 1 and let C_1 be the 64-bit LCG with values a, c according to the corresponding entry of Table 2. Then $A_1.C_1$ has neither the weakness of C_1 (short period low bits) nor the weakness of A_1 (each bit depends on only a few bits of the previous state).

Example 2.28. As a more complicated example, consider the RNG of [PTVF07, p. 342-343]. It can be written as

$$f.(\pi_1(A_{1,\mathrm{l}}.C_3), A_{3,\mathrm{r}}, B_1), \tag{2.40}$$

where π_1 is the projection $\pi_1(x_1, x_2) = x_1$, $A_{1,\mathrm{l}}, C_3, A_{3,\mathrm{r}}, B_1 : \mathbb{Z}_2^{64} \longrightarrow \mathbb{Z}_2^{64}$,

$$A_{1,\mathrm{l}} = X_{k_3} X_{-k_2} X_{k_1}, \tag{2.41a}$$
$$C_3(x) = (a\,x + c) \mod 2^{64}, \tag{2.41b}$$
$$A_{3,\mathrm{r}} = X_{-k_3} X_{k_2} X_{-k_1}, \tag{2.41c}$$
$$B_1(c, x) = \left(\left\lfloor \frac{a\,x + c}{b} \right\rfloor, (a\,x + c) \mod b \right), \tag{2.41d}$$

with the X_{k_i} according to (2.11), the parameters are according to the corresponding entries of Tables 1,2,3 respectively; and

$$f : (\mathbb{Z}_2^{64})^3 \longrightarrow \mathbb{Z}_2^{64}, \quad f(x, v, w) := \big((x + v) \mod 2^{64}\big) \,\mathrm{XOR}\, w. \tag{2.42}$$

From Rem. 2.24(a),(b), we obtain that the period of $f.(\pi_1(A_{1,\mathrm{l}}.C_3), A_{3,\mathrm{r}}, B_1)$ is given by the period of $(C_3, A_{3,\mathrm{r}}, B_1)$ or, more precisely,

$$\begin{aligned} \pi_{f.(\pi_1(A_{1,\mathrm{l}}.C_3), A_{3,\mathrm{r}}, B_1)}(x) &= 2^{64} \cdot (2^{64} - 1) \cdot (4294957665 \cdot 2^{32} - 2)/2 \approx 3.13854 \cdot 10^{57} \\ &\text{for each } x := (x_1, x_2, 0, x_3) \in \mathbb{Z}_2^{64} \times \big(\mathbb{Z}_2^{64} \setminus \{0\}\big) \times \mathbb{Z}_2^{32} \times \big(\mathbb{Z}_2^{32} \setminus \{0\}\big). \end{aligned} \tag{2.43}$$

Following [PTVF07, p. 342-343], we consider the following implementation in *C++* (see explanation below):

```
typedef unsigned long long int Ullong; // type for 64-bit numbers
struct Ran
{
  Ullong u,v,w;
  Ran(Ullong j) : v(4101842887655102017LL), w(1)
  {
    u = j ^ v; int64();
    v = u; int64();
    w = v; int64();
  }
  inline Ullong int64()
  {
    u = u * 2862933555777941757LL + 7046029254386353087LL;  // (1)
    v ^= v >> 17; v ^= v << 31; v ^= v >> 8;                // (2)
    w = 4294957665U*(w & 0xffffffff) + (w >> 32);           // (3)
    Ullong x = u ^ (u << 21); x ^= x >> 35; x ^= x << 4;    // (4)
    return (x + v) ^ w;                                     // (5)
  }
};
```

Variables `u,v,w` are used to iterate C_3, $A_{3,\mathrm{r}}$, B_1, respectively. The constructor function `Ran(Ullong j)` implements one possibility of initializing `u,v,w`, using only the value of `j`. The combined RNG of (2.40) is implemented as the member function `Ullong int64()`, returning the first component of $f.(\pi_1(A_{1,\mathrm{l}}.C_3), A_{3,\mathrm{r}}, B_1)$, using the values of `u,v,w` as input. It uses the variable `x` to store $\pi_1(A_{1,\mathrm{l}}.C_3)$. We observe $u \mapsto C_3(u)$ according to (2.41b) is implemented in `(1)`, $v \mapsto A_{3,\mathrm{r}}(v)$ according to (2.41c) is implemented in `(2)`, $w \mapsto B_1(w)$ according to (2.41d) is implemented in `(3)`, $u \mapsto x := A_{1,\mathrm{l}}(u)$ according to (2.41a) is implemented in `(4)`, and $(x, v, w) \mapsto f(x, v, w)$ according to (2.42) is implemented in `(5)`.

2.4 Statistical Tests

As mentioned before, before applying RNG to serious applications, one should check the quality of their output by submitting it to statistical tests. Many possible tests have been published in the literature, and a thorough treatment of this subject is beyond the scope of this class. Caveat: Even if an RNG has passed n statistical tests, there is no guarantee that it will not fail test number $n + 1$. However, the *probability* that an RNG is a good generator increases with the number of different tests it passes. The DIEHARD test suite [Mar03a], already mentioned several times, can be seen as a minimum requirement for a good RNG. We will only briefly touch on some of the 15 tests in DIEHARD, after a short discussion of chi-square tests.

2.4.1 Chi-Square Tests

The chi-square test, also written as χ^2-test, is a test to investigate if the vector-valued random variable $(Y_1, \ldots, Y_N)$ (also called a random vector) is multinomially distributed according to a given multinomial distribution. This is the expected distribution if the Y_i count the numbers of how often sample s_i has been drawn in n independent drawings from a finite sample space $S = \{s_1, \ldots, s_N\}$, where the probability of drawing s_i is $p_i > 0$.

We recall the probability-theoretic notions relevant for the description of the chi-square test:

Notation 2.29. Given a finite set $S \neq \emptyset$ and $n \in \mathbb{N}$, let

$$\Sigma(S, n) := \left\{ \vec{k} = (k_s)_{s \in S} \in \mathbb{N}_0^S : \sum_{s \in S} k_s = n \right\} \tag{2.44}$$

denote the set of S-indexed tuples of nonnegative integers that have sum precisely n.

Definition 2.30. Let $n, N \in \mathbb{N}$, let $S = \{s_1, \ldots, s_N\}$, $\#S = N$, be a finite sample space, and $\vec{p} := (p_1, \ldots, p_N) \in (\mathbb{R}^+)^N$ with $\sum_{i=1}^n p_i = 1$.

(a) The discrete probability measure on $\Sigma(S, n)$ defined by

$$\mathcal{M}_{n,\vec{p}}\{\vec{k}\} := \frac{n!}{k_1! \cdots k_N!} p_1^{k_1} \cdots p_N^{k_N} \tag{2.45}$$

is called the *multinomial distribution* with parameters n and $\vec{p}$. If $(\Omega, \mathcal{A}, P)$ is a probability space and $Y := (Y_1, \ldots, Y_N) : \Omega \longrightarrow \Sigma(S, n)$ a random vector, then Y is called $\mathcal{M}_{n,\vec{p}}$-distributed if, and only if, $P_Y = \mathcal{M}_{n,\vec{p}}$.

(b) The map

$$V : \Sigma(S, n) \longrightarrow \mathbb{R}_0^+, \quad V(\vec{k}) := \sum_{i=1}^N \frac{(k_i - np_i)^2}{np_i} = -n + \frac{1}{n} \sum_{i=1}^N \frac{k_i^2}{p_i} \tag{2.46}$$

is called the *chi square statistic* to n and $\vec{p}$; one sometimes says that it has $N - 1$ degrees of freedom (as $k_N = n - \sum_{i=1}^{N-1} k_i$).

Notation 2.31. Recalling the gamma function

$$\Gamma : \mathbb{R}^+ \longrightarrow \mathbb{R}^+, \quad \Gamma(t) := \int_0^\infty u^{t-1} e^{-u} \, \mathrm{d}u \,, \tag{2.47}$$

for each $a, b \in \mathbb{R}^+$, let $\gamma_{a,b}$ denote the function

$$\gamma_{a,b} : \mathbb{R}^+ \longrightarrow \mathbb{R}^+, \quad \gamma_{a,b}(x) := \frac{1}{a^b \Gamma(b)} \, x^{b-1} \, e^{-\frac{x}{a}}. \tag{2.48}$$

Remark 2.32. It follows from (2.47), (2.48), and a simple change of variables that

$$\int_0^\infty \gamma_{a,b}(x)\,\mathrm{d}x \;= 1 \tag{2.49}$$

for each $a, b > 0$. If you are worried about $\gamma_{a,b}$ not being defined at 0, just define it to be any value you like – it does not change the value of the integral.

Definition and Remark 2.33. For each $a, b > 0$, the measure on $\mathcal{B}^1$ defined by

$$\Gamma_{a,b} := \gamma_{a,b}\,\lambda_1, \tag{2.50}$$

is called the *gamma distribution* on $\mathbb{R}$ with parameters a and b (here, $\gamma_{a,b}$ is taken to be extended by 0 to all of $\mathbb{R}$). If $(\Omega, \mathcal{A}, P)$ is a probability space and $X : \Omega \longrightarrow \mathbb{R}$ a random variable such that $P_X = \Gamma_{a,b}$, then one says X is $\Gamma_{a,b}$-distributed. For $\Gamma_{a,b}$-distributed X, one checks

$$E(X) = ab, \tag{2.51a}$$
$$V(X) = a^2 b. \tag{2.51b}$$

Of particular interest is the distribution $\chi^2_n := \Gamma_{2,n/2}$ for $n \in \mathbb{N}$. It is called the *chi-square distribution with n degrees of freedom*. Plugging the parameters into (2.51), for χ^2_n-distributed X:

$$E(X) = n, \tag{2.52a}$$
$$V(X) = 2n. \tag{2.52b}$$

—

The actual chi-square test is now based on the following theorem:

Theorem 2.34. *Consider the situation of Def.* 2.30. *It holds that*

$$\lim_{n\to\infty} \mathcal{M}_{n,\vec{p}}\Big\{\vec{k} \in \Sigma(S,n) : V(\vec{k}) \le c\Big\} = \chi^2_{N-1}[0,c] \quad \textit{for each } c > 0. \tag{2.53}$$

Proof. See, e.g., [Geo09, Th. 11.12]. ■

We can now describe how a chi-square test is carried out. To investigate, if the members of a sequence of S-valued i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$, $X_i : \Omega \longrightarrow S$, is distributed according to $\vec{p} = (p_1, \ldots, p_N)$, choose $n \in \mathbb{N}$ sufficiently large (see below), generate $X_1(\omega), \ldots, X_n(\omega)$, and compute $\vec{k} := Y(\omega) \in \Sigma(S,n)$, where $Y := (Y_1, \ldots, Y_N)$ is defined through

$$Y_i : \Omega \longrightarrow \mathbb{N}_0, \quad Y_i(\omega) := \#\big\{\alpha \in \{1,\ldots,n\} : X_\alpha(\omega) = s_i\big\}. \tag{2.54}$$

If the hypothesis regarding the $X_1, X_2, \ldots$ is true, then Y is $\mathcal{M}_{n,\vec{p}}$-distributed (see, e.g., [Geo09, Th. 2.9]). Thus, in view of Th. 2.34, computing $c := V(\vec{k})$ and $\chi^2_{N-1}[0,c]$ for

$\vec{k} = (k_1, \ldots, k_N) := (Y_1(\omega), \ldots, Y_N(\omega))$, one considers the sequence $X_1, X_2, \ldots$ to have failed the test, provided $\chi^2_{N-1}[0, c]$ is very low or very high: If, for example, $\chi^2_{N-1}[0, c] < 0.05$ or $\chi^2_{N-1}[0, c] > 0.95$, then, in both cases, the probability of $c = V(\vec{k})$ is $< 5\%$, given that the $X_1, X_2, \ldots$ are, indeed, i.i.d. and distributed according to $\vec{p} = (p_1, \ldots, p_N)$ (and given n sufficiently large).

As for the question of how large to choose n, one frequently finds the rule

$$n \geq \frac{5}{\min\{p_i : i = 1, \ldots, N\}} \tag{2.55}$$

in the literature (e.g. [Knu98, p. 45], [Geo09, p. 302]), even though I did not find any source providing a reason for this particular value.

One way of applying the just described chi-square test to RNG is based on the following Lem. 2.35.

Lemma 2.35. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $n \in \mathbb{N}$, $S := \{n_1, \ldots, n_k\} \subseteq \{1, \ldots, n\}$, $\#S = k \leq n$, $\pi : \mathbb{Z}_2^n \longrightarrow \mathbb{Z}_2^k$, $\pi(x_1, \ldots, x_n) = (x_{n_1}, \ldots, x_{n_k})$.*

(a) *If $X : \Omega \longrightarrow \mathbb{Z}_2^n$ is a uniformly distributed random variable, then so is $\pi \circ X$.*

(b) *If $(X_i)_{i \in \mathbb{N}}$, $X_i : \Omega \longrightarrow \mathbb{Z}_2^n$ is a sequence of i.i.d. random variables, uniformly distributed, then the same holds for the sequence $(\pi \circ X_i)_{i \in \mathbb{N}}$.*

Proof. (a): Let $x \in \mathbb{Z}_2^k$. One computes

$$P_{\pi \circ X}\{x\} = P_X\big(\pi^{-1}\{x\}\big) = \frac{\#\pi^{-1}\{x\}}{\#\mathbb{Z}_2^n} = \frac{2^{n-k}}{2^n} = 2^{-k} = \frac{1}{\#\mathbb{Z}_2^k}, \tag{2.56}$$

thereby establishing the case.

(b) is just a combination of (a) with Th. B.10. ∎

Example 2.36. Suppose, we have an RNG represented by a map $A : \mathbb{Z}_2^{64} \longrightarrow \mathbb{Z}_2^{64}$ (for instance one of the RNG considered in Sec. 2.3). We want to apply the chi-square test to investigate the hypothesis that a sequence $A_1, A_2, \ldots$ in $\mathbb{Z}_2^{64}$, generated by A, is likely to be the output of a sequence of uniformly i.i.d. random variables. According to Lem. 2.35, under the hypothesis, the projection π of the A_i onto the lowest 10 bits is also uniformly i.i.d., now on $0, \ldots, 2^{10} - 1 = 1023$ instead of on $0, \ldots, 2^{64} - 1$. Thus, the probability p_i for $i \in \{0, \ldots, 1023\}$ is $1/1024$. The chi-square test now consists of the following steps:

(i) According to (2.55), choose $n \geq 5 \cdot 1024 = 5120$.

(ii) For each $i \in \{0, \ldots, 1023\}$, calculate the number k_i of times i occurs in the sequence $\pi(A_1), \ldots, \pi(A_n)$.

(iii) Set $\vec{k} := (k_0, \ldots, k_{1023})$, compute $c := V(\vec{k})$ according to (2.46), and the corresponding $\chi^2_{1023}[0, c]$.

2.4.2 Binary Rank Tests

3 of the 15 tests in the DIEHARD test suite [Mar03a] are binary rank tests. They are another application of the chi-square tests described in the previous section.

For a binary rank test, one uses the output of an RNG to create binary matrices of some fixed size. For example, if the RNG provides 64-bit numbers, then one can use the lower (or the upper) 32 bits to form the rows of a 32×32 binary matrix, i.e. a matrix over $\mathbb{Z}_2$. Let

$$r_k := \#\big\{A \in \mathbb{Z}_2^{32\times 32} : \operatorname{rk}(A) = k\big\} < 2^{32\cdot 32} = 2^{1024} \tag{2.57}$$

be the number of matrices having rank precisely k (it is a not too difficult combinatorial task to compute explicit formulas for the r_k).

Under the hypothesis (the one to be tested) that the output of the RNG is uniformly distributed, so are the generated matrices in $\mathbb{Z}_2^{32\times 32}$, and the corresponding ranks are distributed according to $\vec{p} := (p_0, \ldots, p_{32})$, $p_k := r_k/2^{1024}$, on $S := \{0, \ldots, 32\}$. So we are in a setting where the chi-square test of Sec. 2.4.1 applies.

In practise, since the values for $p_0, \ldots, p_{28}$ are very small as compared to $p_{29}, \ldots, p_{32}$, one lumps the occurrence of ranks $0, \ldots, 28$ into one event, using the 5-element set $S' := \{\leq 28, 29, 30, 31, 32\}$ instead of S.

2.4.3 A Simple Monkey Test: The CAT Test

2 of the 15 tests in the DIEHARD test suite [Mar03a] are monkey tests, however, much more efficient and powerful than the CAT test. Here, we just give the general idea of such tests.

The name of these tests refers to the well-known thought experiment of a monkey (or a group of monkeys) hitting the keys of a typewriter at random. For simplicity, assume that only the 26 capital letters A,...,Z are possible. One can then apply probability theory to the potential literary output. One can, for instance, count the number of keystrokes it takes the monkey to type CAT for the first time, and one can compute the probability distribution for that number.

Now, in random number generation, the RNG takes the role of the monkey. For example, if the output of the RNG is $x \in \mathbb{Z}_2^{64} \cong \mathbb{Z}_{2^{64}}$, uniformly distributed, then

$$U : \mathbb{Z}_{2^{64}} \longrightarrow \{1, \ldots, 26\}, \quad U(x) := \left\lfloor 26 \cdot \frac{x}{2^{64}} \right\rfloor + 1, \tag{2.58}$$

can be used to obtain a uniformly distributed sequence of capital letters. In particular, every three-letter word is as likely to occur as any other. Since there are $26^3 = 17576$ three-letter words, the RNG monkey should, on average, need 17576 keystrokes to spell the first CAT.

3 Simulating Random Variables

So far, namely in Sec. 2, we have studied the creation of random numbers, simulating i.i.d. random variables, uniformly distributed on some finite set S (and, actually, $S = \mathbb{Z}_2^{64}$ (often tacitly identified with $\mathbb{Z}_{2^{64}} \cong \mathbb{Z}_2^{64}$) for all the primary RNG provided explicitly). We have seen that all known RNG (hardware RNG as well as pseudo RNG) have their issues, and one can even embark on philosophical discussions if true RNG should exist or not.

However, from now on, we will disregard such questions, simply assuming that we are somehow able to simulate a sequence of $\mathbb{Z}_2^{64}$-valued, uniformly i.i.d. random variables $U_1, U_2, \ldots$.

The present section is devoted to the problem of transforming the U_i into a sequence of i.i.d. random variables having some other prescribed probability distribution. Note that, due to Th. B.10, preserving independence is usually not an issue.

3.1 Uniform Deviates

We have already implicitly used that, given $\mathbb{Z}_2^{64}$-valued uniform deviates (recall Def. 2.1(b)), we can obtain $[0,1]^d$-valued uniform deviates. The following Lem. 3.1 provides the precise statement this transformation is founded on:

Lemma 3.1. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $n \in \mathbb{N}$, $S := \{0, \ldots, n-1\}$, $U : S \longrightarrow [0,1]$, $U(k) := k/n$.*

(a) *If $X : \Omega \longrightarrow S$ is a uniformly distributed random variable, then $U \circ X$ is approximately uniformly distributed on $[0,1]$ in the sense that*

$$\lambda_1[a,b] - \frac{1}{n} < P_{U\circ X}[a,b] < \lambda_1[a,b] + \frac{1}{n} \tag{3.1}$$

for each $0 \leq a < b \leq 1$.

(b) *If $(X_1, \ldots, X_d)$, $d \in \mathbb{N}$, $X_i : \Omega \longrightarrow S$, is a tuple of uniformly i.i.d. random variables, then $(U \circ X_1, \ldots, U \circ X_d) : \Omega \longrightarrow [0,1]^d$ is approximately uniformly distributed on $[0,1]^d$ in the sense that*

$$\lambda_d[a,b] - O\left(\frac{1}{n}\right) < P_{(U\circ X_1,\ldots,U\circ X_d)}[a,b] < \lambda_d[a,b] + O\left(\frac{1}{n}\right) \tag{3.2}$$

for each $(0,\ldots,0) \leq a < b \leq (1,\ldots,1)$.

(c) *Let $(X_i)_{i\in\mathbb{N}}$, $X_i : \Omega \longrightarrow S$ be a sequence of uniformly i.i.d. random variables; $d \in \mathbb{N}$. For each $i \in \mathbb{N}$, define*

$$U_i : \Omega \longrightarrow [0,1]^d, \quad U_i(\omega) := \big((U \circ X_{(i-1)d+1})(\omega), \ldots, (U \circ X_{id})(\omega)\big). \tag{3.3}$$

Then $(U_i)_{i\in\mathbb{N}}$ is an i.i.d. sequence of random variables, approximately uniformly distributed in the sense of (b).

Proof. Exercise. ■

Remark 3.2. If the $[0,1]$-valued random variable U is uniformly distributed, then one can consider U as $\mathbb{R}$-valued, where the distribution is given by

$$\underset{t\in\mathbb{R}}{\forall} \quad P\{U<t\} = P\{U\le t\} = \begin{cases} 0 & \text{for } t\le 0, \\ t & \text{for } 0\le t\le 1, \\ 1 & \text{for } 1\le t. \end{cases} \tag{3.4}$$

3.2 Inverse Transform Method

Given a probability measure π on $\mathcal{B}^1$, the inverse transform method is designed to simulate random variables with values in (subsets of) $\mathbb{R}$, being distributed according to π. This is accomplished making use of the corresponding cumulative distribution function (CDF) of π, where it turns out to be convenient to use the right-continuous (r.c.) version $F_{\pi,\mathrm{r}}$ of the CDF (see Def. B.28(b) and Th. B.29). For the sake of readability, we will write $F_\pi := F_{\pi,\mathrm{r}}$ in the following.

We can precisely state our goal as follows: Given a $[0,1]$-valued uniformly distributed random variable U, transform U into a random variable X, that is distributed according to F_π. According to the following Prop. 3.3, this can be accomplished using (a generalization of) the inverse of F_π:

Proposition 3.3. *Let $(\Omega,\mathcal{A},P)$ be a probability space, $U:\Omega\longrightarrow[0,1]$ a uniformly distributed random variable, and $F_\pi:\overline{\mathbb{R}}\longrightarrow[0,1]$, $F_\pi(x):=\pi]-\infty,x]$ the r.c. CDF for some probability measure π on $\mathcal{B}^1$. Define*

$$X:\Omega\longrightarrow\overline{\mathbb{R}},\ X:=\tilde{F}_\pi\circ U, \tag{3.5}$$

where

$$\tilde{F}_\pi:\,[0,1]\longrightarrow\overline{\mathbb{R}},\quad \tilde{F}_\pi(x):=\inf\{y\in\mathbb{R}:\,F_\pi(y)\ge x\} \tag{3.6}$$

(it is $\tilde{F}_\pi=F_\pi^{-1}$ if, and only if, F_π is invertible, i.e. if, and only if, F_π is continuous and strictly increasing). Then X is a random variable distributed according to F_π, i.e. $P_X]-\infty,x]=F_\pi(x)$ for each $x\in\mathbb{R}$.

Proof. From Th. B.29, we know F_π is increasing and r.c. We first show

$$\underset{x\in\mathbb{R}}{\forall} \quad A:=\big\{\omega\in\Omega:\,\tilde{F}_\pi(U(\omega))\le x\big\}=B:=\big\{\omega\in\Omega:\,U(\omega)\le F_\pi(x)\big\}: \tag{3.7}$$

If $\omega\in A$, then $\gamma:=\inf\{y\in\mathbb{R}:\,U(\omega)\le F_\pi(y)\}=\tilde{F}_\pi(U(\omega))\le x$. If $\gamma<x$, then $U(\omega)\le F_\pi(x)$ follows as F_π is increasing, showing $\omega\in B$. If $\gamma=x$, then there is $\gamma_n\downarrow\gamma=x$ with $U(\omega)\le F_\pi(\gamma_n)$ for each $n\in\mathbb{N}$. As F_π is r.c., $U(\omega)\le\lim_{n\to\infty}F_\pi(\gamma_n)=F_\pi(\gamma)=F_\pi(x)$, again showing $\omega\in B$. Conversely, if $\omega\in B$, then $U(\omega)\le F_\pi(x)$, such that $\inf\{y\in\mathbb{R}:\,U(\omega)\le F_\pi(y)\}\le x$, showing $\omega\in A$.

From (3.7), one obtains

$$\underset{x\in\mathbb{R}}{\forall} \quad \begin{aligned} P_X]-\infty,x] &= P\Big(U^{-1}\big(\tilde{F}_\pi^{-1}]-\infty,x]\big)\Big) = P(A) \overset{(3.7)}{=} P(B) \\ &= P\Big(U^{-1}\big(]-\infty,F_\pi(x)]\big)\Big) \overset{(3.4)}{=} F_\pi(x), \end{aligned}$$

establishing the case (note that, due to (3.7), the measurability of U implies the measurability of $\tilde{F}_\pi \circ U$). ■

Definition 3.4. Transforming a $[0,1]$-valued uniformly distributed random variable U into X using (3.5) and (3.6) is called the *inverse transform method*.

Example 3.5. Let $a > 0$. Recall the *exponential distribution*, $E_a := \Gamma_{a^{-1},1}$, a special gamma distribution (cf. Def. and Rem. 2.33; from (2.51), one obtains $E(X) = 1/a$ and $V(X) = 1/a^2$ for $X \sim E_a$). The CDF of the exponential distribution is

$$F : \overline{\mathbb{R}} \longrightarrow [0,1], \quad F(x) = \begin{cases} 0 & \text{for } x \in [-\infty, 0], \\ \int_0^x a\, e^{-at}\,\mathrm{d}t \,= [-e^{-at}]_0^x = 1 - e^{-ax} & \text{for } x \in [0,\infty[, \\ 1 & \text{for } x = \infty. \end{cases} \tag{3.8}$$

Given a $[0,1]$-valued uniformly distributed random variable U, we would like to construct an exponentially distributed random variable X via the inverse transform method. Clearly, F is invertible on $[0,\infty[$ with

$$F^{-1} : [0,1[\longrightarrow [0,\infty[, \quad F^{-1}(x) := -a^{-1}\ln(1-x). \tag{3.9}$$

According to Prop. 3.3, we obtain an exponentially distributed random variable X from (3.5), when using

$$\tilde{F} : [0,1] \longrightarrow \{-\infty\} \cup]0,\infty], \quad \tilde{F}(x) := \begin{cases} -\infty & \text{for } x = 0, \\ -a^{-1}\ln(1-x) & \text{for } 0 < x < 1, \\ \infty & \text{for } x = 1. \end{cases} \tag{3.10a}$$

Since U and $1-U$ have the same distribution, one can also replace the $\tilde{F}$ of (3.10a) with

$$\tilde{F} : [0,1] \longrightarrow \{-\infty\} \cup]0,\infty], \quad \tilde{F}(x) := \begin{cases} -\infty & \text{for } x = 1, \\ -a^{-1}\ln x & \text{for } 1 > x > 0, \\ \infty & \text{for } x = 0. \end{cases} \tag{3.10b}$$

Example 3.6. We apply the inverse transform method to obtain random variables with discrete distributions: Let $S = \{s_1, \ldots, s_n\} \subseteq \mathbb{R}$ be a finite set with $s_1 < \cdots < s_n$. Let the probability measure π on S be defined by $\pi(s_i) := p_i \in [0,1]$, $\sum_{i=1}^n p_i = 1$. Then the r.c. CDF is

$$F_\pi : \overline{\mathbb{R}} \longrightarrow [0,1], \quad F_\pi(x) = \begin{cases} 0 & \text{for } x < s_1, \\ q_i & \text{for } s_i \le x < s_{i+1} \text{ and } i \in \{1, \ldots, n-1\}, \\ 1 & \text{for } s_n \le x, \end{cases} \tag{3.11}$$

where $q_i := \sum_{j=1}^{i} p_j$. According to Prop. 3.3, we obtain a random variable X distributed according to π from (3.5), when using

$$\tilde{F}_\pi : [0,1] \longrightarrow \{-\infty\} \cup S, \quad \tilde{F}_\pi(x) := \inf\{y \in \mathbb{R} : F_\pi(y) \geq x\} = \begin{cases} -\infty & \text{for } x = 0, \\ s_{k(x)} & \text{for } x > 0, \end{cases} \tag{3.12}$$

where $k(x) := \min\big\{i \in \{1, \ldots, n\} : x \leq q_i\big\}$.

Example 3.7. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $U : \Omega \longrightarrow [0,1]$ a uniformly distributed random variable, and $F_\pi : \overline{\mathbb{R}} \longrightarrow [0,1]$, $F_\pi(x) := \pi]-\infty, x]$ the r.c. CDF for some probability measure π on $\mathcal{B}^1$. Suppose the random variable $X : \Omega \longrightarrow \mathbb{R}$ is π-distributed, and we are interested in transforming U into some random variable Y, which is distributed according to X *under the condition* $A := \big\{X \in]a,b]\big\}$, where $a, b \in \overline{\mathbb{R}}$, $a < b$, $F_\pi(a) < F_\pi(b)$ (i.e. $P(A) = F_\pi(b) - F_\pi(a) > 0$). Thus, we are looking for some $Y : \Omega \longrightarrow \mathbb{R}$, such that for each $x \in \mathbb{R}$:

$$\begin{aligned} P\{Y \leq x\} = P\{X \leq x | A\} &= \frac{P\big\{X \in]-\infty, x] \cap]a,b]\big\}}{P(A)} \\ &= \begin{cases} 0 & \text{for } x \leq a, \\ \frac{F_\pi(x) - F_\pi(a)}{F_\pi(b) - F_\pi(a)} & \text{for } a \leq x \leq b, \\ 1 & \text{for } b \leq x. \end{cases} \end{aligned} \tag{3.13}$$

Using the inverse transform method, Y can be constructed in two steps, first letting

$$V : \Omega \longrightarrow [F_\pi(a), F_\pi(b)], \quad V := F_\pi(a) + \big(F_\pi(b) - F_\pi(a)\big)\, U, \tag{3.14}$$

and then

$$Y : \Omega \longrightarrow \overline{\mathbb{R}}, \quad Y := \tilde{F}_\pi \circ V, \tag{3.15}$$

with $\tilde{F}$ according to (3.6). To verify (3.13), we compute, for each $x \in \mathbb{R}$:

$$\begin{aligned} P\{Y \leq x\} &= P\Big(V^{-1}\big(\tilde{F}_\pi^{-1}]-\infty, x]\big)\Big) \overset{(*)}{=} P\Big(V^{-1}\big(]-\infty, F_\pi(x)]\big)\Big) \\ &= P\big\{\omega \in \Omega : V(\omega) \leq F_\pi(x)\big\} = P\left\{\omega \in \Omega : U(\omega) \leq \frac{F_\pi(x) - F_\pi(a)}{F_\pi(b) - F_\pi(a)}\right\} \\ &\overset{(3.4)}{=} \begin{cases} 0 & \text{for } x \leq a, \\ \frac{F_\pi(x) - F_\pi(a)}{F_\pi(b) - F_\pi(a)} & \text{for } a \leq x \leq b, \\ 1 & \text{for } b \leq x, \end{cases} \end{aligned} \tag{3.16}$$

showing that Y has, indeed, the desired conditional distribution. At "(*)", (3.7) was used with U replaced by V.

—

The inverse transform method is usually not the most efficient way of obtaining a random variable with a desired distribution. In general, its feasibility and efficiency depend on efficient algorithms for computing $\tilde{F}_\pi$ being at hand or not. If F_π is invertible, then Newton's method can sometimes be useful to compute $\tilde{F}_\pi(u) = F_\pi^{-1}(u)$ as the solution x to the root problem $F_\pi(x) - u = 0$.

The inverse transform method, at least in the form that was presented here, only yields univariate (i.e. 1-dimensional) distributions.

3.3 Acceptance-Rejection Method

The goal of the acceptance-rejection method is still to simulate random variables Y that have some desired given distribution. It is not restricted to $\mathbb{R}$-valued Y, but also works for Y with values in $\mathbb{R}^d$ or even in a more complicated space Ω'.

The idea of this method, also called the von Neumann acceptance-rejection method, is to first simulate values of a more conveniently distributed random variable X. The output $X(\omega)$ is then rejected with a certain probability, where the rejection mechanism is designed such that the accepted values are, indeed, distributed according to the desired distribution, i.e. they can be used to represent Y.

We will restrict ourselves to formulating the acceptance-rejection method for the case, where the distribution of Y is given via a probability density function (PDF, see Def. B.26(b)). It is noted that variants also exist for situations, where Y is not given via a PDF.

The acceptance-rejection method is based on the following proposition:

Proposition 3.8. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\Omega', \mathcal{A}', \mu)$ a measure space, $Y, X : \Omega \longrightarrow \Omega'$ random variables with PDF $f, g : \Omega' \longrightarrow \mathbb{R}_0^+$, respectively (i.e. f PDF for Y and g PDF for X), $g > 0$, and $c > 1$ such that*

$$f(x) \leq c\, g(x) \quad \text{for each } x \in \Omega'. \tag{3.17}$$

If $U : \Omega \longrightarrow [0,1]$ is a uniformly distributed random variable, which is independent of X, then the distribution of Y is the conditional distribution of X under the condition

$$C := \left\{ U \leq \frac{f \circ X}{c(g \circ X)} \right\}, \tag{3.18}$$

that means

$$P\{Y \in A\} = \int_A f \,\mathrm{d}\mu = \frac{P\left\{X \in A,\, U \leq \frac{f \circ X}{c(g \circ X)}\right\}}{P(C)} \quad \text{for each } A \in \mathcal{A}'. \tag{3.19}$$

Proof. Exercise. ■

Definition and Remark 3.9. Given the situation of Prop. 3.8 and a method for generating sequences $(u_1, x_1), (u_2, x_2), \ldots$ of values of independent copies of (U, X), the

acceptance-rejection method returns (i.e. accepts) x_i as a value of Y if, and only if, $u_i \leq f(x_i)/(c\,g(x_i))$ (otherwise, x_i is rejected). Then the distribution of the returned values is that of X under the condition C, with C as in (3.18), and Prop. 3.8 yields that the returned x_i are, indeed, the values of a random variable Y, distributed according to the PDF f.

Remark 3.10. In the situation of Prop. 3.8, (3.17) can not hold with $c < 1$ and it can hold with $c = 1$ only if $f = g$ μ-almost everywhere (exercise). It is desirable to find g such that c is close to 1, such that as few generated values from X as possible need to be rejected. On the other hand, a competing requirement is that g and X need to be efficiently computable. It can be an art to find a suitable g for a given f such that g and X are efficiently computable *and* g makes c small (i.e. close to 1). In this context, if $h(x) := f(x)/(cg(x))$ is computationally costly to evaluate, acceptance-rejection can often be made more efficient by the so-called *squeeze method*: One finds functions h_1, h_2 that can be evaluated more quickly and that form a squeeze of h in the sense that $h_1 \leq h \leq h_2$. One now needs to evaluate $h(x_i)$ only if $h_1(x_i) \leq u_i \leq h_2(x_i)$. Of course, finding good squeezes can also be an art.

Example 3.11. We show how the acceptance-rejection method can be applied to obtain $N(0,1)$-distributed random variables. Other methods for sampling univariate normally distributed random variables (actually, more efficient ones) will be discussed in Sec. 3.4.1 below.

In the present example, the goal is to apply acceptance-rejection to obtain an $N(0,1)$-distributed Y, i.e. a Y with PDF

$$f : \mathbb{R} \longrightarrow [0,1], \quad f(x) := \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}. \tag{3.20}$$

For g, we use the PDF of a so-called *double exponential* distribution, i.e.

$$g : \mathbb{R} \longrightarrow [0,1], \quad g(x) := \frac{1}{2}\, e^{-|x|}. \tag{3.21}$$

Then, for each $x \in \mathbb{R}$,

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}}\, e^{-x^2/2+|x|} \leq \sqrt{\frac{2e}{\pi}} =: c = 1.3154... \approx 1.3155, \tag{3.22}$$

since $-x^2/2 + |x|$ attains its max $1/2$ at $x = 1$. Thus, (3.17) is satisfied and the acceptance-rejection method according to Def. and Rem. 3.9 can be applied, where a value x_i is accepted if, and only if,

$$u_i \leq \frac{f(x_i)}{cg(x_i)} = e^{-x_i^2/2+|x_i|-1/2} = e^{-(|x_i|-1)^2/2}. \tag{3.23}$$

If $(\Omega, \mathcal{A}, P)$ is a probability space, $Z : \Omega \longrightarrow \mathbb{R}_0^+$ is an E_1-distributed (i.e. exponentially distributed, cf. Ex. 3.5) random variable and $V : \Omega \longrightarrow \{-1, 1\}$ is a uniformly

distributed random variable, Z, V mutually independent, then $ZV : \Omega \longrightarrow \mathbb{R}$ is doubly exponentially distributed: Indeed, for each $0 < a < b$,

$$\begin{aligned} P\big\{ZV \in [a,b]\big\} &= P\big\{Z \in [a,b],\, V = 1\big\} = P\big\{Z \in [a,b]\big\} P\{V = 1\} \\ &= \frac{1}{2} \int_a^b e^{-x}\, \mathrm{d}x \;= \int_a^b g(x)\, \mathrm{d}x\,, \end{aligned} \tag{3.24}$$

and analogously for $a < b < 0$.

We already know how to generate sequences $(u_i)_{i\in\mathbb{N}}$, $(v_i)_{i\in\mathbb{N}}$, $(z_i)_{i\in\mathbb{N}}$, such that the u_i represent i.i.d. copies of $[0,1]$-valued uniformly distributed random variables, the v_i represent i.i.d. copies of $\{-1,1\}$-valued uniformly distributed random variables, and the z_i represent i.i.d. copies of $\mathbb{R}_0^+$-valued E_1-distributed random variables (see Ex. 3.5). If we generate the three sequences independently, then, using the acceptance-rejection method, the following algorithm yields $y_1, y_2, \ldots$, representing i.i.d. copies of $N(0,1)$-distributed random variables. We only describe one step of the algorithm:

$$\begin{aligned} &1: \quad \text{generate } u_i \text{ and } z_i \\ &2: \quad h_i := f(z_i)/(cg(z_i)) \\ &3: \quad \text{if } u_i \le h_i \text{ then} \\ &\qquad\quad k := k+1;\ \text{generate } v_k;\ y_k := z_i v_k \\ &4: \quad i := i+1;\ \text{goto } 1 \end{aligned} \tag{3.25}$$

Note that, since f and g are both symmetric with respect to $x = 0$, in Steps 2 and 3, we can use the $z_i \in \mathbb{R}_0^+$ to perform the acceptance test according to (3.23). This means, one can avoid generating v_k for rejected values.

Example 3.12. In Ex. 3.7, the inverse transform method was used to simulate particular conditional distributions. Now consider the general case: Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\Omega', \mathcal{A}')$ a measurable space, $X : \Omega \longrightarrow \Omega'$ a random variable, and $C \in \mathcal{A}'$ with $P\{X \in C\} > 0$. Suppose we know how to simulate the distribution of X, but would like to simulate the distribution of X under the condition $X \in C$. If $x_1, x_2, \ldots$ represent values of i.i.d. copies of X, then one can always obtain $y_1, y_2, \ldots$ representing the conditional distribution by the following brute force algorithm:

$$\begin{aligned} &1: \quad \text{generate } x_i \\ &2: \quad \text{if } x_i \in C \text{ then} \\ &\qquad\quad k := k+1;\ y_k := x_i \\ &3: \quad i := i+1;\ \text{goto } 1 \end{aligned} \tag{3.26}$$

If μ is a measure on $(\Omega', \mathcal{A}')$ and $g : \Omega' \longrightarrow \mathbb{R}^+$ is a PDF for X, then

$$\frac{P\{X \in A \cap C\}}{P\{X \in C\}} = \frac{1}{P\{X \in C\}} \int_A g\, \chi_C \, \mathrm{d}\mu \quad \text{for each } A \in \mathcal{A}', \tag{3.27}$$

i.e. $f := g\,\chi_C / P\{X \in C\}$ is a PDF for the conditional distribution. Since

$$f = \frac{g\,\chi_C}{P\{X \in C\}} \le \frac{g}{P\{X \in C\}}, \tag{3.28}$$

(3.17) holds, and (3.26) is precisely the acceptance-rejection method of Def. and Rem. 3.9, where the acceptance test $u_i \leq P\{X \in C\} f(x_i)/g(x_i)$ does not depend on $u_i > 0$ and reduces to testing $x_i \in C$.

3.4 Normal Random Variables and Vectors

Normal random variables and vectors are of particular importance for mathematical finance applications (cf. (1.21) and (1.25)), and, thus, we devote special attention to techniques for their efficient simulation in the present section.

3.4.1 Simulation of Univariate Normals

In the present section, we consider methods for simulating random variables that have a univariate, i.e. one-dimensional normal distribution $N(\alpha, \sigma^2)$.

Lemma 3.13. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $Z : \Omega \longrightarrow \mathbb{R}$ an $N(0,1)$-distributed random variable. Then, given $\alpha \in \mathbb{R}$, $\sigma > 0$,*

$$X : \Omega \longrightarrow \mathbb{R}, \quad X := \alpha + \sigma Z, \tag{3.29}$$

is $N(\alpha, \sigma^2)$-distributed.

Proof. Letting $f : \mathbb{R} \longrightarrow \mathbb{R}$, $f(x) = \alpha + \sigma x$, we have $X = f \circ Z$. For each $A \in \mathcal{B}^1$, we have

$$P_X(A) = P_{f\circ Z}(A) = P_Z\big(f^{-1}(A)\big) = \frac{1}{\sqrt{2\pi}} \int_{f^{-1}(A)} e^{-\xi^2/2}\, \mathrm{d}\xi = \frac{1}{\sqrt{2\pi\sigma^2}} \int_A e^{-\frac{(x-\alpha)^2}{2\,\sigma^2}}\, \mathrm{d}x\,, \tag{3.30}$$

thereby establishing the case. ∎

In view of Lem. 3.13, we will restrict ourselves to studying methods for simulating $N(0,1)$-distributed random variables.

A first possibility for generating $N(0,1)$-distributed random variables was already described in Ex. 3.11 as an application of the acceptance-rejection method. As alternatives, we will now discuss the so-called *Box-Muller method* [BM58] and its more efficient variant due to Marsaglia and Bray [MB64].

The Box-Muller method is based on the following proposition:

Proposition 3.14. *Let $(\Omega, \mathcal{A}, P)$ be a probability space. If $U_1, U_2 : \Omega \longrightarrow]0,1[$ are independent and uniformly distributed random variables, $U : \Omega \longrightarrow]0,1[^2$, $U := (U_1, U_2)$, and*

$$f :]0,1[^2 \longrightarrow \mathbb{R}^2, \quad f(u,v) := \sqrt{-2\ln u}\,\big(\sin(2\pi v), \cos(2\pi v)\big), \tag{3.31}$$

then $Z := f \circ U : \Omega \longrightarrow \mathbb{R}^2$ is distributed according to the standard bivariate normal distribution (cf. Sec. 3.4.2 below), i.e. for each $a, b \in \mathbb{R}^2$ with $a < b$:

$$P\big\{Z \in [a,b]\big\} = \frac{1}{2\pi} \int_{a_1}^{b_1} \int_{a_2}^{b_2} e^{-\frac{x^2+y^2}{2}}\, \mathrm{d}y\, \mathrm{d}x\,. \tag{3.32}$$

In particular, the two components Z_1 and Z_2 of Z are independent and both $N(0,1)$-distributed.

Proof. The proof is an easy consequence of the change of variables formula, for which we need to compute the Jacobian of f. The derivative is

$$Df(u,v) = \begin{pmatrix} -\frac{\sin(2\pi v)}{u\sqrt{-2\ln u}} & 2\pi\sqrt{-2\ln u}\cos(2\pi v) \\ -\frac{\cos(2\pi v)}{u\sqrt{-2\ln u}} & -2\pi\sqrt{-2\ln u}\sin(2\pi v) \end{pmatrix}, \tag{3.33}$$

yielding the Jacobian determinant

$$\det Df(u,v) = \frac{2\pi}{u}. \tag{3.34}$$

Thus, to use the change of variables $(x,y) := f(u,v)$, we need to compute u in terms of (x,y) – more precisely, we need to compute the first component of f^{-1}. If $(x,y) := f(u,v)$, then

$$x^2+y^2 = -2\ln u\big(\sin^2(2\pi v)+\cos^2(2\pi v)\big) = -2\ln u \quad \Rightarrow \quad u = e^{-\frac{x^2+y^2}{2}}. \tag{3.35}$$

Thus, for each $A \in \mathcal{B}^2$:

$$\begin{aligned} P_Z(A) &= P_{f\circ U}(A) = P_U\big(f^{-1}(A)\big) = \int_{f^{-1}(A)} 1\,\mathrm{d}(u,v) \\ &\overset{(*)}{=} \int_A \Big(\det Df\big(f^{-1}(x,y)\big)\Big)^{-1}\,\mathrm{d}(x,y) \\ &= \int_A \frac{u(x,y)}{2\pi}\,\mathrm{d}(x,y) \;=\; \frac{1}{2\pi}\int_A e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}(x,y)\,, \end{aligned} \tag{3.36}$$

where the change of variables formula has been used at $(*)$ with $(x,y) := f(u,v)$. ■

Definition and Remark 3.15. If $u_1, u_2, \ldots$ denote the output of a sequence of i.i.d. copies of random variables uniformly distributed on $[0,1]$, then the *Box-Muller method* is given by the following algorithm:

$$\begin{aligned} &1: \quad \text{generate } u_i \neq 0 \text{ and } u_{i+1} \\ &2: \quad r := \sqrt{-2\ln u_i};\ \gamma := 2\pi u_{i+1} \\ &3: \quad z_i := r\cos\gamma;\ z_{i+1} := r\sin\gamma \\ &4: \quad i := i+2;\ \text{goto } 1 \end{aligned} \tag{3.37}$$

According to Prop. 3.14, $z_1, z_2, \ldots$ represent the output of i.i.d. copies of $N(0,1)$-distributed random variables.

—

The idea of the Marsaglia-Bray algorithm is to modify (3.37) such that the computationally costly evaluations of cos and sin can be avoided. To that end, they provide a procedure for simulating random variables that are uniformly distributed on the unit circle S.

Remark 3.16. Let

$$S := \big\{(x_1, x_2) \in \mathbb{R}^2 : \, x_1^2 + x_2^2 = 1\big\}. \tag{3.38}$$

It is an exercise to show the conclusion of Prop. 3.14 still holds if U_2 is replaced by a random variable $W = (W_1, W_2) : \, \Omega \longrightarrow S$, where W is uniformly distributed, U_1 and W are independent, and f is replaced by

$$g :]0,1[\times S \longrightarrow \mathbb{R}^2, \quad g(u, w_1, w_2) := \sqrt{-2 \ln u}\,(w_1, w_2). \tag{3.39}$$

—

The Marsaglia-Bray algorithm for simulating random variables uniformly distributed on S is based on the following proposition:

Proposition 3.17. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $U_1, U_2 : \, \Omega \longrightarrow]0,1[$ be independent and uniformly distributed random variables.*

(a) *Letting*

$$V_i : \, \Omega \longrightarrow]-1,1[, \quad V_i := 2U_i - 1, \quad i \in \{1,2\}, \tag{3.40}$$

the V_i are independent and uniformly distributed random variables.

(b) *$V := (V_1, V_2) : \, \Omega \longrightarrow]-1,1[^2$ is a uniformly distributed random variable. In particular, the distribution of V under the condition $\{V \in C\}$,*

$$C := \big\{(x_1, x_2) \in \mathbb{R}^2 : \, x_1^2 + x_2^2 \leq 1\big\}, \tag{3.41}$$

is the uniform distribution on C.

(c) *Letting*

$$U : \, \Omega \longrightarrow [0,2[, \quad U := V_1^2 + V_2^2, \tag{3.42}$$

the distribution of U, under the condition $\{V \in C\}$, is the uniform distribution on $[0,1]$.

(d) *Recalling S from* (3.38) *and letting*

$$W := (W_1, W_2) : \, \Omega \longrightarrow S, \quad W_i : \, \Omega \longrightarrow [0,1], \quad W_i := V_i/\sqrt{U}, \quad i \in \{1,2\}, \tag{3.43}$$

the distribution of W, under the condition $\{0 < U \leq 1\}$, is uniformly distributed on S.

(e) *Under the condition $\{0 < U \leq 1\}$, U and W are independent.*

Proof. (a): Note $V_i = f \circ U_i$ with $f :]0,1[\longrightarrow]-1,1[$, $f(x) := 2x - 1$. The independence is due to Th. B.10; for the distribution note, for each Borel set A in $]-1,1[$:

$$P_{V_i}(A) = P_{f \circ U_i}(A) = P_{U_i}\big(f^{-1}(A)\big) = \int_{f^{-1}(A)} 1 \, \mathrm{d}\xi \; = \frac{1}{2} \int_A 1 \, \mathrm{d}x \, . \tag{3.44}$$

(b) clearly holds as a simple consequence of the independence of V_1 and V_2.

(c): For each $0 \le a < b \le 1$, denote the corresponding annulus by

$$A_{a,b} := \big\{(x_1, x_2) \in \mathbb{R}^2 : \, a \le x_1^2 + x_2^2 \le b\big\}. \tag{3.45}$$

We compute

$$\begin{aligned} \forall_{0\le a<b\le 1} \quad \frac{P\big\{U \in [a,b]\big\}}{P\{V \in C\}} &= \frac{P\big\{V_1^2 + V_2^2 \in [a,b]\big\}}{\frac{1}{4}\lambda_2(C)} = \frac{P\big\{V \in A_{a,b}\big\}}{\frac{1}{4}\pi} = \frac{\frac{1}{4}\lambda_2(A_{a,b})}{\frac{1}{4}\pi} \\ &= \frac{\frac{1}{4}\pi(b-a)}{\frac{1}{4}\pi} = b - a. \end{aligned} \tag{3.46}$$

(d): For each $0 \le \alpha < \beta \le 2\pi$, denote the corresponding segments of S and C by

$$S_{\alpha,\beta} := \big\{(x_1, x_2) \in \mathbb{R}^2 : \, x_1 = \sin\gamma, \, x_2 = \cos\gamma, \, \alpha \le \gamma \le \beta\big\}, \tag{3.47a}$$

$$C_{\alpha,\beta} := \big\{(x_1, x_2) \in \mathbb{R}^2 : \, x_1 = r\sin\gamma, \, x_2 = r\cos\gamma, \, \alpha \le \gamma \le \beta, \, r \in [0,1]\big\}, \tag{3.47b}$$

respectively. We compute

$$\forall_{0\le\alpha<\beta\le 2\pi} \quad \frac{P\big\{W \in S_{\alpha,\beta}\big\}}{P\{V \in C\}} = \frac{P\{V \in C_{\alpha,\beta}\}}{\frac{1}{4}\pi} = \frac{\frac{1}{4}\lambda_2(C_{\alpha,\beta})}{\frac{1}{4}\pi} = \frac{\frac{1}{4}\,\frac{\beta-\alpha}{2\pi}\,\pi}{\frac{1}{4}\pi} = \frac{\beta-\alpha}{2\pi}. \tag{3.48}$$

(e): For each $0 \le a < b \le 1$, $0 \le \alpha < \beta \le 2\pi$:

$$\begin{aligned} \frac{P\big\{U \in [a,b],\, W \in S_{\alpha,\beta}\big\}}{P\{V \in C\}} &= \frac{P\{V \in A_{a,b} \cap C_{\alpha,\beta}\}}{\frac{1}{4}\pi} = \frac{\frac{1}{4}\lambda_2(A_{a,b} \cap C_{\alpha,\beta})}{\frac{1}{4}\pi} = \frac{\frac{1}{4}\,\frac{\beta-\alpha}{2\pi}\pi(b-a)}{\frac{1}{4}\pi} \\ &= \frac{\beta-\alpha}{2\pi}\,(b-a) = \frac{P\big\{U \in [a,b]\big\}}{P\{V \in C\}}\,\frac{P\big\{W \in S_{\alpha,\beta}\big\}}{P\{V \in C\}}, \end{aligned} \tag{3.49}$$

completing the proof of (e) as well as the proof of the proposition. ■

Remark 3.18. If $u_1, u_2, \ldots$ denote the output of a sequence of i.i.d. copies of random variables uniformly distributed on $[0,1]$, then the Marsaglia-Bray variant of the Box-Muller method is given by the following algorithm:

$$\begin{aligned} 1: &\quad \text{generate } u_i \text{ and } u_{i+1} \\ 2: &\quad v_1 := 2u_i - 1; \; v_2 := 2u_{i+1} - 1 \\ 3: &\quad u := v_1^2 + v_2^2 \\ 4: &\quad \text{if } u = 0 \text{ or } u > 1 \\ &\qquad i := i + 2; \text{ goto } 1 \\ 5: &\quad r := \sqrt{\frac{-2\ln u}{u}} \\ 6: &\quad z_k := rv_1; \; z_{k+1} := rv_2 \\ 7: &\quad k := k + 2; \; i := i + 2; \text{ goto } 1 \end{aligned} \tag{3.50}$$

According to Prop. 3.17(c), u is uniformly distributed on $[0,1]$ and $(v_1, v_2)/\sqrt{u}$ is uniformly distributed on S. In consequence, Rem. 3.16 guarantees that the $z_1, z_2, \ldots$ represent the output of i.i.d. copies of $N(0,1)$-distributed random variables.

—

Another alternative for generating univariate normal deviates is given by the algorithm from [Lev92]. The following provides a *C++* implementation, coded as a derivation from the RNG `Ran` of Sec. 2.3.5. Of course, you can also base it on any other (decent) RNG. The following implementation follows [PTVF07, p. 369] (see [Lev92] for an explanation of the algorithm):

```
struct Normaldev : Ran
{
  const double fac(5.42101086242752217E-20);
  double mu,sig;
  Normaldev(double mmu,
            double ssig, Ullong i) : Ran(i), mu(mmu), sig(ssig){}
  double dev()
  {
    double u,v,x,y,q;
    do {
         do
           {
             u = fac*int64();
           } while(0.0 == u);
         v = 1.7156*(fac*int64()-0.5);
         x = u - 0.449871;
         y = abs(v) + 0.386595;
         q = x*x + y*(0.19600*y-0.25472*x);
       } while (q > 0.27597 && (q > 0.27846 || v*v > -4.*log(u)*u*u));
    return mu + sig*v/u;
  }
};
```

3.4.2 Simulation of Multivariate Normals

In the present section, we proceed to consider the simulation of multivariate normals, i.e. of normally distributed random vectors. Let us first recall the definition of multivariate normals in generalization of Def. and Rem. B.35. We begin with some preparations:

Notation 3.19. For each $\alpha \in \mathbb{R}$, it is useful to define the *degenerate* normal distribution as the *Dirac* probability measure with its measure concentrated in α, i.e.

$$\forall_{\alpha\in\mathbb{R}} \quad N(\alpha,0) := \nu_{\alpha,0} := \delta_\alpha : \mathcal{B}^1 \longrightarrow [0,1], \quad \delta_\alpha(A) := \begin{cases} 1 & \text{for } \alpha \in A, \\ 0 & \text{for } \alpha \notin A. \end{cases} \tag{3.51}$$

For $\sigma > 0$, $N(\alpha, \sigma)$ is defined in Def. and Rem. B.35, and we let

$$\mathcal{N}_1 := \left\{ N(\alpha, \sigma) : \alpha \in \mathbb{R},\, \sigma \in \mathbb{R}_0^+ \right\} \tag{3.52}$$

denote the set of all normal distributions on $\mathcal{B}^1$.

Remark 3.20. A probability measure $\nu : \mathcal{B}^1 \longrightarrow [0,1]$ is in $\mathcal{N}_1$ if, and only if, $A \circ \nu \in \mathcal{N}_1$ for each linear map $A : \mathbb{R} \longrightarrow \mathbb{R}$. This property, suitably generalized to higher dimensions, can be used to define multivariate normal distributions:

Definition 3.21. A probability measure $\nu : \mathcal{B}^d \longrightarrow [0,1]$, $d \in \mathbb{N}$, is called a d-dimensional *Gaussian* or *normal distribution* if, and only if, $A \circ \nu \in \mathcal{N}_1$ for each linear map $A : \mathbb{R}^d \longrightarrow \mathbb{R}$. The set of all d-dimensional normal distributions is denoted by $\mathcal{N}_d$. If $(\Omega, \mathcal{A}, P)$ is a probability space and $X : \Omega \longrightarrow \mathbb{R}^d$ a random variable such that $P_X \in \mathcal{N}_d$, then one calls X a *multivariate normal* or a *normally distributed random vector*.

Remark 3.22. A probability measure $\nu : \mathcal{B}^d \longrightarrow [0,1]$, is a normal distribution if, and only if, for each line L through the origin, the orthogonal projection $\pi : \mathbb{R}^d \longrightarrow L$ transforms ν into a (possibly degenerate) normal distribution on L.

Definition and Remark 3.23. If μ is a probability measure measure on $\mathcal{B}^d$, $d \in \mathbb{N}$, then

$$\hat{\mu} : \mathbb{R}^d \longrightarrow \mathbb{C}, \quad \hat{\mu}(x) := \int_{\mathbb{R}^d} e^{i \langle x, y \rangle} \, \mathrm{d}\mu(y) \,, \tag{3.53}$$

is called the *Fourier transform* of μ, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product on $\mathbb{R}^d$. The map $\mu \mapsto \hat{\mu}$ is one-to-one, i.e. μ is uniquely determined by its Fourier transform (see, e.g., [Bau02, Th. 23.4]). If $(\Omega, \mathcal{A}, P)$ is a probability space, and $X : \Omega \longrightarrow \mathbb{R}^d$ is a random vector, then

$$\phi_X : \mathbb{R}^d \longrightarrow \mathbb{C}, \quad \phi_X(x) := \hat{P}_X(x) = E\left(e^{i \langle x, X \rangle}\right), \tag{3.54}$$

is called the *characteristic function* of X.

Theorem 3.24. *A probability measure $\nu : \mathcal{B}^d \longrightarrow [0,1]$ is in $\mathcal{N}_d$, $d \in \mathbb{N}$, if, and only if, there exist a vector $\alpha \in \mathbb{R}^d$ and a symmetric positive semidefinite real $d \times d$ matrix Σ such that its Fourier transform is given by*

$$\hat{\nu} : \mathbb{R}^d \longrightarrow \mathbb{C}, \quad \hat{\nu}(x) = e^{i \langle x, \alpha \rangle - \frac{1}{2} \langle x, \Sigma x \rangle}. \tag{3.55}$$

Moreover, ν is uniquely determined by α and Σ.

Proof. See, e.g., [Bau02, Th. 30.2]). ■

Definition and Remark 3.25. If $d \in \mathbb{N}$, $\alpha \in \mathbb{R}^d$, Σ is a symmetric positive semidefinite real $d \times d$ matrix, and $\nu \in \mathcal{N}_d$ is such that its Fourier transform is given by (3.55), then one writes

$$N(\alpha, \Sigma) := \nu_{\alpha, \Sigma} := \nu. \tag{3.56}$$

One calls $N(0, \mathrm{Id})$ the d-variate *standard* normal distribution. If $(\Omega, \mathcal{A}, P)$ is a probability space and $X : \Omega \longrightarrow \mathbb{R}^d$ a random vector such that $P_X = N(\alpha, \Sigma)$, then one says X is $N(\alpha, \Sigma)$-distributed. For $N(\alpha, \Sigma)$-distributed X, one checks that each component X_i, $i = 1, \ldots, d$, is normally distributed with $E(X_i) = \alpha_i$. So one defines

$$E(X) := \big(E(X_1), \ldots, E(X_d)\big) = \alpha \tag{3.57}$$

and calls it the *expectation vector* of X. Moreover, one verifies that

$$\sigma_{ij} := \mathrm{Cov}(X_i, X_j) \tag{3.58}$$

are precisely the entries of Σ, which justifies calling Σ the *covariance matrix* of X.

Proposition 3.26. *If $d \in \mathbb{N}$, $\alpha \in \mathbb{R}^d$, Σ is a symmetric positive semidefinite real $d \times d$ matrix, then $N(\alpha, \Sigma)$ has a density with respect to λ_d if, and only if, Σ is positive definite, i.e. if, and only if, Σ is invertible. In the latter case, the density is given by*

$$g_{\alpha,\Sigma} : \mathbb{R}^d \longrightarrow \mathbb{R}^+, \quad g_{\alpha,\Sigma}(x) := \frac{1}{\sqrt{(2\pi)^d \det \Sigma}}\, e^{-\frac{1}{2}\langle x-\alpha,\, \Sigma^{-1}(x-\alpha)\rangle}. \tag{3.59}$$

Proof. See, e.g., [Bau02, Th. 30.4]). ■

Lemma 3.27. *Let $(\Omega, \mathcal{A}, P)$ be a probability space. If $d \in \mathbb{N}$, $\alpha \in \mathbb{R}^d$, Σ is a symmetric positive semidefinite real $d \times d$ matrix, the random vector $Z : \Omega \longrightarrow \mathbb{R}^d$ is $N(\alpha, \Sigma)$-distributed, $\beta \in \mathbb{R}^k$, $k \in \mathbb{N}$, and $A : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ is any real $k \times d$ matrix, then*

$$X : \Omega \longrightarrow \mathbb{R}^k, \quad X := \beta + AZ, \tag{3.60}$$

is $N(\beta + A\alpha, A\Sigma A^{\mathrm{t}})$-distributed.

Proof. Exercise. ■

Remark 3.28. In view of Lem. 3.27, to simulate $N(\alpha, \Sigma)$-distributed random vectors X, it suffices to start with a sequence $z_1, z_2, \ldots$ of values representing the output of i.i.d. copies of an $N(0, 1)$-distributed random variable, letting $Z_i := (z_{(i-1)d+1}, \ldots, z_{id})$, $i \in \mathbb{N}$, one obtains a sequences of values representing the output of i.i.d. copies of an $N(0, \mathrm{Id})$-distributed random vector Z, such that $X_i := \alpha + AZ_i$ represent the output of i.i.d. copies of X, provided $\Sigma = AA^{\mathrm{t}}$.

—

According to Rem. 3.28, we have reduced the problem of simulating multivariate normals to the problem of decomposing a symmetric positive semidefinite real matrix Σ in the form $\Sigma = AA^{\mathrm{t}}$. This can actually always be accomplished such that A is lower triangular, which is desirable, as it allows efficient matrix multiplications $X_i = \alpha + AZ_i$.

Definition 3.29. Let Σ be a symmetric positive semidefinite real $d \times d$ matrix, $d \in \mathbb{N}$. A decomposition

$$\Sigma = AA^{\mathrm{t}} \tag{3.61}$$

is called a *Cholesky decomposition* of Σ if, and only if, A is a left or lower triangular matrix. While useful in our context, this definition is slightly nonstandard, as the name Cholesky decomposition is often restricted to the case where Σ is positive definite.

Theorem 3.30. *Let Σ be a symmetric positive semidefinite real $d \times d$ matrix, $d \in \mathbb{N}$. Then there exists a Cholesky decomposition of Σ in the sense of Def.* 3.61. *Moreover, A can be chosen such that all its diagonal entries are nonnegative. If Σ is positive definite, then the diagonal entries of A can be chosen to be all positive and this determines A uniquely.*

Proof. For the positive definite case see, e.g., [Pla10, Th. 4.24]. The general (positive semidefinite) case can be deduced from the positive definite case as follows: If Σ is symmetric positive semidefinite, then each $\Sigma_n := \Sigma + \frac{1}{n}\,\mathrm{Id}$, $n \in \mathbb{N}$, is positive definite:

$$x^{\mathrm{t}}\Sigma_n x = x^{\mathrm{t}}\Sigma x + \frac{x^{\mathrm{t}}x}{n} > 0 \quad \text{for each } 0 \neq x \in \mathbb{R}^d. \tag{3.62}$$

Thus, for each $n \in \mathbb{N}$, there exists a lower triangular matrix A_n with positive diagonal entries such that $\Sigma_n = A_n A_n^{\mathrm{t}}$.

On the other hand, Σ_n converges to Σ with respect to each norm on $\mathbb{R}^{d^2}$ (since all norms on $\mathbb{R}^{d^2}$ are equivalent). In particular, $\lim_{n\to\infty} \|\Sigma_n - \Sigma\|_2 = 0$, where $\|A\|_2$ denotes the spectral norm of the matrix A. Recall $\|A\|_2 = r(A)$ for each symmetric A, where

$$r(A) := \max\big\{|\lambda| : \lambda \in \mathbb{C} \text{ and } \lambda \text{ is eigenvalue of } A\big\} \tag{3.63}$$

is the spectral radius of A (if A is symmetric, then all eigenvalues of A are real). Thus,

$$\|A_n\|_2^2 = r(\Sigma_n) = \|\Sigma_n\|_2, \tag{3.64}$$

such that $\lim_{n\to\infty} \|\Sigma_n - \Sigma\|_2 = 0$ implies that the set $K := \{A_n : n \in \mathbb{N}\}$ is bounded with respect to $\|\cdot\|_2$. Thus, the closure of K in $\mathbb{R}^{d^2}$ is compact, which implies $(A_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(A_{n_k})_{k\in\mathbb{N}}$, converging to some matrix $A \in \mathbb{R}^{d^2}$. As this convergence is with respect to the norm topology on $\mathbb{R}^{d^2}$, each entry of the A_{n_k} must converge (in $\mathbb{R}$) to the respective entry of A. In particular, A is lower triangular with all nonnegative diagonal entries. It only remains to show $AA^{\mathrm{t}} = \Sigma$. However,

$$\Sigma = \lim_{k\to\infty} \Sigma_{n_k} = \lim_{k\to\infty} A_{n_k} A_{n_k}^{\mathrm{t}} = AA^{\mathrm{t}}, \tag{3.65}$$

which establishes the case. ■

Example 3.31. The decomposition

$$\begin{pmatrix} 0 & 0 \\ \sin x & \cos x \end{pmatrix} \begin{pmatrix} 0 & \sin x \\ 0 & \cos x \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{for each } x \in \mathbb{R} \tag{3.66}$$

shows a symmetric positive semidefinite matrix (which is not positive definite) can have uncountably many different Cholesky decompositions.

Theorem 3.32. *Let Σ be a symmetric positive semidefinite real $d \times d$ matrix, $d \in \mathbb{N}$. Define the index set*

$$I := \big\{(i,j) \in \{1,\dots,d\}^2 : j \leq i\big\}. \tag{3.67}$$

Then a matrix $A = (A_{ij})$ providing a Cholesky decomposition $\Sigma = AA^{\mathrm{t}}$ of $\Sigma = (\sigma_{ij})$ is obtained via the following algorithm, defined recursively over I, using the order $(1,1) < (2,1) < \cdots < (d,1) < (2,2) < \cdots < (d,2) < \cdots < (d,d)$ (which corresponds to traversing the lower half of Σ by columns from left to right):

$$A_{11} := \sqrt{\sigma_{11}}. \tag{3.68a}$$

For $(i,j) \in I \setminus \{(1,1)\}$:

$$A_{ij} := \begin{cases} \left(\sigma_{ij} - \sum_{k=1}^{j-1} A_{ik}A_{jk}\right)/A_{jj} & \text{for } i > j \text{ and } A_{jj} \neq 0, \\ 0 & \text{for } i > j \text{ and } A_{jj} = 0, \\ \sqrt{\sigma_{ii} - \sum_{k=1}^{i-1} A_{ik}^2} & \text{for } i = j. \end{cases} \tag{3.68b}$$

Proof. A lower triangular $d \times d$ matrix A provides a Cholesky decomposition of Σ if, and only if,

$$AA^{\mathrm{t}} = \begin{pmatrix} A_{11} & & & \\ A_{21} & A_{22} & & \\ \vdots & \vdots & \ddots & \\ A_{d1} & A_{d2} & \dots & A_{dd} \end{pmatrix} \begin{pmatrix} A_{11} & A_{21} & \dots & A_{d1} \\ & A_{22} & \dots & A_{d2} \\ & & \ddots & \vdots \\ & & & A_{dd} \end{pmatrix} = \Sigma, \tag{3.69}$$

i.e. if, and only if, the $d(d+1)/2$ lower half entries of A constitute a solution to the following (nonlinear) system of $d(d+1)/2$ equations:

$$\sum_{k=1}^{j} A_{ik}A_{jk} = \sigma_{ij}, \quad (i,j) \in I. \tag{3.70a}$$

Using the order on I introduced in the statement of the theorem, (3.70a) takes the form

$$\begin{aligned} A_{11}^2 &= \sigma_{11}, \\ A_{21}A_{11} &= \sigma_{21}, \\ &\vdots \\ A_{d1}A_{11} &= \sigma_{d1}, \\ A_{21}^2 + A_{22}^2 &= \sigma_{22}, \\ A_{31}A_{21} + A_{32}A_{22} &= \sigma_{32}, \\ &\vdots \\ A_{d1}^2 + \cdots + A_{dd}^2 &= \sigma_{dd}. \end{aligned} \tag{3.70b}$$

From Th. 3.30, we know (3.70) must have at least one solution with $A_{jj} \geq 0$ for each $j \in \{1, \dots, d\}$. In particular, $\sigma_{jj} \geq 0$ for each $j \in \{1, \dots, d\}$ (this is also immediate from Σ being positive semidefinite, since $\sigma_{jj} = e_j^{\mathrm{t}} \Sigma e_j$, where e_j denotes the jth standard unit vector of $\mathbb{R}^d$). We need to show that (3.68) yields a solution to (3.70). The proof

is carried out by induction on d. For $d = 1$, we have $\sigma_{11} \geq 0$ and $A_{11} = \sqrt{\sigma_{11}}$, i.e. there is nothing to prove. Now let $d > 1$. If $\sigma_{11} > 0$, then

$$A_{11} = \sqrt{\sigma_{11}}, \quad A_{21} = \sigma_{21}/A_{11}, \quad \ldots, \quad A_{d1} = \sigma_{d1}/A_{11} \tag{3.71a}$$

is the unique solution to the first d equations of (3.70b) satisfying $A_{11} > 0$, and this solution is provided by (3.68). If $\sigma_{11} = 0$, then $\sigma_{21} = \cdots = \sigma_{d1} = 0$: Otherwise, let $s := (\sigma_{21} \ \ldots \ \sigma_{d1})^{\mathrm{t}} \in \mathbb{R}^{d-1} \setminus \{0\}$, $\alpha \in \mathbb{R}$, and note

$$(\alpha,\, s^{\mathrm{t}}) \begin{pmatrix} 0 & s^{\mathrm{t}} \\ s & \Sigma_{d-1} \end{pmatrix} \begin{pmatrix} \alpha \\ s \end{pmatrix} = (\alpha,\, s^{\mathrm{t}}) \begin{pmatrix} s^{\mathrm{t}} s \\ \alpha s + \Sigma_{d-1} s \end{pmatrix} = 2\alpha \|s\|^2 + s^{\mathrm{t}} \Sigma_{d-1} s < 0$$

for $\alpha < -s^{\mathrm{t}} \Sigma_{d-1} s / (2\|s\|^2)$, in contradiction to Σ being positive semidefinite. Thus,

$$A_{11} = A_{21} = \cdots = A_{d1} = 0 \tag{3.71b}$$

is a particular solution to the first d equations of (3.70b), and this solution is provided by (3.68). We will now denote the solution to (3.70) given by Th. 3.30 by $B_{11}, \ldots, B_{dd}$ to distinguish it from the A_{ij} constructed via (3.68).

In each case, $A_{11}, \ldots, A_{d1}$ are given by (3.71), and, for $(i, j) \in I$ with $i, j \geq 2$, we define

$$\tau_{ij} := \sigma_{ij} - A_{i1} A_{j1} \quad \text{for each } (i, j) \in J := \big\{(i, j) \in I : \, i, j \geq 2\big\}. \tag{3.72}$$

To be able to proceed by induction, we show that the symmetric $(d-1) \times (d-1)$ matrix

$$T := \begin{pmatrix} \tau_{22} & \ldots & \tau_{d2} \\ \vdots & \ddots & \vdots \\ \tau_{d2} & \ldots & \tau_{dd} \end{pmatrix} \tag{3.73}$$

is positive semidefinite. If $\sigma_{11} = 0$, then (3.71b) implies $\tau_{ij} = \sigma_{ij}$ for each $(i, j) \in J$ and T is positive semidefinite, as Σ being positive semidefinite implies

$$\begin{pmatrix} x_2 & \ldots & x_d \end{pmatrix} T \begin{pmatrix} x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 0 & x_2 & \ldots & x_d \end{pmatrix} \Sigma \begin{pmatrix} 0 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \geq 0 \tag{3.74}$$

for each $(x_2, \ldots, x_d) \in \mathbb{R}^{d-1}$. If $\sigma_{11} > 0$, then (3.71a) holds as well as $B_{11} = A_{11}$, $\ldots$, $B_{d1} = A_{d1}$. Thus, (3.72) implies

$$\tau_{ij} := \sigma_{ij} - A_{i1} A_{j1} = \sigma_{ij} - B_{i1} B_{j1} \quad \text{for each } (i, j) \in J. \tag{3.75}$$

Then (3.70) with A replaced by B together with (3.75) implies

$$\sum_{k=2}^{j} B_{ik} B_{jk} = \tau_{ij} \quad \text{for each } (i, j) \in J \tag{3.76}$$

or, written in matrix form,

$$BB^{\mathrm{t}} = \begin{pmatrix} B_{22} & & \\ \vdots & \ddots & \\ B_{d2} & \dots & B_{dd} \end{pmatrix} \begin{pmatrix} B_{22} & \dots & B_{d2} \\ & \ddots & \vdots \\ & & B_{dd} \end{pmatrix} = \begin{pmatrix} \tau_{22} & \dots & \tau_{d2} \\ \vdots & \ddots & \vdots \\ \tau_{d2} & \dots & \tau_{dd} \end{pmatrix} = T, \tag{3.77}$$

which, once again, establishes T to be positive semidefinite (cf. [Koe03, Sec. 6.2.3]).

By induction, we now know the algorithm of (3.68) yields a (possibly different from (3.77) for $\sigma_{11} > 0$) decomposition of T:

$$CC^{\mathrm{t}} = \begin{pmatrix} C_{22} & & \\ \vdots & \ddots & \\ C_{d2} & \dots & C_{dd} \end{pmatrix} \begin{pmatrix} C_{22} & \dots & C_{d2} \\ & \ddots & \vdots \\ & & C_{dd} \end{pmatrix} = \begin{pmatrix} \tau_{22} & \dots & \tau_{d2} \\ \vdots & \ddots & \vdots \\ \tau_{d2} & \dots & \tau_{dd} \end{pmatrix} = T \tag{3.78}$$

or

$$\sum_{k=2}^{j} C_{ik}C_{jk} = \tau_{ij} = \sigma_{ij} - A_{i1}A_{j1} \quad \text{for each } (i,j) \in J, \tag{3.79}$$

where

$$C_{22} := \sqrt{\tau_{22}} = \begin{cases} \sqrt{\sigma_{22}} & \text{for } \sigma_{11} = 0, \\ B_{22} & \text{for } \sigma_{11} > 0, \end{cases} \tag{3.80a}$$

and, for $(i,j) \in J \setminus \{(2,2)\}$,

$$C_{ij} := \begin{cases} \left(\tau_{ij} - \sum_{k=2}^{j-1} C_{ik}C_{jk}\right)/C_{jj} & \text{for } i > j \text{ and } C_{jj} \neq 0, \\ 0 & \text{for } i > j \text{ and } C_{jj} = 0, \\ \sqrt{\tau_{ii} - \sum_{k=2}^{i-1} C_{ik}^2} & \text{for } i = j. \end{cases} \tag{3.80b}$$

Substituting $\tau_{ij} = \sigma_{ij} - A_{i1}A_{j1}$ from (3.72) into (3.80) and comparing with (3.68), an induction over J with respect to the order introduced in the statement of the theorem shows $A_{ij} = C_{ij}$ for each $(i,j) \in J$. In particular, since all C_{ij} are well-defined by induction, all A_{ij} are well-defined by (3.68) (i.e. all occurring square roots exist as real numbers). It also follows that $\{A_{ij} : (i,j) \in I\}$ is a solution to (3.70): The first d equations are satisfied according to (3.71); the remaining $(d-1)\,d/2$ equations are satisfied according to (3.79) combined with $C_{ij} = A_{ij}$. This concludes the proof that (3.68) furnishes a solution to (3.70). ■

Remark 3.33. Even though, in exact arithmetic, the algorithm (3.68) does work for every symmetric positive semidefinite real $d \times d$ matrix Σ, it is sensitive to round-off errors: If Σ is not positive definite, then some A_{jj} must be 0. However, due to round-off errors, it might be very small, but nonzero, resulting in division by very small numbers for the subsequent A_{ij}. Due to this circumstance, if it is known a priori that Σ is not positive definite, one might want to avoid applying (3.68) directly to Σ. If Σ is not positive definite, then it has rank $k < d$ and there exists a symmetric positive definite $k \times k$ matrix $\tilde{\Sigma}$ and a $d \times k$ matrix B such that

$$\Sigma = B\tilde{\Sigma}B^{\mathrm{t}} : \tag{3.81}$$

From Linear Algebra, we know there exist $d \times d$ matrices Σ_1 and B_1, B_1 invertible, such that $\Sigma = B_1 \Sigma_1 B_1^{\mathrm{t}}$, where one can even obtain $\Sigma_1 = \begin{pmatrix} \mathrm{Id}_k & 0 \\ 0 & 0 \end{pmatrix}$ (see [Koe03, Secs. 3.5.6, 6.2.3]), and one can use the first k columns of B_1 for B. Of course, one should not expect B to be numerically easier to obtain than a Cholesky decomposition of Σ. However, if a decomposition of the form (3.81) is somehow known a priori (not necessarily with $\tilde{\Sigma} = \mathrm{Id}_k$, just with $\tilde{\Sigma}$ positive definite), then it might make sense to apply (3.68) to $\tilde{\Sigma}$, yielding a Cholesky decomposition $\tilde{\Sigma} = AA^{\mathrm{t}}$. Then, using Lem. 3.27, one first simulates an $N(0, \mathrm{Id})$-distributed Z, letting $X := \alpha + BAZ$. Since $\Sigma = BAA^{\mathrm{t}}B^{\mathrm{t}}$, Lem. 3.27 shows X is $N(\alpha, \Sigma)$-distributed. However, note than one now also pays an additional matrix multiplication in $X := \alpha + BAZ$.

4 Simulating Stochastic Processes

As already seen in the motivating Sections 1.2 and 1.3, quantities relevant to Mathematical Finance such as the stock price often come in the form of stochastic processes. It is, thus, important to have tools to simulate the paths of such processes. We begin by recalling the precise definitions of stochastic process and path.

Definition 4.1. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\Omega', \mathcal{A}')$ a measurable space, and I and index set. A *stochastic process* is simply any family $(X_t)_{t \in I}$ of random variables $X_t : \Omega \longrightarrow \Omega'$ (we will mostly be interested in the cases $I = \mathbb{R}_0^+$ and $I = [0, T]$, $T > 0$, interpreting t as time; for $I = \mathbb{R}_0^+$, we usually write $(X_t)_{t \geq 0}$ instead of $(X_t)_{t \in \mathbb{R}_0^+}$). For each $\omega \in \Omega$, the function $P_\omega : I \longrightarrow \Omega'$, $P_\omega(t) := X_t(\omega)$ is called a *path* of the process.

—

Given a stochastic process $(X_t)_{t \in I}$, we are now interested in simulating sample paths $t \mapsto X_t(\omega)$. Usually, for $I \subseteq \mathbb{R}$, this means specifying $t_0 < t_1 < \cdots < t_k$ and computing a finite sequence $(x_0, \ldots, x_k) \in (\Omega')^{k+1}$, representing $(X_{t_0}(\omega), \ldots, X_{t_k}(\omega))$.

Definition 4.2. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\Omega', \mathcal{A}')$ a measurable space, and $(X_t)_{t \in I}$ a corresponding stochastic process. Given a finite $\emptyset \neq J \subseteq I$, consider the product map

$$X_J := \bigotimes_{t \in J} X_t : \Omega \longrightarrow (\Omega')^J, \quad X_J(\omega) := \big(X_t(\omega)\big)_{t \in J}. \tag{4.1}$$

(a) Then the *joint distribution* P_J of the $(X_t)_{t \in J}$ is defined as the push-forward measure $P_J := P_{X_J}$ on $(\Omega')^J$.

(b) A method for simulating X_J with the correct distribution P_J is called *exact* on J with respect to $(X_t)_{t \in I}$.

4.1 Brownian Motion

4.1.1 Definition and Basic Properties

Some of the most important stochastic processes in the context of mathematical finance applications are those referred to as Brownian motions, and we already encountered the standard Brownian motion in Sections 1.2 and 1.3. Here is the definition:

Definition 4.3. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $d \in \mathbb{N}$. A stochastic process $(W_t)_{t\geq 0}$, $W_t : \Omega \longrightarrow \mathbb{R}^d$ is called a d-dimensional *Brownian motion* (sometimes also referred to as a *Wiener process*, which justifies the common notation W_t) if, and only if, the following conditions (i)–(iii) hold

(i) For almost every $\omega \in \Omega$, the path $t \mapsto W_t(\omega)$ is continuous.

(ii) Each family of increments $\big(W_{t_0}, W_{t_1} - W_{t_0}, \ldots, W_{t_k} - W_{t_{k-1}}\big)$ with $0 \leq t_0 < t_1 < \cdots < t_k$ is independent.

(iii) For each $0 \leq s < t$, the increment $W_t - W_s$ is $N\big(0, (t-s)\,\mathrm{Id}\,\big)$-distributed.

A Brownian motion is a *standard* Brownian motion if, and only if,

$$W_0 = 0 \quad \text{almost surely.} \tag{4.2}$$

If $\alpha \in \mathbb{R}^d$ and Σ is a symmetric positive semidefinite real $d \times d$ matrix, then the stochastic process $(W_t)_{t\geq 0}$ is said to be a Brownian motion with *drift* α and *covariance matrix* Σ if, and only if, it satisfies (i)-(iii) above with (iii) replaced by

(iii)′ For each $0 \leq s < t$, the increment $W_t - W_s$ is $N\big((t-s)\alpha, (t-s)\Sigma\big)$-distributed.

A Brownian motion on $[0, T]$, $T > 0$, is a Brownian motion on $\mathbb{R}_0^+$, restricted to $[0, T]$.

Lemma 4.4. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $d \in \mathbb{N}$. If $(W_t)_{t\geq 0}$ is a d-dimensional Brownian motion satisfying Def.* 4.3(i)-(iii)*, $\alpha \in \mathbb{R}^d$, Σ is a symmetric positive semidefinite real $d \times d$ matrix, and $\Sigma = AA^{\mathrm{t}}$, then $(X_t)_{t\geq 0}$ with*

$$\underset{t\in\mathbb{R}_0^+}{\forall} \quad X_t := \alpha t + AW_t \tag{4.3}$$

is a d-dimensional Brownian motion with drift α and covariance matrix Σ.

Proof. If $\omega \in \Omega$ and the path $t \mapsto W_t(\omega)$ is continuous, then $t \mapsto X_t(\omega) = \alpha t + AW_t(\omega)$ is also continuous, showing Def. 4.3(i) holds. Given $0 \leq t_0 < t_1 < \cdots < t_k$, the independence of $\big(X_{t_0}, X_{t_1} - X_{t_0}, \ldots, X_{t_k} - X_{t_{k-1}}\big)$ follows from the independence of $\big(W_{t_0}, W_{t_1} - W_{t_0}, \ldots, W_{t_k} - W_{t_{k-1}}\big)$ via Th. B.10, i.e. Def. 4.3(ii) holds. Finally, the validity of Def. 4.3(iii)′ is an immediate consequence of Lem. 3.27. ■

Remark 4.5. The stochastic process given by (4.3) is a solution to the stochastic differential equation (SDE)

$$\mathrm{d}X_t = \alpha\,\mathrm{d}t\, + A\,\mathrm{d}W_t\,. \tag{4.4}$$

This allows to generalize the notion of Brownian motion to the case of a time-dependent drift $\alpha(t)$ and a time-dependent covariance matrix $\Sigma(t)$: A Brownian motion with drift $\alpha(t)$ and covariance matrix $\Sigma(t) = A(t)A(t)^{\mathrm{t}}$ is a solution to the SDE

$$\mathrm{d}X_t = \alpha(t)\,\mathrm{d}t\, + A(t)\,\mathrm{d}W_t. \tag{4.5}$$

One can show that such solutions have almost surely continuous paths, independent increments in the sense of Def. 4.3(ii) and, for each $0 \le s < t$, the increment $X_t - X_s$ is $N\big(\int_s^t \alpha(u)\,\mathrm{d}u\,, \int_s^t \Sigma(u)\,\mathrm{d}u\,\big)$-distributed.

Remark 4.6. If $(X_t)_{t\ge 0}$ is a d-dimensional standard Brownian motion ($X_0 = 0$ a.s.) with drift α and covariance matrix Σ, then combining $X_0 = 0$ with Def. 4.3(iii)$'$ implies X_t is $N\big(t\alpha, t\Sigma\big)$-distributed for each $t \ge 0$.

Remark 4.7. Let $(X_t)_{t\ge 0}$ be a 1-dimensional standard Brownian motion with drift $\alpha \in \mathbb{R}$ and variance σ^2. Given $0 \le t_0 < t_1 < \cdots < t_k$, we are interested in the joint distribution of $(X_{t_i})_{i=0}^k$. Since $(X_{t_0}, \ldots, X_{t_k})$ is a linear transformation of $\big(X_{t_0}, X_{t_1} - X_{t_0}, \ldots, X_{t_k} - X_{t_{k-1}}\big)$,

$$\begin{pmatrix} X_{t_0} \\ X_{t_1} \\ \vdots \\ X_{t_k} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & \ldots & & 1 \end{pmatrix} \begin{pmatrix} X_{t_0} \\ X_{t_1} - X_{t_0} \\ \vdots \\ X_{t_k} - X_{t_{k-1}} \end{pmatrix},$$

we already know the joint distribution is multivariate normal by Def. 4.3(iii)$'$ and Lem. 3.27. The expectation vector is $\alpha(t_0, \ldots, t_k)$. Moreover, if $0 \le s < t$, then

$$\begin{aligned} \mathrm{Cov}(X_s, X_t) &= E(X_s X_t) - E(X_s)E(X_t) = E\big((X_t - X_s)X_s + X_s^2\big) - \alpha^2 st \\ &= E\big((X_t - X_s)X_s\big) - \alpha^2 s(t-s) + E(X_s^2) - \alpha^2 s^2 \\ &= \mathrm{Cov}(X_s, X_t - X_s) + \mathrm{Cov}(X_s, X_s) \\ &= \mathrm{Cov}(X_s, X_s) = \sigma^2 s = \sigma^2 \min\{s, t\}. \end{aligned} \tag{4.6}$$

Thus, the $(k+1) \times (k+1)$ covariance matrix of the joint distribution is

$$\Sigma = (\sigma_{ij})_{(i,j)\in\{1,\ldots,k+1\}^2}, \quad \sigma_{ij} = \sigma^2\, \min\{t_{i-1}, t_{j-1}\}. \tag{4.7}$$

Its Cholesky decomposition is $AA^{\mathrm{t}} = \Sigma$ with

$$A = (A_{ij}) = \sigma \begin{pmatrix} \sqrt{t_0} & 0 & \ldots & 0 \\ \sqrt{t_0} & \sqrt{t_1 - t_0} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{t_0} & \sqrt{t_1 - t_0} & \ldots & \sqrt{t_k - t_{k-1}} \end{pmatrix}, \tag{4.8}$$

i.e.

$$A_{ij} = \begin{cases} \sigma\sqrt{t_0} & \text{for } j = 1, \\ \sigma\sqrt{t_{j-1} - t_{j-2}} & \text{for } i \geq j > 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

Indeed, for each $i \geq j \geq 1$:

$$\sum_{\mu=1}^{j} A_{i\mu} A_{j\mu} = \begin{cases} \sigma^2 t_0 & \text{for } j = 1, \\ \sigma^2 t_0 + \sigma^2 \sum_{\mu=2}^{j} (t_{\mu-1} - t_{\mu-2}) = \sigma^2 t_{j-1} & \text{for } j > 1. \end{cases} \tag{4.10}$$

4.1.2 1-Dimensional Brownian Motion via Random Walk

The first goal in the present section is to simulate a 1-dimensional Brownian motion $(W_t)_{t\geq 0}$ with initial value $W_0 \equiv w_0 \in \mathbb{R}$. Given $0 = t_0 < t_1 < \cdots < t_k$, the idea of the *random walk* construction is to start at w_0 and to randomly walk from w_0 to some w_1, representing the value of W_{t_1}, and so on, until arriving at some w_k, representing the value of W_{t_k}. This random walk should be constructed such that the resulting method is exact on $\{t_0, \ldots, t_k\}$ in the sense of Def. 4.2(b). While simulating a complicated process exactly may be difficult or impossible, Brownian motions are still sufficiently simple, such that an exact simulation is easily achieved, using the properties of the increments:

Start with a sequence $z_1, z_2, \ldots$ of values representing the output of i.i.d. copies of an $N(0,1)$-distributed random variable (see Sec. 3.4.1 above for methods to generate the z_i). Define $(w_0, \ldots, w_k)$ via the following recursion:

$$w_0 : \quad \text{given initial value,} \tag{4.11a}$$

$$w_i := w_{i-1} + z_i \sqrt{t_i - t_{i-1}} \quad \text{for } i = 1, \ldots, k. \tag{4.11b}$$

The same construction actually still works to simulate a Brownian motion $(X_t)_{t\geq 0}$ with drift α and variance σ^2. The recursion then becomes

$$x_0 : \quad \text{given initial value,} \tag{4.12a}$$

$$x_i := x_{i-1} + \alpha(t_i - t_{i-1}) + z_i \, \sigma \sqrt{t_i - t_{i-1}} \quad \text{for } i = 1, \ldots, k, \tag{4.12b}$$

and for time-dependent α and σ:

$$x_0 : \quad \text{given initial value,} \tag{4.13a}$$

$$x_i := x_{i-1} + \int_{t_{i-1}}^{t_i} \alpha(u)\,\mathrm{d}u + z_i \sqrt{\int_{t_{i-1}}^{t_i} \sigma^2(u)\,\mathrm{d}u} \quad \text{for } i = 1, \ldots, k. \tag{4.13b}$$

Remark 4.8. If $Z_1, Z_2, \ldots$ are i.i.d., $N(0,1)$-distributed random variables, then, letting

$$X_0 :\equiv x_0, \tag{4.14a}$$

$$X_i := X_{i-1} + \int_{t_{i-1}}^{t_i} \alpha(u)\,\mathrm{d}u + Z_i \sqrt{\int_{t_{i-1}}^{t_i} \sigma^2(u)\,\mathrm{d}u} \quad \text{for } i = 1, \ldots, k, \tag{4.14b}$$

the $X_i - X_{i-1}$ are $N\big(\int_{t_{i-1}}^{t_i} \alpha(u)\,\mathrm{d}u\,, \int_{t_{i-1}}^{t_i} \sigma^2(u)\,\mathrm{d}u\big)$-distributed and independent by Th. B.10, verifying the exactness on $\{t_0, \dots, t_k\}$ of the random walk construction.

—

Depending on α and σ, the integrals in (4.13b) might not be easily computable. So, in practise, one might want to replace them via quadrature formulas, in the simplest case by approximating α and σ as constantly equal to $\alpha(t_{i-1})$ and $\sigma(t_{i-1})$ on $[t_{i-1}, t_i]$, respectively. This leads to replacing (4.13b) with

$$x_i := x_{i-1} + \alpha(t_{i-1})(t_{i-1} - t_i) + z_i\,\sigma(t_{i-1})\sqrt{t_{i-1} - t_i} \quad \text{for } i = 1, \dots, k. \tag{4.15}$$

However, the quadrature now introduces a so-called discretization error and, in general, the simulation is no longer exact (i.e. the joint distribution of $(X_{t_0}, \dots, X_{t_k})$ is no longer simulated exactly).

4.1.3 1-Dimensional Brownian Motion via Multivariate Normals

One can also base simulations of 1-dimensional Brownian motions on Rem. 4.7. Not surprisingly, we will essentially recover the random walk construction of the previous section. However, it is still instructive to see how this works. As before, let $0 = t_0 < t_1 < \cdots < t_k$ be given.

For simplicity, as in Rem. 4.7, let us consider a standard Brownian motion with constant α, σ. From Rem. 4.7, we know the joint distribution of $(X_{t_1}, \dots, X_{t_k})$ is $N(\beta, \Sigma)$ with expectation $\beta = \alpha(t_1, \dots, t_k)$ and covariance matrix Σ and Cholesky factor A given by (4.7) and (4.8), respectively (now starting at t_1 instead of t_0).

Thus, with $z_1, z_2, \dots$ as in Sec. 4.1.2, $(z_1, \dots, z_k)$ simulates an $N(0, \mathrm{Id})$-distributed random vector, and, according to Rem. 3.28,

$$\begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} := \alpha \begin{pmatrix} t_1 \\ \vdots \\ t_k \end{pmatrix} + \sigma \begin{pmatrix} \sqrt{t_1} & 0 & \dots & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \dots & \sqrt{t_k - t_{k-1}} \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_k \end{pmatrix} \tag{4.16}$$

simulates an $N(\beta, \Sigma)$-distributed random vector, i.e. it exactly simulates the desired joint distribution. Recalling (4.12b), namely

$$x_i := x_{i-1} + \alpha(t_i - t_{i-1}) + z_i\,\sigma\sqrt{t_i - t_{i-1}}, \quad \text{for } i = 1, \dots, k, \tag{4.17}$$

we see (4.16) and (4.12b) are equivalent. However, it is actually not advisable to compute $(x_1, \dots, x_k)$ via the above matrix multiplication – the recursion is much more efficient.

4.1.4 1-Dimensional Brownian Motion via Brownian Bridge

In the present section, we consider a *standard* Brownian motion $(W_t)_{t \geq 0}$ (i.e. $W_0 \equiv 0$), in general, with drift $\alpha \in \mathbb{R}$ and variance $\sigma^2 > 0$ (except for the formulation of the

Markov property in Th. 4.10, which can be stated, whithout additional difficulty, for d-dimensional Brownian motions, both standard and nonstandard).

We still consider $0 = t_0 < t_1 < \cdots < t_k$. The random walk construction of Sec. 4.1.2 obtains the $w_1, \ldots, w_k$, simulating $(W_{t_1}, \ldots, W_{t_k})$, from left to right, i.e. from 1 through k in increasing order. However, it is actually possible to obtain the w_i using an arbitrary order on $\{1, \ldots, k\}$, and it is sometimes desirable to have this flexibility at hand.

Simulating the random variable W_u of a Brownian motion conditional on the distributions of W_s and W_t with $s < u < t$ is referred to as a *Brownian bridge*. This kind of sampling is obviously required to obtain $w_1, \ldots, w_k$ for a nonincreasing order on $\{1, \ldots, k\}$, which is therefore called a Brownian bridge construction. The conditional sampling is feasible for Brownian motions due to the following theorem regarding conditional normal distributions:

Theorem 4.9 (Conditioning Formula for Multivariate Normals)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $d, k \in \mathbb{N}$, $k < d$. Suppose the random vector $Z : \Omega \longrightarrow \mathbb{R}^d$ is $N(\alpha, \Sigma)$-distributed, $\alpha \in \mathbb{R}^d$, Σ a symmetric positive semidefinite real $d \times d$ matrix. Partition $Z = (Z_{[1]}, Z_{[2]})$ into an $\mathbb{R}^k$-valued map $Z_{[1]} := (Z_1, \ldots, Z_k)$ and an $\mathbb{R}^{d-k}$-valued map $Z_{[2]} := (Z_{k+1}, \ldots, Z_d)$. In the same way, also partition*

$$\alpha = \begin{pmatrix} \alpha_{[1]} \\ \alpha_{[2]} \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{[11]} & \Sigma_{[12]} \\ \Sigma_{[21]} & \Sigma_{[22]} \end{pmatrix}, \tag{4.18}$$

i.e., in particular, $\alpha_{[1]} \in \mathbb{R}^k$, $\alpha_{[2]} \in \mathbb{R}^{d-k}$, $\Sigma_{[11]}$ is a $k \times k$ matrix, $\Sigma_{[12]}$ is a $k \times (d-k)$ matrix, $\Sigma_{[21]}$ is a $(d-k) \times k$ matrix, and $\Sigma_{[22]}$ is a $(d-k) \times (d-k)$ matrix.

Then, for each $x \in \mathbb{R}^{d-k}$, the distribution of $Z_{[1]}$ under the condition $\{Z_{[2]} = x\}$ is

$$N\left(\alpha_{[1]} + \Sigma_{[12]}\Sigma_{[22]}^{-1}(x - \alpha_{[2]}),\, \Sigma_{[11]} - \Sigma_{[12]}\Sigma_{[22]}^{-1}\Sigma_{[21]}\right), \tag{4.19}$$

where, in general, $\Sigma_{[22]}^{-1}$ denotes the generalized inverse of $\Sigma_{[22]}$ as defined in [Eat83, p. 87] *(it coincides with the inverse for invertible $\Sigma_{[22]}$).*

Proof. See [Eat83, Prop. 3.13]. ■

And we need one more property of Brownian motions, namely the so-called *Markov property* (it will not be formulated in its most general form, but merely in the form needed here):

Theorem 4.10 (Markov Property of Brownian Motions)**.** *Let $\alpha \in \mathbb{R}^d$, $d \in \mathbb{N}$, Σ a symmetric positive semidefinite real $d \times d$ matrix. Let $(X_t)_{t \geq 0}$ be a d-dimensional Brownian motion with drift α and covariance matrix Σ, and $0 \leq t_1 < \cdots < t_k$, $k \in \mathbb{N}$, $k \geq 2$, $s \in [t_{i-1}, t_i]$ for some $i \in \{2, \ldots, k\}$. Then, for each $(x_1, \ldots, x_k) \in (\mathbb{R}^d)^k$, the following identity of conditional distributions holds:*

$$\big(X_s | X_{t_1} = x_1, \ldots, X_{t_k} = x_k\big) = \big(X_s | X_{t_{i-1}} = x_{i-1},\, X_{t_i} = x_i\big). \tag{4.20}$$

In other words, conditioning X_s on all times $t_1, \ldots, t_k$ is the same as conditioning X_s merely on the two times immediately before and after s.

Proof. For example, the statement follows from combining [Bau02, Th. 40.6(3)] with [Bau02, Cor. 42.4]. ■

In the nth step of a Brownian bridge construction, we need to obtain w_j, representing W_{t_j}. If t_j is bigger than all t_i previously dealt with, then w_j is obtained from the w_i belonging to the largest previous t_i via the random walk construction of Sec. 4.1.2. If, on the other hand, t_j falls between previously considered t_i, then (4.20) says that we only need to determine $\big(W_{t_j} | W_{t_M} = w_M,\, W_{t_N} = w_N\big)$, where M is the largest index such that $t_M < t_j$ and w_M has already been constructed, whereas N is the smallest index such that $t_j < t_N$ and w_N has already been constructed. Letting $u := t_M < s := t_j < t := t_N$, we use the following proposition, which is derived from Th. 4.9.

Proposition 4.11. *If $(W_t)_{t\geq 0}$ is a 1-dimensional standard Brownian motion with drift $\alpha \in \mathbb{R}$ and variance $\sigma^2 > 0$, then, for each $0 \leq u < s < t$ and $x, y \in \mathbb{R}$, the conditional distribution $\big(W_s | W_u = x,\, W_t = y\big)$ is independent of the drift α and given by*

$$N\left(\frac{(t-s)x + (s-u)y}{t-u},\, \frac{\sigma^2 (s-u)(t-s)}{t-u}\right). \tag{4.21}$$

Proof. According to Rem. 4.7, the (unconditional) distribution of (W_u, W_s, W_t) is

$$N\left(\alpha \begin{pmatrix} u \\ s \\ t \end{pmatrix},\, \sigma^2 \begin{pmatrix} u & u & u \\ u & s & s \\ u & s & t \end{pmatrix}\right). \tag{4.22}$$

To apply Th. 4.9, we permute the order of the entries: The distribution of (W_s, W_u, W_t) is

$$N\left(\alpha \begin{pmatrix} s \\ u \\ t \end{pmatrix},\, \sigma^2 \begin{pmatrix} s & u & s \\ u & u & u \\ s & u & t \end{pmatrix}\right). \tag{4.23}$$

In terms of the notation from (4.18), we have

$$\alpha_{[1]} = \alpha\, s, \quad \alpha_{[2]} = \begin{pmatrix} \alpha\, u \\ \alpha\, t \end{pmatrix}, \quad \Sigma_{[11]} = \sigma^2\, s, \quad \Sigma_{[12]} = \sigma^2\, (u, s), \tag{4.24a}$$

$$\Sigma_{[21]} = \sigma^2 \begin{pmatrix} u \\ s \end{pmatrix}, \quad \Sigma_{[22]} = \sigma^2 \begin{pmatrix} u & u \\ u & t \end{pmatrix}. \tag{4.24b}$$

For $u > 0$, $\Sigma_{[22]}$ is invertible, where

$$\Sigma_{[22]}^{-1} = \frac{1}{\sigma^2\,(ut - u^2)} \begin{pmatrix} t & -u \\ -u & u \end{pmatrix} = \frac{1}{\sigma^2\,(t-u)} \begin{pmatrix} t/u & -1 \\ -1 & 1 \end{pmatrix}. \tag{4.24c}$$

For $u = 0$, $\Sigma_{[22]} = \sigma^2 \begin{pmatrix} 0 & 0 \\ 0 & t \end{pmatrix}$ is not invertible and the generalized inverse according to [Eat83, p. 87] is

$$\Sigma_{[22]}^{-1} = \frac{1}{\sigma^2\, t} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.24d}$$

Thus, according to Th. 4.9, the expectation value of the conditional distribution is, for $u > 0$,

$$\alpha_{[1]} + \Sigma_{[12]}\Sigma_{[22]}^{-1}\begin{pmatrix} x - \alpha\, u \\ y - \alpha\, t \end{pmatrix} = \alpha s + \frac{(u,s)}{t-u}\begin{pmatrix} tx/u - t\alpha - y + \alpha t \\ -x + \alpha u + y - \alpha t \end{pmatrix}$$
$$= \alpha s + \frac{tx - uy - sx + \alpha su + sy - \alpha st}{t-u} = \frac{(t-s)x + (s-u)y}{t-u}, \tag{4.25a}$$

as claimed in (4.21); and, for $u = 0$, taking into account $x = 0$ for our standard Brownian motion,

$$\alpha_{[1]} + \Sigma_{[12]}\Sigma_{[22]}^{-1}\begin{pmatrix} 0 \\ y - \alpha\, t \end{pmatrix} = \alpha s + \frac{(0,s)}{t}\begin{pmatrix} 0 \\ y - \alpha t \end{pmatrix} = \frac{\alpha st + sy - \alpha st}{t} = \frac{sy}{t}, \tag{4.25b}$$

again as claimed in (4.21). According to Th. 4.9, the variance of the conditional distribution, divided by σ^2, is, for $u > 0$,

$$\sigma^{-2}\left(\Sigma_{[11]} - \Sigma_{[12]}\Sigma_{[22]}^{-1}\Sigma_{[21]}\right) = s - \frac{(u,s)}{t-u}\begin{pmatrix} t-s \\ -u+s \end{pmatrix} = \frac{st - su - ut + su + su - s^2}{t-u}$$
$$= \frac{(s-u)(t-s)}{t-u}, \tag{4.26a}$$

and for $u = 0$,

$$\sigma^{-2}\left(\Sigma_{[11]} - \Sigma_{[12]}\Sigma_{[22]}^{-1}\Sigma_{[21]}\right) = s - \frac{(0,s)}{t}\begin{pmatrix} 0 \\ s \end{pmatrix} = \frac{s(t-s)}{t}, \tag{4.26b}$$

completing the proof. ■

4.1.5 Simulating d-Dimensional Brownian Motions

The techniques of random walk and Brownian bridge, used to simulate 1-dimensional Brownian motions, can also be applied to simulate d-dimensional Brownian motions. If the d-dimensional Brownian motion has covariance matrix Σ, then a decomposition $\Sigma = AA^{\mathrm{t}}$ is needed (see Th. 3.32 and Rem. 3.33 regarding how to obtain such a decomposition).

In the most simple case, i.e. $\alpha = 0$ (no drift) and $\Sigma = \mathrm{Id}$, the random walk construction (4.11) translates directly to the d-dimensional case, with z_i and w_i now being vectors in $\mathbb{R}^d$ (each z_i having to be filled with d random numbers, representing i.i.d. $N(0,1)$-distributed random variables). This is equivalent to independently simulating each component of the d-dimensional Brownian motion using the 1-dimensional (4.11).

The generalization of (4.12) to d dimensions is

$$x_0: \quad \text{given initial vector}, \tag{4.27a}$$
$$x_i := x_{i-1} + \alpha(t_i - t_{i-1}) + \sqrt{t_i - t_{i-1}}\, A\, z_i, \quad \text{for } i = 1, \ldots, k, \tag{4.27b}$$

with $z_i, x_i \in \mathbb{R}^d$, which is computationally more expensive, as it involves a matrix multiplication in each step. The computational expenditure becomes even bigger if α or, especially, if Σ depends on t: Here, (4.13) generalizes to

$$x_0 : \quad \text{given initial vector}, \tag{4.28a}$$

$$x_i := x_{i-1} + \int_{t_{i-1}}^{t_i} \alpha(u)\,\mathrm{d}u \; + A(t_{i-1}, t_i)\, z_i, \quad \text{for } i = 1, \ldots, k, \tag{4.28b}$$

$$\text{where} \quad A(t_{i-1}, t_i) A(t_{i-1}, t_i)^{\mathrm{t}} = \int_{t_{i-1}}^{t_i} \Sigma(u)\,\mathrm{d}u\,, \tag{4.28c}$$

i.e., in addition to the matrix multiplication in (4.28b), a matrix factorization according to (4.28c) needs to be done in each step (not to mention performing the integrals).

For a d-dimensional standard Brownian motion $(W_t)_{t\geq 0}$ without drift and $\Sigma = \mathrm{Id}$, the Brownian bridge construction can simply be applied independently to each component of W_t (as was the case for the corresponding random walk construction described above). To obtain a d-dimensional Brownian motion $(X_t)_{t\geq 0}$ with drift vector α and covariance matrix Σ via a Brownian bridge construction, one still applies the construction to $(W_t)_{t\geq 0}$ and obtains X_t by the usual representation $X_t = \alpha\, t + A\, W_t$ with $\Sigma = A\, A^{\mathrm{t}}$.

4.1.6 Simulating Geometric Brownian Motions

Another important type of stochastic process occurring in models of mathematical finance is the so-called geometric Brownian motion.

Definition 4.12. Following [Gla04, Sec. 3.2.1], we define an $\mathbb{R}^+$-valued stochastic process $(S_t)_{t\geq 0}$ to be a 1-dimensional *geometric Brownian motion* with drift $\alpha \in \mathbb{R}$ and variance σ^2, $\sigma > 0$, if, and only if, $(S_t)_{t\geq 0}$ is a solution to the SDE

$$\frac{\mathrm{d}S_t}{S_t} = \alpha\,\mathrm{d}t \; + \sigma\,\mathrm{d}W_t\,, \tag{4.29}$$

where $(W_t)_{t\geq 0}$ is a 1-dimensional standard Brownian motion with $\alpha = 0$ and $\sigma = 1$.

Caveat 4.13. There is a *drift shift* between the drift of a geometric Brownian motion and its corresponding Brownian motion: Let $(S_t)_{t\geq 0}$ denote a 1-dimensional geometric Brownian motion with drift $\alpha \in \mathbb{R}$ and variance σ^2, $\sigma > 0$. Letting

$$\underset{t\in\mathbb{R}_0^+}{\forall} \quad X_t := \ln S_t, \tag{4.30}$$

Itô's formula (C.3) yields

$$\mathrm{d}X_t \; = \left(\frac{\alpha S_t}{S_t} - \frac{\sigma^2}{2\,S_t^2}\,S_t^2\right)\,\mathrm{d}t \; + \frac{\sigma S_t}{S_t}\,\mathrm{d}W_t \; = \left(\alpha - \frac{\sigma^2}{2}\right)\,\mathrm{d}t \; + \sigma\,\mathrm{d}W_t\,, \tag{4.31}$$

showing $(X_t)_{t\geq 0}$ constitutes a 1-dimensional Brownian motion with drift $\alpha - \frac{1}{2}\sigma^2$ and variance σ^2. It is an exercise to show the converse, namely that $(X_t)_{t\geq 0}$ defined by (4.30)

being a 1-dimensional Brownian motion with drift $\alpha - \frac{1}{2}\sigma^2$ and variance σ^2 implies $(S_t)_{t\geq 0}$ to be a 1-dimensional geometric Brownian motion with drift $\alpha \in \mathbb{R}$ and variance σ^2. Using obvious notation, we can summarize the above as

$$(S_t)_{t\geq 0} \text{ is } \mathrm{GBM}(\alpha, \sigma^2) \quad \Leftrightarrow \quad (X_t)_{t\geq 0} \text{ is } \mathrm{BM}(\alpha - \sigma^2/2, \sigma^2). \tag{4.32}$$

Remark 4.14. The relation (4.32) allows to exploit methods for simulating Brownian motions to also simulate geometric Brownian motions: For example, if $(X_t)_{t\geq 0}$ is a 1-dimensional Brownian motion with drift $\alpha - \frac{1}{2}\sigma^2$ and deviation σ, then, for $0 \leq s < t$,

$$X_t - X_s = \left(\alpha - \frac{1}{2}\sigma^2\right)(t-s) + Z\sigma\sqrt{t-s}, \tag{4.33}$$

where the random variable Z is $N(0,1)$-distributed. From Caveat 4.13, we know $S_t := \exp(X_t)$ defines a 1-dimensional geometric Brownian motion $(S_t)_{t\geq 0}$ with drift α and variance σ^2. Exponentiating (4.33) yields

$$S_t = S_s \exp\left(\left(\alpha - \frac{1}{2}\sigma^2\right)(t-s) + Z\sigma\sqrt{t-s}\right). \tag{4.34}$$

Let $0 = t_0 < t_1 < \cdots < t_k$ and let $z_1, z_2, \ldots$ represent the output of i.i.d. copies of an $N(0,1)$-distributed random variable. Together with the independence of the increments, (4.34) implies the random walk construction

$$s_0 > 0: \quad \text{given initial value,} \tag{4.35a}$$

$$s_i := s_{i-1} \exp\left(\left(\alpha - \frac{1}{2}\sigma^2\right)(t_i - t_{i-1}) + z_i\,\sigma\,\sqrt{t_i - t_{i-1}}\right) \quad \text{for } i = 1, \ldots, k, \tag{4.35b}$$

provides an exact simulation on $\{t_0, \ldots, t_k\}$ in the sense of Def. 4.2(b) for the 1-dimensional geometric Brownian motion $(S_t)_{t\geq 0}$. Not surprisingly, (4.35) is precisely the exponentiated form of (4.12), with α replaced by $\alpha - \frac{1}{2}\sigma^2$.

Definition 4.15. Following [Gla04, Sec. 3.2.3], we define an $(\mathbb{R}^+)^d$-valued stochastic process $(S_t)_{t\geq 0}$, $d \in \mathbb{N}$, to be a d-dimensional *geometric Brownian motion* with drift $\alpha \in \mathbb{R}^d$ and symmetric positive semidefinite real $d \times d$ covariance matrix Σ if, and only if, $S_t = ((S_t)_1, \ldots, (S_t)_d)$ and $(S_t)_{t\geq 0}$ is a solution to the system of SDE

$$\underset{i\in\{1,\ldots,d\}}{\forall} \quad \frac{\mathrm{d}(S_t)_i}{(S_t)_i} = \alpha_i\,\mathrm{d}t + \sigma_i\,\mathrm{d}(W_i)_t\,, \tag{4.36}$$

where $\sigma_i > 0$, each $((W_i)_t)_{t\geq 0}$ is a standard Brownian motion with $\tilde{\alpha}_i = 0$, $\tilde{\sigma}_i = 1$, and such that

$$\underset{i,j\in\{1,\ldots,d\}}{\forall} \quad \rho_{ij} := \mathrm{Cor}\left((W_i)_t, (W_j)_t\right) = \frac{\mathrm{Cov}\left((W_i)_t, (W_j)_t\right)}{\sqrt{V\left((W_i)_t\right)}\sqrt{V\left((W_j)_t\right)}} = \frac{\mathrm{Cov}\left((W_i)_t, (W_j)_t\right)}{t}, \tag{4.37}$$

$$\Sigma = \left(\sigma_i\sigma_j\rho_{ij}\right). \tag{4.38}$$

Caveat 4.16. In Def. 4.15, calling α the drift and Σ the covariance matrix of the the d-dimensional geometric Brownian motion $(S_t)_{t\geq 0}$ is not entirely canonical – according to (4.36), one would have $\big(\alpha_1(S_t)_1, \dots, \alpha_d(S_t)_d\big)$ as the actual drift vector of S_t, and it is an exercise to show that, for constant (deterministic) $S_0 \in \mathbb{R}^d$, (4.36) and (4.37) imply the actual covariances are

$$\underset{i,j\in\{1,\dots,d\}}{\forall} \quad \operatorname{Cov}\big((S_t)_i, (S_t)_j\big) = (S_0)_i (S_0)_j \, e^{(\alpha_i+\alpha_j)\,t}(e^{\rho_{ij}\sigma_i\sigma_j} - 1). \tag{4.39}$$

Remark 4.17. Similar to the 1-dimensional case, the relation between geometric Brownian motions and Brownian motions allows to obtain simulation methods for geometric Brownian motions also in the d-dimensional situation. Noting that, in the situation of Def. 4.15, $\big(\sigma_1(W_1)_t, \dots, \sigma_d(W_d)_t\big)$ is a Brownian motion with drift 0 and covariance matrix Σ, (4.36) can be written as

$$\underset{i\in\{1,\dots,d\}}{\forall} \quad \frac{\mathrm{d}(S_t)_i}{(S_t)_i} = \alpha_i \,\mathrm{d}t \, + A_i \,\mathrm{d}W_t \,, \tag{4.40}$$

where $(W_t)_{t\geq 0}$ denotes a d-dimensional standard Brownian motion with drift 0 and trivial covariance matrix Id, and A_i is the ith row of A with $\Sigma = AA^{\mathrm{t}}$. An argument analogous to the one in Rem. 4.14 yields, for $0 \leq s < t$,

$$\begin{aligned} \underset{i\in\{1,\dots,d\}}{\forall} \quad (S_t)_i &= (S_s)_i \exp\left(\left(\alpha_i - \frac{1}{2}\sigma_i^2\right)(t-s) + \sqrt{t-s}\, A_i \, Z\right) \\ &= (S_s)_i \exp\left(\left(\alpha_i - \frac{1}{2}\sigma_i^2\right)(t-s) + \sqrt{t-s}\, \sum_{j=1}^{d} A_{ij} \, Z_j\right), \end{aligned}$$

where the random vector Z is $N(0, \mathrm{Id})$-distributed. Once again, let $0 = t_0 < t_1 < \cdots < t_k$ and let $z_1, z_2, \dots$ represent the output of i.i.d. copies of an $N(0, \mathrm{Id})$-distributed random vector. Together with the independence of the increments, (4.41) implies the random walk construction

$$s_0 \in (\mathbb{R}^+)^d : \quad \text{given initial vector,} \tag{4.41a}$$

$$(s_l)_i := (s_{l-1})_i \exp\left(\left(\alpha_i - \frac{1}{2}\sigma_i^2\right)(t_l - t_{l-1}) + \sqrt{t_l - t_{l-1}} \sum_{j=1}^{d} A_{ij} \, (z_l)_j\right)$$

$$\text{for } i = 1, \dots, d \text{ and } l = 1, \dots, k, \tag{4.41b}$$

provides an exact simulation on $\{t_0, \dots, t_k\}$ for the d-dimensional geometric Brownian motion $(S_t)_{t\geq 0}$. Note that one obtains (4.41) from (4.27) by exponentiating each component after replacing α_i by $\alpha_i - \frac{1}{2}\sigma_i^2$.

4.2 Gaussian Short Rate Models

4.2.1 Short Rates and Bond Pricing

As a slightly more concrete application of simulating stochastic processes to financial mathematics, let us consider certain stochastic processes modeling so-called short rates.

In the motivational Sections 1.2 and 1.3, the interest rate (short rate) r was treated as a deterministic quantity, i.e. it was modeled as an $\mathbb{R}_0^+$-valued function. However, it is usually more realistic to consider r as a stochastic quantity to be modeled by a stochastic process (i.e. by a random variable-valued function, each random variable being, in turn, $\mathbb{R}_0^+$-valued).

If the short rate r_t is instantaneously compounded, then an investment deposit earning interest rate r_u at time u grows from a value of 1 to a value of

$$\beta_t = \exp\left(\int_0^t r_u \, \mathrm{d}u\right) \quad \text{at time } t. \tag{4.42}$$

As before, a time variable occurring as a subscript indicates a stochastic process, i.e. the integral in (4.42) is a random variable-valued integral. Under so-called risk-neutral pricing, if a derivative security pays X at time T, its price at $t = 0$ is the expected value

$$E\left(\frac{X}{\beta_T}\right) = E\left(X \, \exp\left(-\int_0^T r_u \, \mathrm{d}u\right)\right). \tag{4.43}$$

In particular, one is often interested in the resulting price $B(0,T)$ of a bond at $t = 0$ if the bond pays $X = 1$ at $t = T$:

$$B(0,T) := E\left(\exp\left(-\int_0^T r_u \, \mathrm{d}u\right)\right). \tag{4.44}$$

We will now proceed to consider stochastic processes used to model the short rate r. For background and scope of the respective models, please consult the literature on mathematical modeling of finance. Here, we will not be concerned with a deeper discussion of the models and their validity.

Only two types of models will be discussed here, namely continuous-time Ho-Lee and Vasicek models. All these models fall into the class of *Gaussian* models, i.e., in each case, $(r_t)_{t\geq 0}$ constitutes a Gaussian stochastic process (a process such that the joint distribution of $(r_{t_1}, \ldots, r_{t_k})$ is (multivariate) normal for each finite sequence $t_1, \ldots, t_k \geq 0$).

4.2.2 Ho-Lee Models

Definition 4.18. An $\mathbb{R}$-valued stochastic process $(r_t)_{t\geq 0}$ is given by a *Ho-Lee model* if, and only if, the process constitutes a solution to the SDE

$$\mathrm{d}r_t \; = g(t) \, \mathrm{d}t \; + \sigma \, \mathrm{d}W_t \, , \tag{4.45}$$

with $\sigma > 0$, a locally integrable (deterministic) function $g : \mathbb{R}_0^+ \longrightarrow \mathbb{R}$, and $(W_t)_{t\geq 0}$ denoting a 1-dimensional standard Brownian motion with drift 0 and variance 1.

Remark 4.19. Comparing (4.45) with (4.5), one observes that any short rate $(r_t)_{t\geq 0}$ given by a Ho-Lee model is merely a 1-dimensional Brownian motion with time-dependent drift $g(t)$ and variance σ^2:

$$r_t = r_0 + \int_0^t g(s)\,\mathrm{d}s \, + \sigma\, W_t. \tag{4.46}$$

Thus, it can be simulated (exactly) by (4.13), provided the antiderivative of g is available, or by (4.15) (incurring a discretization error), if the antiderivative of g is not at hand.

—

A short rate $(r_t)_{t\geq 0}$ given by a Ho-Lee model is still sufficiently simple, such that the bond price $B(0,T)$ according to (4.44) can be computed in closed form. To compute $B(0,T)$, we need the following result about normally distributed random variables:

Proposition 4.20. *If $(\Omega, \mathcal{A}, P)$ is a probability space and the random variable $X : \Omega \longrightarrow \mathbb{R}$ is $N(\alpha,\sigma^2)$-distributed, $\alpha \in \mathbb{R}$, $\sigma \in \mathbb{R}_0^+$, then*

$$E(e^X) = e^{\alpha + \frac{\sigma^2}{2}}. \tag{4.47}$$

Proof. If $\sigma = 0$, then

$$E(e^X) = \int_{-\infty}^{\infty} e^x \,\mathrm{d}P_X(x) \, = \int_{-\infty}^{\infty} e^x \,\mathrm{d}\delta_\alpha(x) \, = e^\alpha \tag{4.48a}$$

as claimed. If $\sigma > 0$, then

$$\begin{aligned}
&E(e^X)\\
&= \int_{-\infty}^{\infty} e^x \,\mathrm{d}P_X(x) \, = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^x \, e^{-\frac{(x-\alpha)^2}{2\,\sigma^2}} \,\mathrm{d}x\\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - (2\alpha + 2\sigma^2)x + \alpha^2}{2\,\sigma^2}\right) \mathrm{d}x\\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{2\alpha\sigma^2 + \sigma^4}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - (2\alpha + 2\sigma^2)x + \alpha^2 + 2\alpha\sigma^2 + \sigma^4}{2\,\sigma^2}\right) \mathrm{d}x\\
&= e^{\alpha + \frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{\left(x - (\alpha + \sigma^2)^2\right)^2}{2\,\sigma^2}\right) \mathrm{d}x \, = e^{\alpha + \frac{\sigma^2}{2}},
\end{aligned} \tag{4.48b}$$

completing the proof. ■

In addition to Prop. 4.20, the computation of $B(0,T)$ needs the following result, which states that integrals of Brownian motions are normally distributed (one can also show this, more generally, for integrals of Gaussian processes):

Proposition 4.21. *If $(\Omega, \mathcal{A}, P)$ is a probability space and the $\mathbb{R}$-valued stochastic process $(X_t)_{t\geq 0}$, $X_t : \Omega \longrightarrow \mathbb{R}$, is a 1-dimensional Brownian motion with drift $\alpha(t)$ and variance*

$\sigma^2(t)$, *where* $\alpha : \mathbb{R}_0^+ \longrightarrow \mathbb{R}$ *is locally integrable and* $\sigma : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ *is locally square-integrable, then*

$$\underset{T\in\mathbb{R}_0^+}{\forall} \quad I_T : \Omega \longrightarrow \mathbb{R}, \quad I_T(\omega) := \int_0^T X_t(\omega)\,\mathrm{d}t\,, \tag{4.49}$$

is a normally distributed random variable.

Proof. The statement of the proposition is a simple consequence of the integration by parts formula for Itô integrals of Th. C.3, which yields

$$\begin{aligned} I_T &= \int_0^T X_t\,\mathrm{d}t \;= T\,X_T - \int_0^T s\,\alpha(s)\,\mathrm{d}s \;- \int_0^T s\,\sigma(s)\,\mathrm{d}W_s \\ &= T\,X_0 + T\int_0^T \alpha(s)\,\mathrm{d}s \;+ T\int_0^T \sigma(s)\,\mathrm{d}W_s \;- \int_0^T s\,\alpha(s)\,\mathrm{d}s \;- \int_0^T s\,\sigma(s)\,\mathrm{d}W_s \\ &= T\,X_0 + \int_0^T (T-s)\,\alpha(s)\,\mathrm{d}s \;+ \int_0^T (T-s)\,\sigma(s)\,\mathrm{d}W_s\,, \end{aligned}$$

showing $I_T = Y_T$, where $(Y_t)_{t\geq 0}$ is a 1-dimensional Brownian motion with drift $(T-t)\alpha(t)$ and variance $(T-t)^2\sigma^2(t)$. Thus, I_T is normally distributed. ■

Theorem 4.22. *Fix some* $T \geq 0$*. If the short rate* $(r_t)_{t\geq 0}$ *is given by a Ho-Lee model, i.e. one having the form* (4.46)*, and* $r_0 \in \mathbb{R}$*, then the bond price according to* (4.44) *is*

$$B(0,T) = \exp\left(-r_0\,T - \int_0^T\int_0^t g(s)\,\mathrm{d}s\;\mathrm{d}t \;+ \frac{\sigma^2\,T^3}{6}\right). \tag{4.50}$$

Proof. We already know that $(r_t)_{t\geq 0}$ is a Brownian motion given by (4.46). Let $(\Omega, \mathcal{A}, P)$ be a probability space such that $r_t : \Omega \longrightarrow \mathbb{R}$ for each $t \geq 0$. According to Prop. 4.21,

$$I_T : \Omega \longrightarrow \mathbb{R}, \quad I_T(\omega) := \int_0^T r_t(\omega)\,\mathrm{d}t \;= r_0\,T + \int_0^T\int_0^t g(s)\,\mathrm{d}s\;\mathrm{d}t \;+ \sigma\int_0^T W_t(\omega)\,\mathrm{d}t\,, \tag{4.51}$$

constitutes a normally distributed random variable. We proceed to compute its expected value and variance.

The expected value is

$$E(I_T) = E\left(\int_0^T r_t\,\mathrm{d}t\right) = \int_\Omega\int_0^T r_t(\omega)\,\mathrm{d}t\;\mathrm{d}\omega \;= r_0\,T + \int_0^T\int_0^t g(s)\,\mathrm{d}s\;\mathrm{d}t\,, \tag{4.52}$$

since

$$E\left(\int_0^T W_t\,\mathrm{d}t\right) = \int_\Omega\int_0^T W_t(\omega)\,\mathrm{d}t\;\mathrm{d}\omega \overset{\text{Fubini}}{=} \int_0^T\int_\Omega W_t(\omega)\,\mathrm{d}\omega\;\mathrm{d}t \;= 0, \tag{4.53}$$

as $E(W_t) = 0$ for each $t \geq 0$. The Fubini theorem applies, since one can assume the Brownian motion $(\omega, t) \mapsto W_t(\omega)$ to be $\mathcal{A}\otimes\mathbb{B}^1$-measurable.

Next, we turn our attention to the variance. We find

$$V(I_T) = V\left(\int_0^T r_t \,\mathrm{d}t\right) = \sigma^2 V\left(\int_0^T W_t \,\mathrm{d}t\right), \tag{4.54}$$

since the first two summands in the right-hand side of (4.51) do not depend on $\omega \in \Omega$. We further compute

$$\begin{aligned} V\left(\int_0^T W_t \,\mathrm{d}t\right) &\overset{(4.53)}{=} E\left(\left(\int_0^T W_t \,\mathrm{d}t\right)^2\right) = \int_\Omega \int_0^T \int_0^T W_s(\omega) W_t(\omega) \,\mathrm{d}s \;\mathrm{d}t \;\mathrm{d}\omega \\ &= 2\int_\Omega \int_0^T \int_0^t W_s(\omega) W_t(\omega) \,\mathrm{d}s \;\mathrm{d}t \;\mathrm{d}\omega \\ &\overset{\text{Fubini}}{=} 2\int_0^T \int_0^t \int_\Omega W_s(\omega) W_t(\omega) \,\mathrm{d}\omega \;\mathrm{d}s \;\mathrm{d}t \\ &= 2\int_0^T \int_0^t \mathrm{Cov}(W_s, W_t) \,\mathrm{d}s \;\mathrm{d}t \overset{(4.6)}{=} 2\int_0^T \int_0^t s \,\mathrm{d}s \;\mathrm{d}t \\ &= \frac{T^3}{3}. \end{aligned} \tag{4.55}$$

Combining (4.52), (4.54), and (4.55) yields

$$I_T = \int_0^T r_t \,\mathrm{d}t \sim N\left(r_0\, T + \int_0^T \int_0^t g(s) \,\mathrm{d}s \;\mathrm{d}t\,,\, \frac{\sigma^2 T^3}{3}\right), \tag{4.56a}$$

and

$$-I_T = -\int_0^T r_t \,\mathrm{d}t \sim N\left(-r_0\, T - \int_0^T \int_0^t g(s) \,\mathrm{d}s \;\mathrm{d}t\,,\, \frac{\sigma^2 T^3}{3}\right). \tag{4.56b}$$

Finally, an application of (4.47) implies

$$B(0,T) = E(e^{-I_T}) = \exp\left(-r_0\, T - \int_0^T \int_0^t g(s) \,\mathrm{d}s \;\mathrm{d}t + \frac{\sigma^2 T^3}{6}\right), \tag{4.57}$$

completing the proof. ■

4.2.3 Vasicek Models

Definition 4.23. An $\mathbb{R}$-valued stochastic process $(r_t)_{t\geq 0}$ is given by a *Vasicek model* if, and only if, the process constitutes a solution to the SDE

$$\mathrm{d}r_t = \alpha\big(b(t) - r_t\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t\,, \tag{4.58}$$

with $\alpha, \sigma > 0$, a locally integrable (deterministic) function $b : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$, and $(W_t)_{t\geq 0}$ denoting a 1-dimensional standard Brownian motion with drift 0 and variance 1. Solutions to (4.58) are also known as *Ornstein-Uhlenbeck* processes.

Remark 4.24. According to (4.58), the drift of a short rate given by a Vasicek model is positive for $b(t) > r_t$ and negative for $b(t) < r_t$. Thus, especially for constant b, b can be interpreted as a long-term interest rate, that the short rate is pulled toward, α controlling the strength of the pull.

Remark 4.25. According to Itô's formula [Gla04, Th. B.1.1], if $(r_t)_{t\geq 0}$ is given by a Vasicek model, then, for each $0 \leq u < t$,

$$r_t = e^{-\alpha(t-u)} r_u + \alpha \int_u^t e^{-\alpha(t-s)} b(s)\,\mathrm{d}s + \sigma \int_u^t e^{-\alpha(t-s)}\,\mathrm{d}W_s\,, \tag{4.59}$$

which, for $r_0 \in \mathbb{R}_0^+$, defines a Gaussian process. Moreover, the distribution of r_t under the condition $\{r_u = x \in \mathbb{R}_0^+\}$ is $N(\alpha_t, \sigma_t^2)$, where

$$\alpha_t = e^{-\alpha(t-u)} x + \alpha \int_u^t e^{-\alpha(t-s)} b(s)\,\mathrm{d}s\,, \tag{4.60a}$$

$$\sigma_t^2 = \sigma^2 \int_u^t e^{-2\alpha(t-s)}\,\mathrm{d}s = \frac{\sigma^2}{2\alpha}\left(1 - e^{-2\alpha(t-u)}\right), \tag{4.60b}$$

which, for $0 = t_0 < t_1 < \cdots < t_k$ and $z_1, z_2, \ldots$ representing the output of i.i.d. copies of an $N(0,1)$-distributed random variable, provides the recursion

$$r_0 \in \mathbb{R}_0^+ : \quad \text{given initial value,} \tag{4.61a}$$

$$r_i := e^{-\alpha(t_i - t_{i-1})} r_{i-1} + \alpha \int_{t_{i-1}}^{t_i} e^{-\alpha(t_i - s)} b(s)\,\mathrm{d}s + \frac{z_i\,\sigma}{\sqrt{2\alpha}} \sqrt{1 - e^{-2\alpha(t_i - t_{i-1})}} \quad \text{for } i = 1, \ldots, k, \tag{4.61b}$$

for an exact simulation of $(r_t)_{t\geq 0}$. Carrying out the integral in (4.61b) exactly might not always be feasible. Approximating b as constantly equal to $b(t_{i-i})$ on $[t_{i-1}, t_i]$, we can compute the resulting approximation of the integral:

$$\alpha \int_{t_{i-1}}^{t_i} e^{-\alpha(t_i - s)} b(s)\,\mathrm{d}s \approx \alpha\, b(t_{i-i}) \left[\frac{e^{\alpha(s-t_i)}}{\alpha}\right]_{t_{i-1}}^{t_i} = b(t_{i-i})\,\left(1 - e^{-\alpha(t_i - t_{i-1})}\right). \tag{4.62}$$

In addition, depending on the situation, it might also be worth gaining some efficiency (possibly at the price of sacrificing some accuracy) by using the approximation $e^x \approx 1 + x$ in (4.61b) as well as (4.62), replacing it with the (no longer exact) recursion formula

$$\begin{aligned} r_i &:= \big(1 - \alpha(t_i - t_{i-1})\big)\, r_{i-1} + b(t_{i-i})\,\alpha\,(t_i - t_{i-1}) + z_i\,\sigma\,\sqrt{t_i - t_{i-1}} \\ &= r_{i-1} + \alpha\,\big(b(t_{i-i}) - r_{i-1}\big)(t_i - t_{i-1}) + z_i\,\sigma\,\sqrt{t_i - t_{i-1}}. \end{aligned} \tag{4.63}$$

Remark 4.26. For Ho-Lee models, the bond price $B(0,T)$ could be computed as a closed-form formula, cf. (4.50). Even though, Vasicek models are, in general, more complicated, for a short rate $(r_t)_{t\geq 0}$ given by a Vasicek model, $B(0,T)$ according to (4.44) can still be computed in closed form. The computation is carried out in [Gla04, pp. 113-114] and yields

$$B(0,T) = \exp\big(-A(0,T)\,r_0 + C(0,T)\big), \tag{4.64}$$

where

$$A(0,T) := \frac{1-e^{-\alpha T}}{\alpha}, \tag{4.65a}$$

$$C(0,T) := -\alpha \int_0^T \int_0^u e^{-\alpha(u-s)} b(s)\,\mathrm{d}s\,\mathrm{d}u + \frac{\sigma^2}{2\alpha^2}\left(T + \frac{1-e^{-2\alpha T}}{2\alpha} + \frac{2(e^{-\alpha T}-1)}{\alpha}\right). \tag{4.65b}$$

5 Variance Reduction Techniques

5.1 Control Variates

Idea: Use the errors in estimates of known quantities to reduce the error in an estimate of an unknown quantity.

Let $(\Omega, \mathcal{A}, P)$ be a probability space and assume all random variables occurring subsequently in this section are defined on Ω, if nothing else is indicated.

Suppose, the goal is to estimate the expected value $E(Y)$ of a real-valued and square-integrable random variable Y, whereas we already know $E(X)$ for another real-valued and square-integrable random variable X with positive variance $V(X) > 0$. If X and Y are "not too different" in a sense that we will have to make precise below, then knowledge about X can be used to improve the method for estimating $E(Y)$.

Assume $Y_1, Y_2, \ldots$ is a sequence of i.i.d. copies of Y. Then, letting

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i \quad \text{for each } n \in \mathbb{N}, \tag{5.1}$$

the strong law of large numbers Th. B.32 implies

$$\bar{Y}_n \to E(Y) \quad \text{almost surely.} \tag{5.2}$$

That is why $\bar{Y}_n$ is often used as a first estimate for $E(Y)$. Our goal is to improve on this estimate, employing X. To this end, we assume $X_1, X_2, \ldots$ is a sequence of i.i.d. copies of X, with the additional property that the sequence of pairs $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d. copies of (X, Y). Analogous to (5.1), let

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \quad \text{for each } n \in \mathbb{N}, \tag{5.3}$$

such that

$$\bar{X}_n \to E(X) \quad \text{almost surely.} \tag{5.4}$$

For each $b \in \mathbb{R}$, we define the random variables

$$Y_i(b) : \Omega \longrightarrow \mathbb{R}, \quad Y_i(b) := Y_i - b\left(X_i - E(X)\right) \quad \text{for each } i \in \mathbb{N}, \tag{5.5}$$

and

$$\bar{Y}_n(b) : \Omega \longrightarrow \mathbb{R}, \quad \bar{Y}_n(b) := \frac{1}{n}\sum_{i=1}^{n} Y_i(b) = \bar{Y}_n - b\left(\bar{X}_n - E(X)\right) \quad \text{for each } n \in \mathbb{N}. \tag{5.6}$$

One calls the $\bar{Y}_n(b)$ a *control variate estimator* of $E(Y)$ with *control* $\bar{X}_n - E(X)$.

This estimator is unbiased:

$$\begin{aligned} E\big(\bar{Y}_n(b)\big) &= E\Big(\bar{Y}_n - b\left(\bar{X}_n - E(X)\right)\Big) = E(\bar{Y}_n) - b\left(E(X) - E(X)\right) \\ &= E(\bar{Y}_n) = E(Y). \end{aligned} \tag{5.7}$$

This estimator is also consistent, since, in the sense of convergence almost surely:

$$\lim_{n\to\infty} \bar{Y}_n(b) = \lim_{n\to\infty} \Big(\bar{Y}_n - b\left(\bar{X}_n - E(X)\right)\Big) = E(Y). \tag{5.8}$$

We denote the standard deviations

$$\sigma_X := \sqrt{V(X)}, \quad \sigma_Y := \sqrt{V(Y)}, \quad \sigma(b) := \sqrt{V\big(Y_i(b)\big)}, \tag{5.9}$$

and the correlation

$$\rho_{XY} := \mathrm{Cor}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X\sigma_Y}, \tag{5.10}$$

where it is noted that $X, Y \in L^2(P)$ implies $XY \in L^1(P)$ (i.e. XY is integrable) by the Hölder inequality. For each $i \in \mathbb{N}$, we compute the variance $\sigma^2(b)$ of $Y_i(b)$, confirming it does not depend on i:

$$\begin{aligned} \sigma^2(b) := V\big(Y_i(b)\big) &= E\Big(\big(Y_i(b) - E(Y_i(b))\big)^2\Big) = E\Big(\big(Y_i - E(Y) - b\left(X_i - E(X)\right)\big)^2\Big) \\ &= \sigma_Y^2 - 2b\,E\Big(\big(Y_i - E(Y)\big)\big(X_i - E(X)\big)\Big) + b^2\sigma_X^2 \\ &= \sigma_Y^2 - 2b\,\sigma_X\sigma_Y\rho_{X,Y} + b^2\sigma_X^2. \end{aligned} \tag{5.11}$$

The pairwise independence of the Y_i implies $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ and, thus,

$$V(\bar{Y}_n) = \frac{\sigma_Y^2}{n} \quad \text{for each } n \in \mathbb{N}. \tag{5.12}$$

Pairwise independence of the (X_i, Y_i) and Th. B.10 imply the pairwise independence of the $Y_i(b)$, in particular, $\mathrm{Cov}\left(Y_i(b), Y_j(b)\right) = 0$ for $i \neq j$. In consequence,

$$V\big(\bar{Y}_n(b)\big) = \frac{\sigma^2(b)}{n} \quad \text{for each } n \in \mathbb{N}. \tag{5.13}$$

Combining (5.11) – (5.13), we see the control variate estimator $\bar{Y}_n(b)$ achieves a *variance reduction* if, and only if,

$$\sigma_Y^2 - 2b\,\sigma_X\sigma_Y\rho_{X,Y} + b^2\sigma_X^2 = \sigma^2(b) = n\,V\big(\bar{Y}_n(b)\big) < n\,V(\bar{Y}_n) = \sigma_Y^2, \tag{5.14a}$$

i.e. if, and only if,

$$b^2\sigma_X < 2b\,\sigma_Y\rho_{X,Y}. \tag{5.14b}$$

The derivative $(\sigma^2)'(b) = -2\sigma_X\sigma_Y\rho_{X,Y} + 2\sigma_X^2\,b$ has zero

$$b_{\min} := \frac{\sigma_Y\rho_{X,Y}}{\sigma_X} = \frac{\mathrm{Cov}(X,Y)}{V(X)} \tag{5.15}$$

and $(\sigma^2)'' \equiv 2\sigma_X^2 > 0$, showing the variance of $\bar{Y}_n(b)$ becomes minimal at $b_{\min}$.

Plugging $b_{\min}$ into (5.11) yields the following ratio between the variance of the optimally controlled estimator and the variance of the uncontrolled estimator, called the *variance reduction ratio*:

$$\frac{V\big(\bar{Y}_n(b_{\min})\big)}{V(\bar{Y}_n)} = \frac{\sigma^2(b_{\min})}{\sigma_Y^2} = \frac{\sigma_Y^2 - 2\,\sigma_Y^2\rho_{X,Y}^2 + \sigma_Y^2\rho_{X,Y}^2}{\sigma_Y^2} = 1 - \rho_{X,Y}^2. \tag{5.16}$$

Remark 5.1. (a) Thus, the effectiveness of a control variate as measured by the variance reduction ratio depends on the strength of the correlation $\rho_{X,Y}$ between X and Y. The sign of $\rho_{X,Y}$ does not affect the ratio, as it is absorbed in $b_{\min}$.

(b) Taking the variance as a measure for the accuracy of the estimators $\bar{Y}_n$ and $\bar{Y}_n(b_{\min})$, respectively, the number of steps is reduced from n for the uncontrolled estimator to $n\,(1-\rho_{X,Y}^2)$ for the controlled estimator. If the computational effort for $\bar{Y}_n$ and $\bar{Y}_n(b_{\min})$ is approximately the same, then $1/(1-\rho_{X,Y}^2)$ is the speed-up achieved by using the control variate estimator as compared to using the uncontrolled one.

(c) Due to the form of (5.16), the usefulness of X drops rapidly with $|\rho_{X,Y}|$: For example, $1 - 0.95^2 = 0.0975$, $1 - 0.9^2 = 0.19$, $1 - 0.7^2 = 0.51$, i.e. the approximate speed-up is reduced from a factor of 10 to 5 to 2.

In (5.16) and Rem. 5.1, we made use of the optimal coefficient $b_{\min}$ given by (5.15). However, since our initial goal was to estimate the *unknown* quantity $E(Y)$, it is not likely that the quantities σ_Y and $\rho_{X,Y}$ are better known than $E(Y)$, i.e. $b_{\min}$ will usually be unavailable in practice. A possible solution is the introduction of the approximating random variables

$$B_n := \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \quad \text{for each } n \in \mathbb{N} \tag{5.17}$$

and the estimator

$$\bar{Z}_n := \bar{Y}_n - \frac{1}{n}\sum_{i=1}^n B_n\big(X_i - E(X)\big) \quad \text{for each } n \in \mathbb{N}. \tag{5.18}$$

Remark 5.2. (a) It is an exercise to show that the strong law of large numbers Th. B.32 implies, in the sense of convergence almost surely,

$$\lim_{n\to\infty} B_n = b_{\min}. \tag{5.19}$$

(b) In general, the estimator $\bar{Z}_n$ is no longer unbiased, where the bias is

$$E(\bar{Z}_n) - E(Y) = -E\left(\frac{1}{n}\sum_{i=1}^{n} B_n\big(X_i - E(X)\big)\right)$$
$$= -E\left(\frac{B_n}{n}\sum_{i=1}^{n}\big(X_i - E(X)\big)\right) = -E\Big(B_n\big(\bar{X}_n - E(X)\big)\Big). \quad (5.20)$$

In general, the bias in (5.20) is nonzero since B_n and $\bar{X}_n$ are not independent. Due to the law of large numbers and (5.19), $\lim_{n\to\infty} B_n\big(\bar{X}_n - E(X)\big) = 0$ almost surely, such that the introduced bias does usually not pose a significant problem if n is sufficiently large. It is, however, an issue for small samples (i.e. small n).

Example 5.3. In Sec. 1.1, it was described how to use estimators for the expected value of random variables to numerically compute integrals. The present example illustrates the use of control variates in this (rather academic) situation: Suppose, we want to use Monte Carlo to compute the integral $\int_0^1 f(x)\,\mathrm{d}x$ for

$$f:\,[0,1]\longrightarrow\mathbb{R},\quad f(x) := 4\sqrt{1-x^2}. \quad (5.21)$$

Letting $U_1, U_2, \ldots$ denote a sequence of i.i.d. random variables, uniformly distributed on $[0,1]$, and recalling (1.5), in terms of the language of the present section, we set

$$Y_i := f \circ U_i \quad \text{for each } i \in \mathbb{N}. \quad (5.22)$$

To obtain a control, we start with a function that is "close" to f, but easy to integrate, say,

$$h:\,[0,1]\longrightarrow\mathbb{R},\quad h(x) := 4 - 4x, \quad (5.23)$$

and set

$$X_i := h \circ U_i \quad \text{for each } i \in \mathbb{N}. \quad (5.24)$$

Since the U_i are i.i.d., so are the X_i and the Y_i by Th. B.10. As the pairs (U_i, U_i) are also i.i.d., so are the pairs (X_i, Y_i), once again by Th. B.10, verifying the correct setting for control variates.

Furthermore,

$$E(X_i) = \int_0^1 h(x)\,\mathrm{d}x \;= 4\left[x - \frac{x^2}{2}\right]_0^1 = 2 \quad (5.25)$$

and, for each $b \in \mathbb{R}$,

$$Y_i(b) := Y_i - b\,\big(X_i - E(X_i)\big) = f \circ U_i - b(h \circ U_i - 2) \quad \text{for each } i \in \mathbb{N}. \quad (5.26)$$

Moreover,

$$\sigma^2_{X_i} = -4 + \int_0^1 h^2(x)\,\mathrm{d}x \;= -4 + 16\left[x - x^2 + \frac{x^3}{3}\right]_0^1 = \frac{4}{3}. \quad (5.27)$$

The example is academic, since the antiderivative of f is not too hard to find – it is given by

$$F:\,[0,1]\longrightarrow\mathbb{R},\quad F(x) = 2\left(x\sqrt{1-x^2} + \arcsin(x)\right), \quad (5.28)$$

as is easily verified by differentiation, yielding the exact integral

$$E(Y_i) = \int_0^1 f(x)\,\mathrm{d}x \;= 2\arcsin(1) = \pi. \tag{5.29}$$

On the other hand, this allows us to actually compute the optimal control coefficient $b_{\min}$ and resulting variance reduction, which we now do, out of curiosity and for the purpose of illustration. The covariance is

$$\begin{aligned}\mathrm{Cov}(X_i, Y_i) &= E(X_iY_i) - E(X_i)E(Y_i) = -2\pi + \int_0^1 (fh)(x)\,\mathrm{d}x \\ &= -2\pi + 4\pi - 16\int_0^1 x\sqrt{1-x^2}\,\mathrm{d}x \;= 2\pi + \frac{16}{3}\left[(1-x^2)^{\frac{3}{2}}\right]_0^1 \\ &= 2\pi - \frac{16}{3} \approx 0.95,\end{aligned} \tag{5.30}$$

implying

$$b_{\min} = \frac{\mathrm{Cov}(X_i, Y_i)}{\sigma^2_{X_i}} \approx \frac{3\cdot 0.95}{4} \approx 0.71\,. \tag{5.31}$$

Since

$$\sigma^2_{Y_i} = -\pi^2 + \int_0^1 f^2(x)\,\mathrm{d}x \;= -\pi^2 + 16\int_0^1 (1-x^2)\,\mathrm{d}x \;= -\pi^2 + \frac{32}{3} \approx 0.797, \tag{5.32}$$

the variance reduction ratio is $1 - \rho^2_{X_iY_i} = 1 - \mathrm{Cov}(X_i, Y_i)^2/(\sigma^2_{X_i}\sigma^2_{Y_i}) \approx 0.15$.

Without the knowledge of $b_{\min}$, one might have tried $b = 1$. While less effective than using $b_{\min}$, $b = 1$ still achieves a significant variance reduction:

$$\begin{aligned}V\big(Y_i(1)\big) &= \int_0^1 \big(f(x) - h(x) + 2\big)^2\,\mathrm{d}x \;- \left(\int_0^1 \big(f(x) - h(x) + 2\big)\,\mathrm{d}x\right)^2 \\ &= \int_0^1 \Big(f^2(x) - 2(fh)(x) + h^2(x) + 4f(x) - 4h(x) + 4\Big)\,\mathrm{d}x \;- \pi^2 \\ &= \frac{32}{3} - 8\pi + \frac{32}{3} + \frac{16}{3} + 4\pi - 4 - \pi^2 \\ &= \frac{68}{3} - 4\pi - \pi^2 \approx 0.231,\end{aligned} \tag{5.33}$$

i.e. less than one third of $\sigma^2_{Y_i}$.

Example 5.4. While a good control variate will usually need to take advantage of problem-dependent features, since virtually all simulations start with an i.i.d. sequence $U_1, U_2, \ldots$ of uniformly distributed, $[0,1]$-valued random variables, a control based on $X_i := U_i$, $E(X_i) = \frac{1}{2}$ is virtually always available. Similarly, for simulations based on an i.i.d. sequence $Z_1, Z_2, \ldots$ of $N(0,1)$-distributed random variables, one can always use a control based on $X_i := Z_i$, $E(X_i) = 0$. Such controls are sometimes referred to as *primitive controls*.

Example 5.5. Control variates can often be used in the mathematical finance application, where Y represents the unknown price of a derivative security for some underlying asset S. A reasonable mathematical finance model should at least be arbitrage-free, and one can show that, for an arbitrage-free model, the discounted price of S, denoted by $\tilde{S}$, should be a martingale, such that $E(\tilde{S}_t) = s_0$ for each time $t \geq 0$, where s_0 is the initial value for the asset price. If, as in Sec. 1.2, we consider a risk-neutral model with constant continuously compounded interest rate r, then $\tilde{S}_t = e^{-rt}S_t$, $E(e^{-rt}S_t) = s_0$, $E(S_t) = e^{rt}s_0$. If we are interested in $E(Y_T)$ for some $T > 0$, then the Y_i should be i.i.d. copies of Y_T and the X_i should be i.i.d. copies of S_T such that the control variate estimator of (5.6) becomes

$$\underset{b\in\mathbb{R}}{\forall} \quad \underset{n\in\mathbb{N}}{\forall} \quad \bar{Y}_n(b) = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - b\left(\bar{X}_i - E(S_T)\right)\Big) = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - b\left(\bar{X}_i - e^{rt}s_0\right)\Big), \quad (5.34)$$

where one would replace b with B_n from (5.17), if a good idea for choosing b is not available. If Y represents the price of a European call option (which was denoted by C in Sec. 1.2), then $Y_T = e^{-rT}(S_T - K)^+$, where K is the option's strike price (cf. (1.14)). In this case, the correlation between Y_T and S_T, and, thus, the effectiveness of the control variate, are determined by the distribution of S_T and by the size of K (where the correlation is perfect for $K = 0$ and typically low for large K).

5.2 Antithetic Variates

The basic setting is still the same as in the previous section on control variates. In particular, all random variables are assumed to be defined on the probability space $(\Omega, \mathcal{A}, P)$, unless indicated otherwise. The goal is to estimate the unknown $E(Y)$ for the random variable Y. Once again, assume $Y_1, Y_2, \ldots$ is a sequence of i.i.d. copies of the real-valued and square-integrable random variable Y.

The idea for control variates was to reduce variance by modifying the simple estimator (5.1) by using i.i.d. copies of a random variable X with known $E(X)$. For antithetic variates, the idea is to use $X \sim Y$ (i.e. X, Y identically distributed), but X, Y *not* independent. As in Sec. 5.1, let $X_1, X_2, \ldots$ be i.i.d. copies of X, where one requires the sequence of pairs $(X_1, Y_1), (X_2, Y_2), \ldots$ to be i.i.d. as well. The aim is to use the correlation between X_i and Y_i to reduce variance. As before, the Hölder inequality implies that $XY, X_1Y_1, X_2Y_2, \ldots$ are integrable.

To proceed, define the simple estimators $\bar{Y}_n$ and $\bar{X}_n$ as in (5.1) and (5.3), respectively, plus the new estimator

$$\underset{n\in\mathbb{N}}{\forall} \quad \bar{Y}_{n,\mathrm{A}} := \frac{\bar{Y}_n + \bar{X}_n}{2}. \quad (5.35)$$

Clearly, $\bar{Y}_{n,\mathrm{A}}$ is unbiased and consistent. To assess if the estimator $\bar{Y}_{n,\mathrm{A}}$ constitutes an improvement over $\bar{Y}_n$, it is incorrect to compare the variances $V(\bar{Y}_{n,\mathrm{A}})$ and $V(\bar{Y}_n)$, since $\bar{Y}_{n,\mathrm{A}}$ has twice as many summands as $\bar{Y}_n$. So one needs to compare $V(\bar{Y}_{n,\mathrm{A}})$ and $V(\bar{Y}_{2n})$:

The estimator $\bar{Y}_{n,\mathrm{A}}$ achieves a variance reduction if, and only if,

$$V(\bar{Y}_{n,\mathrm{A}}) \overset{(*)}{=} \frac{1}{n} V\left(\frac{X_i + Y_i}{2}\right) < V(\bar{Y}_{2n}) = \frac{\sigma_Y^2}{2n}, \tag{5.36}$$

where, for the equality at $(*)$, it was used that, for $i, j \in \mathbb{N}$ with $i \neq j$, the pairs (X_i, Y_i), (X_j, Y_j) are assumed independent, i.e. $X_i + Y_i$ and $X_j + Y_j$ are independent by Th. B.10, implying $\mathrm{Cov}(X_i + Y_i, X_j + Y_j) = 0$. Clearly, (5.36) is equivalent to

$$V(X + Y) < 2\, V(Y). \tag{5.37}$$

Since

$$V(X + Y) = V(X) + V(Y) + 2\,\mathrm{Cov}(X, Y) = 2V(Y) + 2\,\mathrm{Cov}(X, Y), \tag{5.38}$$

both (5.36) and (5.37) are equivalent to

$$\mathrm{Cov}(X, Y) < 0. \tag{5.39}$$

This motivates the following definition:

Definition 5.6. The real-valued integrable random variables X, Y with integrable XY are called *antithetic* if, and only if, X and Y are identically distributed with $\mathrm{Cov}(X, Y) < 0$.

—

Antithetic random variables can typically be obtained from monotonicity properties of the involved random variables, provided such monotonicity properties are present. In simple situations, the following Th. 5.7 can be applied. Many related, more general, theorems can be found in the literature.

Theorem 5.7. *Let U be a real-valued random variable with range $\mathcal{R}(U)$. If $h, k : \mathcal{R}(U) \longrightarrow \mathbb{R}$ are both integrable, hk is also integrable, and h, k are both nondecreasing or both nonincreasing, then, defining random variables*

$$Y := k \circ U, \quad X := h \circ U, \tag{5.40}$$

one obtains

$$\mathrm{Cov}(X, Y) \geq 0. \tag{5.41}$$

Proof. Due to the integrability hypotheses, $E(X), E(Y), E(XY)$ all exist. Let V be a random variable such that U, V are i.i.d. Due to the monotonicity hypotheses,

$$\big(h(U) - h(V)\big)\big(k(U) - k(V)\big) \geq 0. \tag{5.42}$$

Thus, we can estimate

$$\begin{aligned} 0 &\leq E\Big(\big(h(U) - h(V)\big)\big(k(U) - k(V)\big)\Big) \\ &= E\big(h(U)k(U)\big) - E\big(h(U)k(V)\big) - E\big(h(V)k(U)\big) + E\big(h(V)k(V)\big) \\ &\overset{(*)}{=} 2\,E\big(h(U)k(U)\big) - 2\,E\big(h(U)\big)\,E\big(k(U)\big) \\ &= 2\,\mathrm{Cov}\,\big(h(U), k(U)\big) = 2\,\mathrm{Cov}(X, Y), \end{aligned} \tag{5.43}$$

thereby establishing the case. At "(∗)", it was used that, by Th. B.10, the independence of U, V implies the independence of $h(U), k(V)$ and of $k(U), h(V)$. ■

Corollary 5.8. *Let U be a random variable, uniformly distributed on $[0,1]$. If $k : [0,1] \longrightarrow \mathbb{R}$ is square-integrable and nondecreasing or nonincreasing, then, defining random variables*

$$Y := k \circ U, \quad X := k \circ (1-U), \tag{5.44}$$

one obtains

$$\operatorname{Cov}(X,Y) \leq 0. \tag{5.45}$$

Proof. We can apply Th. 5.7 with k and

$$h : [0,1] \longrightarrow \mathbb{R}, \quad h(x) := -k(1-x), \tag{5.46}$$

since $k \in L^2[0,1]$ implies $h \in L^2[0,1]$ and $kh \in L^1[0,1]$ by the Hölder inequality, and since h has the same monotonicity property as k. Theorem 5.7 yields

$$-\operatorname{Cov}(X,Y) = \operatorname{Cov}(-X,Y) \geq 0, \tag{5.47}$$

proving the corollary. ■

Example 5.9. Let us, once again, consider the problem of Ex. 5.3, i.e. the problem of using Monte Carlo to compute the integral $\int_0^1 f(x)\,\mathrm{d}x$ with f given by (5.21), now employing the method of antithetic variates. If U is a random variable, uniformly distributed on $[0,1]$, then the fact that f is decreasing and Cor. 5.8 suggest using

$$Y := f \circ U, \quad X := f \circ (1-U). \tag{5.48}$$

The corresponding estimator is

$$\bar{Y}_{n,\mathrm{A}} := \frac{1}{2n} \sum_{i=1}^{n} \big(f(U_i) + f(1-U_i) \big) \tag{5.49}$$

if the U_i are i.i.d. copies of U. Out of curiosity, one can compute $E(XY)$ numerically,

$$E(XY) = \int_0^1 f(x) f(1-x)\,\mathrm{d}x = \int_0^1 \sqrt{x(2-x)(1-x^2)}\,\mathrm{d}x \approx 0.581, \tag{5.50}$$

yielding

$$\operatorname{Cov}(X,Y) = E(XY) - \pi^2 \approx -0.578 \tag{5.51}$$

and the variance reduction ratio

$$\begin{aligned} \frac{V(\bar{Y}_{n,\mathrm{A}})}{V(\bar{Y}_{2n})} &\overset{(5.36)}{=} \frac{V(X+Y)}{2\,V(Y)} = \frac{2V(Y) + 2\operatorname{Cov}(X,Y)}{2\,V(Y)} \\ &= 1 + \frac{\operatorname{Cov}(X,Y)}{V(Y)} \overset{(5.51),(5.32)}{\approx} 0.275. \end{aligned} \tag{5.52}$$

5.3 Stratified Sampling

We start with a motivating example:

Example 5.10. Suppose we want to poll the job approval of the head of government, who happens to belong to party A. We might poll $n = 1000$ people and ask each person if they approve of the job the head of government is doing. If we model the answer of the ith person by a $\{0, 1\}$-valued random variable X_i, where

$$X_i = \begin{cases} 1 & \text{for answer "yes"}, \\ 0 & \text{for answer "no"}, \end{cases} \tag{5.53}$$

then *simple sampling* means computing $\frac{1}{n}\sum_{i=1}^{n} X_i$, which is the fraction of people having answered "yes".

Now suppose we have also asked each person, which party they voted for in the last election. If it turns out that 50% had voted for party A, 15% for B, 10% for C, and 25% had voted for others or not at all, whereas the true outcome of the election had been 35% A, 20% B, 10% for C, and 35% others or none, then, considering the head of government belongs to A, the fraction of people in the sample answering "yes" is likely to be higher than in the population overall.

This cause of sampling error could be eliminated by *stratifying* the sampling such that the sample of 1000 people being polled has precisely the right number from each stratum, namely, in the described example, 350 A voters, 200 B voters, 100 C voters, and 350 people who voted for others or not at all.

—

Let us consider a corresponding situation more abstractly, where we remain in a similar general setting as in the previous sections. However, we will introduce an intermediate layer, i.e., in addition to the probability space $(\Omega, \mathcal{A}, P)$, a measurable space $(\mathcal{S}, \mathcal{B})$, with the random variable $Y : \Omega \longrightarrow \mathcal{S}$ is considered, where the goal is to estimate $E(f \circ Y)$ for some measurable function $f : \mathcal{S} \longrightarrow \mathbb{R}$. If $Y_1, Y_2, \ldots$ are i.i.d. copies of Y, then our simple estimator now is $\frac{1}{n}\sum_{i=1}^{n} f(Y_i)$.

Definition 5.11. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\mathcal{S}, \mathcal{B})$ a measurable space, $Y : \Omega \longrightarrow \mathcal{S}$ a random variable.

(a) A finite sequence $(S_1, \ldots, S_M)$, $M \in \mathbb{N}$, of pairwise disjoint measurable subsets of $\mathcal{S}$, is called a sequence of *strata* for Y with corresponding probabilities $(a_1, \ldots, a_M)$, if and only if, there exists a P_Y-null set S_0 such that

$$\mathcal{S} = S_0 \,\dot\cup \bigdot_{i=1,\ldots,M} S_i, \tag{5.54a}$$

and

$$\forall_{i=1,\ldots,M} \quad a_i = P_Y(S_i) = P\{Y \in S_i\} > 0, \tag{5.54b}$$

i.e. strata partition $\mathcal{S}$ (up to a P_Y-null set) into disjoint subsets of positive probability with respect to the distribution of Y.

(b) Let $f : \mathcal{S} \longrightarrow \mathbb{R}$ be measurable such that $f \circ Y \in L^2(P)$. Moreover, let $(S_1, \ldots, S_M)$ be strata for Y with corresponding probabilities $(a_1, \ldots, a_M)$ as defined in (a), let $n, n_1, \ldots, n_M \in \mathbb{N}$ be such that

$$n = \sum_{i=1}^{M} n_i, \tag{5.55}$$

and, for each $i \in \{1, \ldots, M\}$, let $Y_1^{(i)}, \ldots, Y_{n_i}^{(i)} : \Omega \longrightarrow S_i$ be i.i.d. random variables such that the entire family $(Y_k^{(i)})_{i \in \{1,\ldots,M\}, k \in \{1,\ldots,n_i\}}$ is independent, and

$$\underset{i=1,\ldots,M}{\forall} \quad \underset{k=1,\ldots,n_i}{\forall} \quad Y_k^{(i)} \sim Y | S_i, \tag{5.56a}$$

i.e.

$$\underset{k=1,\ldots,n_i}{\forall} \quad \underset{B \in \mathcal{B}}{\forall} \quad P\{Y_k^{(i)} \in B \cap S_i\} = \frac{P\{Y \in B \cap S_i\}}{a_i}. \tag{5.56b}$$

Defining

$$\underset{i=1,\ldots,M}{\forall} \quad T_i : \Omega \longrightarrow \mathbb{R}, \quad T_i := \frac{1}{n_i} \sum_{k=1}^{n_i} f(Y_k^{(i)}), \tag{5.57}$$

the random variable

$$T : \Omega \longrightarrow \mathbb{R}, \quad T := \sum_{i=1}^{M} a_i T_i, \tag{5.58}$$

is called the *stratified estimator* of $E(f \circ Y)$ corresponding to the strata $S_1, \ldots, S_M$.

Note that the stratified estimator does not only depend on the strata and the $a_1, \ldots, a_M$, but also on $n, n_1, \ldots, n_M$. One speaks of *proportional allocation* provided $n_i = n a_i$ holds for each $i = 1, \ldots, M$.

Remark 5.12. From Def. 5.11, we can draw some simple conclusions:

(a) As an immediate consequence of (5.54a) and (5.54b), one obtains

$$\sum_{i=1}^{M} a_i = 1. \tag{5.59}$$

(b) From (5.57) and (5.56), one obtains

$$\underset{i=1,\ldots,M}{\forall} \quad E(T_i) = E(f \circ Y_1^{(i)}) = \int_{S_i} f \, \mathrm{d}P_{Y_1^{(i)}} = \int_{S_i} \frac{f}{a_i} \, \mathrm{d}P_Y =: \frac{E_i(f \circ Y)}{a_i}. \tag{5.60}$$

(c) The stratified estimator T of (5.58) is unbiased:

$$E(T) = \sum_{i=1}^{M} a_i E(T_i) \overset{\text{(b)}}{=} \sum_{i=1}^{M} \int_{S_i} f \, \mathrm{d}P_Y = \int_{\mathcal{S}} f \, \mathrm{d}P_Y = E(f \circ Y). \tag{5.61}$$

(d) For each $i = 1, \dots, M$, due to the independence of the $Y_1^{(i)}, \dots, Y_{n_i}^{(i)}$, the maps $f \circ Y_1^{(i)}, \dots, f \circ Y_{n_i}^{(i)}$ are also independent by Th. B.10, and the variance of T_i is

$$\begin{aligned} V(T_i) &= \frac{V(f \circ Y_1^{(i)})}{n_i} = \frac{1}{n_i}\left(E\left((f \circ Y_1^{(i)})^2\right) - \left(E(f \circ Y_1^{(i)})\right)^2\right) \\ &\overset{(5.60)}{=} \frac{1}{n_i}\left(\int_{S_i} f^2 \,\mathrm{d}P_{Y_1^{(i)}} - \left(\frac{E_i(f \circ Y)}{a_i}\right)^2\right) \\ &= \frac{1}{n_i}\left(\int_{S_i} \frac{f^2}{a_i} \,\mathrm{d}P_Y - \left(\frac{E_i(f \circ Y)}{a_i}\right)^2\right). \end{aligned} \tag{5.62}$$

(e) The independence of the $Y_k^{(i)}$ implies the independence of $T_1, \dots, T_M$ (exercise), such that

$$V(T) = \sum_{i=1}^{M} a_i^2 V(T_i). \tag{5.63}$$

Stratified estimators can be useful to reduce variance according to the following result:

Theorem 5.13. *In the situation of Def.* 5.11*, the variance of the stratified estimator T using proportional allocation (i.e. $n_i = na_i$ for each $i = 1, \dots, M$) satisfies*

$$V\left(\frac{1}{n}\sum_{i=1}^{n} f(Y_i)\right) = V(T) + \frac{1}{n}\sum_{i=1}^{M} a_i \left(\frac{E_i(f \circ Y)}{a_i} - E(f \circ Y)\right)^2, \tag{5.64}$$

where the $E_i(f \circ Y)$ are as defined in (5.60)*. In particular, the variance of the proportionally allocated stratified estimator is strictly less than the variance of the simple estimator if, and only if, at least one of the summands in* (5.64) *is positive.*

Proof. We compute, starting with the right-hand side of (5.64),

$$\begin{aligned} & V(T) + \frac{1}{n}\sum_{i=1}^{M} a_i \left(\frac{E_i(f \circ Y)}{a_i} - E(f \circ Y)\right)^2 \\ \overset{(5.63),(5.62)}{=}\ & \sum_{i=1}^{M} a_i^2 \frac{1}{n_i}\left(\int_{S_i} \frac{f^2}{a_i} \,\mathrm{d}P_Y - \left(\frac{E_i(f \circ Y)}{a_i}\right)^2\right) \\ & + \frac{1}{n}\sum_{i=1}^{M} a_i \left(\frac{E_i(f \circ Y)}{a_i}\right)^2 - \frac{2E(f \circ Y)}{n}\sum_{i=1}^{M} E_i(f \circ Y) + \frac{E(f \circ Y)^2}{n}\sum_{i=1}^{M} a_i \\ \overset{n_i = na_i}{=}\ & \sum_{i=1}^{M} a_i^2 \frac{1}{na_i}\left(\int_{S_i} \frac{f^2}{a_i} \,\mathrm{d}P_Y - \left(\frac{E_i(f \circ Y)}{a_i}\right)^2\right) \\ & + \frac{1}{n}\sum_{i=1}^{M} a_i \left(\frac{E_i(f \circ Y)}{a_i}\right)^2 - \frac{2E(f \circ Y)}{n} E(f \circ Y) + \frac{E(f \circ Y)^2}{n} \\ =\ & \frac{1}{n}\int_{\mathcal{S}} f^2 \,\mathrm{d}P_Y - \frac{E(f \circ Y)^2}{n} = \frac{V(f \circ Y)}{n} = V\left(\frac{1}{n}\sum_{i=1}^{n} f(Y_i)\right), \end{aligned} \tag{5.65}$$

which establishes the case. ■

Remark 5.14. The assumption of proportional allocation in Th. 5.13 might seem like a strong restriction, since, in particular, it means that na_i has to be an integer for each i. However, in practise, it is often possible to taylor the a_i by choosing the strata S_i appropriately. And one can also often still obtain a variance reduction if one rounds na_i to the closest integer, but the computations become more technical.

—

Even though proportional allocation will usually result in a variance reduction according to Th. 5.13, this choice for the n_i and, thus, for the stratified estimator, is, in general, not optimal, as can be seen from the following result.

Theorem 5.15. *Let $Y : \Omega \longrightarrow \mathcal{S}$ and $f : \mathcal{S} \longrightarrow \mathbb{R}$ be as in Def. 5.11, in particular, $f \circ Y \in L^2(P)$. Fix strata $S_1, \ldots, S_M$, $M \in \mathbb{N}$, for Y with probabilities $a_1, \ldots, a_M$. Also fix $n \in \mathbb{N}$. Define*

$$\forall_{i=1,\ldots,M} \quad \sigma_i := \sqrt{V(f \circ Y_1^{(i)})}. \tag{5.66}$$

If

$$\forall_{i=1,\ldots,M} \quad n_i := \frac{a_i \sigma_i}{\sum_{j=1}^M a_j \sigma_j}\, n \in \mathbb{N}, \tag{5.67}$$

then this provides a stratified estimator of minimal variance. More precisely, if for each tuple $(m_1, \ldots, m_M) \in \mathbb{N}^M$ such that $n = \sum_{i=1}^M m_i$, the corresponding stratified estimator is denoted by $T(m_1, \ldots, m_M)$, then, letting

$$D := \frac{\sum_{j=1}^M a_j \sigma_j}{n}, \tag{5.68}$$

we have

$$V\Big(T(m_1, \ldots, m_M)\Big) \geq V\left(T\left(\frac{a_1\sigma_1}{D}, \ldots, \frac{a_M\sigma_M}{D}\right)\right). \tag{5.69}$$

Proof. The proof is a consequence of the Cauchy-Schwartz inequality, which, for the Euclidean scalar product on $\mathbb{R}^M$, reads

$$\forall_{x,y\in\mathbb{R}^M} \quad \left(\sum_{i=1}^M x_i y_i\right)^2 = \langle x, y\rangle^2 \leq \|x\|_2^2 \, \|y\|_2^2 = \left(\sum_{i=1}^M x_i^2\right)\left(\sum_{i=1}^M y_i^2\right). \tag{5.70}$$

The right-hand side of the inequality in (5.69) is

$$\begin{aligned} V\left(T\left(\frac{a_1\sigma_1}{D}, \ldots, \frac{a_M\sigma_M}{D}\right)\right) &\overset{(5.63),(5.62)}{=} \sum_{i=1}^M a_i^2 \frac{\sigma_i^2}{n_i} \overset{(5.67),(5.68)}{=} \sum_{i=1}^M a_i^2 \frac{\sigma_i^2 \, D}{a_i \sigma_i} \\ &= \sum_{i=1}^M a_i \, \sigma_i \, D \overset{(5.68)}{=} n \, D^2. \end{aligned} \tag{5.71}$$

We now apply the Cauchy-Schwartz inequality with $x_i = \sqrt{m_i}$ and $y_i = a_i\sigma_i/\sqrt{m_i}$ to obtain

$$\begin{aligned} n\,D^2 &\overset{(5.68)}{=} \frac{1}{n}\left(\sum_{i=1}^{M} a_i\,\sigma_i\right)^2 \overset{(5.70)}{\leq} \frac{1}{n}\left(\sum_{i=1}^{M} m_i\right)\left(\sum_{i=1}^{M}\frac{a_i^2\,\sigma_i^2}{m_i}\right) \\ &\overset{(5.63),(5.62)}{=} V\Big(T(m_1,\ldots,m_M)\Big). \end{aligned} \tag{5.72}$$

Combining (5.71) and (5.72) proves (5.69). ■

Remark 5.16. While interesting from a theoretical point of view, the applicability of Th. 5.15 is limited, as the values σ_i are usually not known. Also, the values n_i defined in (5.67) are usually not integers; but one can obtain versions of Th. 5.15 that account for rounding.

Example 5.17. The problem of using Monte Carlo to compute the integral $\int_0^1 f(x)\,\mathrm{d}x$ with f given by (5.21) was previously considered in Ex. 5.3 and Ex. 5.9. We now want to apply stratified sampling to this situation. For the strata, we use $S_i := [\frac{i-1}{M}, \frac{i}{M}[$, $M \in \mathbb{N}$, $i = 1,\ldots,M$, which yield a decomposition of $\mathcal{S} := [0,1]$. If $Y := U : \Omega \longrightarrow [0,1]$ is uniformly distributed, then $a_i = 1/M$ for each $i = 1,\ldots,M$. Given $n, n_1, \ldots, n_M \in \mathbb{N}$ such that $n = n_1 + \cdots + n_M$, each $Y_k^{(i)}$, $i \in \{1,\ldots,M\}$, $k \in \{1,\ldots,n_i\}$ is uniformly distributed on S_i. Proportional allocation means using $n_i = na_i = n/M$, which is possible as long as n is a multiple of M.

In this academic example, one is actually able to compute the σ_i as defined in (5.66), since the antiderivative F of f is available (cf. (5.28)). However, even for this simple example, the computations quickly become involved an unattractive. Still, for illustration purposes, let us at least compute the variance reduction achieved for $M = 2$ and proportional allocation. Using the mentioned antiderivative F of f as stated in (5.28), we obtain

$$E_1(f\circ U) = \int_0^{\frac{1}{2}} f(x)\,\mathrm{d}x \;= F(1/2) - F(0) \approx 1.9132, \tag{5.73a}$$

$$E_2(f\circ U) = \int_{\frac{1}{2}}^{1} f(x)\,\mathrm{d}x \;= F(1) - F(1/2) \approx 1.2284, \tag{5.73b}$$

and, thus,

$$\frac{1}{n}\sum_{i=1}^{2} a_i\left(\frac{E_i(f\circ Y)}{a_i} - E(f\circ Y)\right)^2 \overset{(5.29)}{\approx} \frac{(0.5\cdot 1.91-\pi)^2 + (0.5\cdot 1.23-\pi)^2}{2n}$$

$$\approx \frac{0.469}{n}, \tag{5.73c}$$

$$V\left(\frac{1}{n}\sum_{i=1}^{n} f(Y_i)\right) \overset{(5.32)}{\approx} \frac{0.797}{n}, \tag{5.73d}$$

$$\frac{V(T)}{V\left(\frac{1}{n}\sum_{i=1}^{n} f(Y_i)\right)} \approx \frac{0.797-0.469}{0.797} = \frac{0.328}{0.797} \approx 0.412, \tag{5.73e}$$

such that $M = 2$ and proportional allocation does already achieve a noticeable variance reduction in this case.

Example 5.18. Let $(W_t)_{t\geq 0}$ denote a 1-dimensional standard Brownian motion with drift $\alpha \in \mathbb{R}$ and variance $\sigma^2 > 0$. Motivated by the Bond price formula (4.44), let us consider the problem of estimating

$$E\left(\exp\left(-\int_0^T W_u \,\mathrm{d}u\right)\right), \quad T > 0, \tag{5.74}$$

even though $(W_t)_{t\geq 0}$ does not constitute a realistic short rate model. For the measurable space $(\mathcal{S}, \mathcal{B})$, we use $\mathcal{S} := C[0,T]$, the set of real-valued continuous functions, with $\mathcal{B}$ being the Borel σ-algebra on $\mathcal{S}$ induced by the topology of uniform convergence (i.e. the $\|\cdot\|_\infty$-norm topology).

To construct the strata, we fix $M \in \mathbb{N}$, $M \geq 2$, and decompose $\mathbb{R}$ into M intervals of positvie length. More precisely, choose numbers

$$-\infty =: s_0 < s_1 < \cdots < s_{M-1} < s_M := \infty \tag{5.75a}$$

and set

$$\underset{i=1,\ldots,M}{\forall} \quad I_i :=]s_{i-1}, s_i]. \tag{5.75b}$$

Then the strata are

$$\underset{i=1,\ldots,M}{\forall} \quad S_i := \big\{g \in C[0,T] : \, g(0) = 0 \ \wedge \ g(T) \in I_i\big\}. \tag{5.75c}$$

We choose a random variable $Y : \Omega \longrightarrow \mathcal{S}$ such that $Y \sim \mathrm{BM}(\alpha, \sigma^2)$, which is supposed to mean that letting

$$\underset{t\in[0,T]}{\forall} \quad W_t : \Omega \longrightarrow \mathbb{R}, \quad W_t(\omega) := Y(\omega)(t), \tag{5.76}$$

yields a 1-dimensional standard Brownian motion $(W_t)_{t\geq 0}$ with drift α and variance σ^2. That such a Y exists is shown, e.g., in [Bau02, Th. 23.4] (see [Bau02, Sec. 38, Sec. 39, Cor. 40.4]) – there even exists a probability measure μ on $(\mathcal{S}, \mathcal{B})$ such that one can choose $(\Omega, \mathcal{A}, P) = (\mathcal{S}, \mathcal{B}, \mu)$ and $Y := \mathrm{Id}$. If we let

$$f : \mathcal{S} \longrightarrow \mathbb{R}, \quad f(g) := \exp\left(-\int_0^T g(u)\,\mathrm{d}u\right), \tag{5.77}$$

then

$$E\left(\exp\left(-\int_0^T W_u \,\mathrm{d}u\right)\right) = \int_\Omega \exp\left(-\int_0^T Y(\omega)(u)\,\mathrm{d}u\right) \mathrm{d}\omega \ = E(f \circ Y), \tag{5.78}$$

i.e. the setting is suitable for stratified sampling.

From Rem. 4.6, we know W_T is $N\big(T\alpha, T\sigma^2\big)$-distributed, i.e.

$$\begin{aligned} a_i &= P\{Y \in S_i\} = P\big\{\omega \in \Omega : Y(\omega)(T) = W_T(\omega) \in I_i\big\} \\ \underset{i=1,\ldots,M}{\forall} \qquad &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{s_{i-1}}^{s_i} e^{-\frac{(x-T\alpha)^2}{2\,T\sigma^2}}\,\mathrm{d}x \;=\; \frac{1}{\sqrt{2\pi}} \int_{\frac{s_{i-1}-T\alpha}{\sigma\sqrt{T}}}^{\frac{s_i-T\alpha}{\sigma\sqrt{T}}} e^{-\xi^2/2}\,\mathrm{d}\xi \;>\; 0, \end{aligned} \tag{5.79}$$

as required for stratified sampling.

To obtain a stratified estimate for (5.74), we can now proceed as follows: Choose $n, K \in \mathbb{N}$ such that $n = KM$. The idea is to make all $a_i = 1/M$ and then use proportional allocation. To this end, let

$$g : \overline{\mathbb{R}} \longrightarrow \overline{\mathbb{R}}, \quad g(x) := T\alpha + \sigma\sqrt{T}\,x, \tag{5.80a}$$

$$g^{-1} : \overline{\mathbb{R}} \longrightarrow \overline{\mathbb{R}}, \quad g^{-1}(x) = \frac{x - T\alpha}{\sigma\sqrt{T}}, \tag{5.80b}$$

$$\underset{i=1,\ldots,M}{\forall} \quad \tau_i := g^{-1}(s_i) = \frac{s_i - T\alpha}{\sigma\sqrt{T}}. \tag{5.80c}$$

If $\Phi : \overline{\mathbb{R}} \longrightarrow [0,1]$, $\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\xi^2/2}\,\mathrm{d}\xi$ denotes the CDF of the standard normal distribution, then

$$\begin{aligned} &\left(\underset{i=1,\ldots,M}{\forall} \quad \frac{1}{M} = a_i \overset{(5.79)}{=} \Phi(\tau_i) - \Phi(\tau_{i-1}) \right) \\ \Rightarrow \quad &\left(\underset{i=1,\ldots,M}{\forall} \quad \tau_i = \Phi^{-1}\left(\frac{i}{M}\right) \right). \end{aligned} \tag{5.81}$$

In practise, one will need code to evaluate Φ and Φ^{-1}. Then one obtains the τ_i from (5.81), and one can employ the inverse transform method as in Ex. 3.7 to obtain i.i.d. random numbers $\iota_k^{(i)} \in I_i$, $k = 1, \ldots, n_i = na_i = K$, that are distributed according to W_T under the condition I_i, i.e. according to $N\big(T\alpha, T\sigma^2\big)|I_i$: If the random variable $U : \Omega \longrightarrow [0,1]$ is uniformly distributed, then, analogously to (3.14) and (3.15), one defines

$$\begin{aligned} \underset{i=1,\ldots,M}{\forall} \qquad V_i : \Omega \longrightarrow [\Phi(\tau_{i-1}), \Phi(\tau_i)], \qquad V_i &:= \Phi(\tau_{i-1}) + \big(\Phi(\tau_i) - \Phi(\tau_{i-1})\big)\,U \\ &= \frac{i-1}{M} + \frac{U}{M}, \end{aligned} \tag{5.82a}$$

$$\underset{i=1,\ldots,M}{\forall} \qquad W_i : \Omega \longrightarrow \overline{\mathbb{R}}, \qquad W_i := \Phi^{-1} \circ V_i. \tag{5.82b}$$

Then,

$$\begin{aligned} U \sim \mathrm{Unif}[0,1] \qquad &\overset{\text{Ex. 3.7}}{\Rightarrow} \qquad W_i \sim N(0,1)\,|\,]\tau_{i-1}, \tau_i] \\ &\overset{\text{Lem. 3.13}}{\Rightarrow} \qquad g(W_i) \sim N\big(T\alpha, T\sigma^2\big)|I_i. \end{aligned} \tag{5.83}$$

Ideally, for the stratified estimator

$$\sum_{i=1}^{M} \frac{a_i}{n_i} \sum_{k=1}^{n_i} f(Y_k^{(i)}) \overset{n_i = n a_i = K}{=} \sum_{i=1}^{M} \frac{1}{n} \sum_{k=1}^{K} f(Y_k^{(i)}) \tag{5.84}$$

according to (5.57) and (5.58), one would like, for each $\iota_k^{(i)} \in I_i$, to sample i.i.d. $Y_k^{(i)} \in S_i$ such that $Y_k^{(i)} \sim \mathrm{BM}(\alpha, \sigma^2) \,|\, \{W_T = \iota_k^{(i)}\}$ and then compute $f(Y_k^{(i)})$ with f as defined in (5.77). However, an exact evaluation of $f(Y_k^{(i)})$ is not at hand, so one uses a quadrature formula, for example, by means of the composite rectangle rule

$$f(Y_k^{(i)}) \approx \exp\left(-h \sum_{\nu=0}^{N-1} Y_k^{(i)}(\nu\, h) \right), \tag{5.85}$$

where $h = T/N$, $N \in \mathbb{N}$, is some step size; and the values $Y_k^{(i)}(\nu\, h)$ are readily obtained using the Brownian bridge construction of Sec. 4.1.4.

5.4 Importance Sampling

Importance sampling is another technique for reducing variance when estimating the expected value of random variables. Here the general setting is the same as previously in Sec. 5.3. Given a probability space $(\Omega, \mathcal{A}, P)$, a measurable space $(\mathcal{S}, \mathcal{B})$, a random variable $Y : \Omega \longrightarrow \mathcal{S}$, and a measurable function $f : \mathcal{S} \longrightarrow \mathbb{R}$, the goal is to estimate $E(f \circ Y) = \int_{\mathcal{S}} f(x)\, \mathrm{d}P_Y(x)$ and to improve over the simple estimator $\frac{1}{n} \sum_{i=1}^{n} f(Y_i)$.

The idea is to change the distribution of Y (more precisely, to replace Y by some suitable random variable X) such that more weight is given to "important" outcomes. We will see that the resulting importance sampling estimator is unbiased, but examples will show one has to use care – while an apt choice of importance sampling estimator can, depending on the situation, reduce the variance by orders of magnitude, a bad choice might also increase the variance by orders of magnitude – it can even make the variance infinite (see Ex. 5.22(a) below).

The idea is to make use of the Radon-Nikodym Th. A.72: According to Th. A.72 and Th. A.73, if ν is another probability measure on $(\mathcal{S}, \mathcal{B})$ such that $\nu \ll P_Y$ (i.e. ν is absolutely continuous with respect to P_Y, i.e. $P_Y(B) = 0$ implies $\nu(B) = 0$), then ν has a unique density g, called the Radon-Nikodym derivative of ν with respect to P_Y,

$$g : \mathcal{S} \longrightarrow \mathbb{R}_0^+, \quad g = \frac{\mathrm{d}\nu}{\mathrm{d}P_Y}, \tag{5.86}$$

cf. Def. A.74.

Definition 5.19. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\mathcal{S}, \mathcal{B})$ a measurable space, and $Y, X : \Omega \longrightarrow \mathcal{S}$ a random variables, satisfying $\nu := P_X \ll P_Y$. If g is as in (5.86), namely the Radon-Nikodym derivative of P_X with respect to P_Y, i.e. $P_X = g P_Y$, and $X_1, X_2, \ldots$ are i.i.d. copies of X, then

$$\underset{n \in \mathbb{N}}{\forall} \quad \bar{Y}_{n,g} := \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} \tag{5.87}$$

is called the *importance sampling estimator* of $E(f \circ Y)$ based on g.

Remark 5.20. Note that (5.87) is well-defined P-almost everywhere, since $g(X_i) > 0$ P-almost everywhere:

$$P\{g \circ X_i = 0\} = P\big\{X_i \in g^{-1}\{0\}\big\} = P_X\big(g^{-1}\{0\}\big) = \int_{g^{-1}\{0\}} g \,\mathrm{d}P_Y = 0. \tag{5.88}$$

Remark 5.21. Consider the situation of Def. 5.19.

(a) One has

$$\begin{aligned} E(f \circ Y) &= \int_{\mathcal{S}} f(x) \,\mathrm{d}P_Y(x) = \int_{\mathcal{S}} \frac{f(x)}{g(x)}\, g(x) \,\mathrm{d}P_Y(x) = \int_{\mathcal{S}} \frac{f(x)}{g(x)} \,\mathrm{d}P_X(x) \\ &= E\left(\frac{f}{g} \circ X\right) \end{aligned} \tag{5.89}$$

and

$$\underset{n\in\mathbb{N}}{\forall} \quad E(\bar{Y}_{n,g}) = E\left(\frac{f}{g} \circ X\right) = E(f \circ Y), \tag{5.90}$$

showing the importance sampling estimator is unbiased.

(b) The variance of the importance sampling estimator (5.87) is, for $n \in \mathbb{N}$,

$$\begin{aligned} V(\bar{Y}_{n,g}) &= \frac{1}{n} V\left(\frac{f(X)}{g(X)}\right) = \frac{1}{n}\left(E\left(\frac{f^2}{g^2} \circ X\right) - \big(E(f \circ Y)\big)^2\right) \\ &= \frac{1}{n}\left(\int_{\mathcal{S}} \frac{f(x)^2}{g(x)^2} \,\mathrm{d}P_X(x) - \big(E(f \circ Y)\big)^2\right) \\ &= \frac{1}{n}\left(\int_{\mathcal{S}} \frac{f(x)^2}{g(x)^2}\, g(x) \,\mathrm{d}P_Y(x) - \big(E(f \circ Y)\big)^2\right) \\ &= \frac{1}{n}\left(\int_{\mathcal{S}} \frac{f(x)^2}{g(x)} \,\mathrm{d}P_Y(x) - \big(E(f \circ Y)\big)^2\right). \end{aligned} \tag{5.91}$$

Example 5.22. We return once more to our standard example of using Monte Carlo to compute the integral $\int_0^1 f(x) \,\mathrm{d}x$ with f given by (5.21), previously considered in Ex. 5.3, Ex. 5.9, and Ex. 5.17, this time using importance sampling. We will base the importance sampling estimator on different functions g. First, for the convenience of the reader, f is restated:

$$f : [0,1] \longrightarrow \mathbb{R}, \quad f(x) := 4\sqrt{1 - x^2}.$$

We also recall that with uniformly distributed $Y := U : \Omega \longrightarrow [0,1]$, it is $\int_0^1 f(x) \,\mathrm{d}x = E(f \circ Y)$.

(a) As mentioned above, a poor choice of g can actually increase the variance, and this first example is designed to illustrate a worst-case scenario: Consider

$$g : [0,1] \longrightarrow \mathbb{R}, \quad g(x) := 2x. \tag{5.92}$$

Letting $X : \Omega \longrightarrow [0,1]$ be distributed according to g (i.e. $P_X = gP_Y = g\lambda_1$) with i.i.d. copies $X_1, X_2, \ldots$ (which could be easily generated from i.i.d. copies of $Y = U$ using the inverse transform method), the importance sampling estimator based on g is

$$\underset{n\in\mathbb{N}}{\forall} \quad \bar{Y}_{n,g} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} = \frac{1}{n}\sum_{i=1}^{n} \frac{4\sqrt{1-X_i^2}}{2X_i}. \tag{5.93}$$

Bearing in mind the computation of Rem. 5.21(b), we compute the variance

$$\begin{aligned} V\left(\frac{f}{g}\circ X\right) &= \int_0^1 \frac{f(x)^2}{g(x)}\,\mathrm{d}P_Y(x) \;-\; \big(E(f\circ Y)\big)^2 \overset{(5.29)}{=} \int_0^1 \frac{16\,(1-x^2)}{2x}\,\mathrm{d}x \;-\pi^2 \\ &= 8\int_0^1 \frac{1}{x}\,\mathrm{d}x \;- 8\int_0^1 x\,\mathrm{d}x \;-\pi^2 = \infty, \end{aligned} \tag{5.94}$$

explaining what was meant by g leading to a worst-case scenario.

(b) A better choice of g turns out to be

$$g:\,[0,1] \longrightarrow \mathbb{R}, \quad g(x) := \frac{4-2x}{3}. \tag{5.95}$$

Once again, letting $X : \Omega \longrightarrow [0,1]$ be distributed according to g with i.i.d. copies $X_1, X_2, \ldots$ (which can still be easily generated from i.i.d. copies of $Y = U$ using the inverse transform method), the importance sampling estimator based on g is

$$\underset{n\in\mathbb{N}}{\forall} \quad \bar{Y}_{n,g} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} = \frac{1}{n}\sum_{i=1}^{n} \frac{12\sqrt{1-X_i^2}}{4-2X_i}. \tag{5.96}$$

To compare variances, we recall from (5.32) that $V(f\circ Y) \approx 0.797$ and compute

$$\begin{aligned} V\left(\frac{f}{g}\circ X\right) &= \int_0^1 \frac{24\,(1-x^2)}{2-x}\,\mathrm{d}x \;-\pi^2 \\ &= 24\left[-6+2x+\frac{x^2}{2}+3\ln(x-2)\right]_0^1 - \pi^2 \approx 0.224, \end{aligned} \tag{5.97}$$

showing that this g reduces variance by more than a factor 3.5 when compared with the simple estimator.

Now that we have seen an example of a g that failed when used in an importance sampling estimator and one example of a g that worked, let us look more closely at the general variance formula (5.91) to better understand what makes the difference between a good and a bad g:

Remark 5.23. To make the variance $V(\bar{Y}_{n,g})$ small, according to (5.91), one has to make the integral in the last line of (5.91) small. Indeed, for $E(f\circ Y) \neq 0$, it is minimized by the choice

$$g := \frac{f}{E(f\circ Y)} \quad \Rightarrow \quad V(\bar{Y}_{n,g}) = \frac{1}{n}\left(E(f\circ Y)\int_{\mathcal{S}} \frac{f(x)^2}{f(x)}\,\mathrm{d}P_Y(x) \;-\; \big(E(f\circ Y)\big)^2\right) = 0. \tag{5.98}$$

Alas, this choice of g is not available if $E(f \circ Y)$ is unknown.

Still, a good importance sampling estimator will have to have g large where f is large (this is, actually, where the name "importance sampling" comes from – a good g gives more weight to the "important" values of f). As seen from Ex. 5.22(a), it is crucial not to have $g \ll f^2$ anywhere, whereas $g \gg f^2$ for some values would not make the variance large. Bearing this in mind, one should roughly choose g proportional to f.

6 Simulation of SDE

6.1 Setting

If nothing else is stated, all random variables are assumed to be defined on the probability space $(\Omega, \mathcal{A}, P)$, which, for technical reasons, we also assume to be complete. Moreover, let $T > 0$.

The goal is to numerically simulate solutions to the SDE

$$\mathrm{d}X_t = a(t, X_t)\,\mathrm{d}t + b(t, X_t)\,\mathrm{d}W_t \tag{6.1a}$$

with initial condition

$$X_0 = X_{\text{init}}, \tag{6.1b}$$

where, in the most general situation, the stochastic process $(X_t)_{t\in[0,T]}$ is $\mathbb{R}^d$-valued, $(W_t)_{t\in[0,T]}$ is an m-dimensional standard Brownian motion with drift vector $\alpha = 0$ and covariance matrix $\Sigma = \mathrm{Id}$, the given random variable $X_{\text{init}} : \Omega \longrightarrow \mathbb{R}^d$ is independent of the family $(W_t)_{t\in[0,T]}$, and the maps

$$a : [0,T] \times \mathbb{R}^d \longrightarrow \mathbb{R}^d, \quad b : [0,T] \times \mathbb{R}^d \longrightarrow \mathbb{R}^{d\times m} \tag{6.2}$$

are measurable. The first summand on the right-hand side of (6.1a) is often referred to as the *drift term*, whereas the second summand is called the *diffusion term*.

Under suitable hypotheses, it is known that the SDE with initial condition (6.1) admits a unique strong solution. In preparation to state this result, we need some definitions:

Definition 6.1. An $\mathbb{R}^d$-valued stochastic process $(X_t)_{t\in[0,T]}$, $d \in \mathbb{N}$, is called an *Itô process* if, and only if, it can be represented as

$$\underset{t\in[0,T]}{\forall} \quad X_t = X_0 + \int_0^t a_u\,\mathrm{d}u + \int_0^t b_u\,\mathrm{d}W_u\,, \tag{6.3}$$

typically written in the shorthand form

$$\underset{t\in[0,T]}{\forall} \quad \mathrm{d}X_t = a_t\,\mathrm{d}t + b_t\,\mathrm{d}W_t\,, \tag{6.4}$$

where $(W_t)_{t\in[0,T]}$ denotes a k-dimensional standard Brownian motion with drift vector $\alpha = 0$ and covariance matrix $\Sigma = \mathrm{Id}$, $k \in \mathbb{N}$, and the stochastic processes

$$a_t : \Omega \longrightarrow \mathbb{R}^d, \quad b_t : \Omega \longrightarrow \mathbb{R}^{d\times k} \tag{6.5}$$

satisfy

$$P\left\{\omega \in \Omega : \int_0^T \left\|a_t(\omega)\right\| \mathrm{d}t \, < \infty\right\} = 1 \tag{6.6a}$$

and

$$P\left\{\omega \in \Omega : \int_0^T \left\|b_t(\omega)\right\|^2 \mathrm{d}t \, < \infty\right\} = 1, \tag{6.6b}$$

respectively. Note that conditions (6.6) do not depend on which particular norms on $\mathbb{R}^d$ and $\mathbb{R}^{d\times k}$ are chosen.

Definition 6.2. An Itô process $(X_t)_{t\in[0,T]}$ is called a *strong solution* to (6.1) if, and only if,

$$X_0 = X_{\text{init}} \quad \text{almost surely,} \tag{6.7a}$$

$$\underset{t\in[0,T]}{\forall} \quad X_t = X_0 + \int_0^t a(u, X_u)\,\mathrm{d}u \, + \int_0^t b(u, X_u)\,\mathrm{d}W_u\,, \tag{6.7b}$$

$$\underset{\omega\in\Omega}{\forall} \quad \text{the path } t \mapsto X_t(\omega) \text{ is continuous,} \tag{6.7c}$$

and

$$\underset{t\in[0,T]}{\forall} \quad X_t \text{ is } \mathcal{F}_t\text{-measurable,} \tag{6.7d}$$

where the increasing family of σ-algebras $(\mathcal{F}_t)_{t\in[0,T]}$, $\mathcal{F}_t \subseteq \mathcal{A}$, is constructed as follows:

$$\underset{t\in[0,T]}{\forall} \quad \mathcal{G}_t := \sigma\big(X_{\text{init}}, (W_s)_{s\in[0,t]}\big), \tag{6.8a}$$

$$\mathcal{G}_\infty := \sigma\left(\bigcup_{t\in[0,T]} \mathcal{G}_t\right), \tag{6.8b}$$

$$\mathcal{N} := \left\{N \in \mathcal{A} : \underset{G\in\mathcal{G}_\infty}{\exists} \big(N \subseteq G \,\wedge\, P(G) = 0\big)\right\}, \tag{6.8c}$$

$$\underset{t\in[0,T]}{\forall} \quad \mathcal{F}_t := \sigma(\mathcal{N} \cup \mathcal{G}_t). \tag{6.8d}$$

Condition (6.7b) is usually stated by saying the process $(X_t)_{t\in[0,T]}$ is *adapted* to the *filtration* $(\mathcal{F}_t)_{t\in[0,T]}$.

Theorem 6.3. *The SDE with initial condition* (6.1) *has a strong solution if the following conditions are satisfied:*

(a) *X_{init} is independent of $(W_t)_{t\in[0,T]}$ and satisfies $E\big(\|X_{\text{init}}\|^2\big) < \infty$.*

(b) *Both a and b are measurable and globally Lipschitz continuous with respect to x. More precisely, the Lipschitz continuity means a, b satisfy*

$$\underset{L\geq 0}{\exists} \quad \underset{\substack{t\in[0,T],\\ x,y\in\mathbb{R}^d}}{\forall} \quad \big\|a(t,x) - a(t,y)\big\| + \big\|b(t,x) - b(t,y)\big\| \leq L\,\big\|x - y\big\|. \tag{6.9}$$

(c) *a and b satisfy a linear growth condition with respect to x, which can be expressed in the form*

$$\exists_{K\geq 0} \quad \forall_{\substack{t\in[0,T],\\ x\in\mathbb{R}^d}} \quad \big\|a(t,x)\big\|^2 + \big\|b(t,x)\big\|^2 \leq K\big(1+\|x\|^2\big). \tag{6.10}$$

Moreover, this strong solution $(X_t)_{t\in[0,T]}$ satisfies

$$\forall_{t\in[0,T]} \quad E\big(\|X_t\|^2\big) \leq C\left(1+E\big(\|X_{\text{init}}\|^2\big)\right)e^{C\,t} < \infty \tag{6.11}$$

with a constant $C \geq 0$ depending only on L, K, and T; and the strong solution $(X_t)_{t\in[0,T]}$ is unique in the sense that, if $(Y_t)_{t\in[0,T]}$ is any strong solution to (6.1), then

$$P\big\{\omega\in\Omega:\ X_t(\omega) = Y_t(\omega) \text{ for each } t\in[0,T]\big\} = 1. \tag{6.12}$$

Proof. We just give the idea of the proof and refer to the literature for details. The uniqueness proof is based on Gronwall's inequality. Regarding existence, a standard proof for the existence of solutions to deterministic ordinary differential equations is to use the integral form of the equation to define a sequence of approximations to the solution, and then obtain a solution as a limit of an approximating sequence. In the SDE case, one can proceed analogously, defining an approximating sequence by

$$\forall_{t\in[0,T]} \quad X_t^{(0)} := X_{\text{init}}, \tag{6.13a}$$

$$\forall_{\substack{t\in[0,T],\\ k\in\mathbb{N}_0}} \quad X_t^{(k+1)} := X_{\text{init}} + \int_0^t a(u, X_u^{(k)})\,\mathrm{d}u + \int_0^t b(u, X_u^{(k)})\,\mathrm{d}W_u\,. \tag{6.13b}$$

For details of the proof see, e.g., [Øk03, Th. 5.2.1] or [KS98, Ch. 5, Ths. 2.5, 2.9] (proof partially as problems, solutions included). ■

6.2 The Euler Scheme

Definition 6.4. In the situation of (6.1), given discrete times $0 = t_0 < t_1 < \cdots < t_N \leq T$, $N\in\mathbb{N}$, and an i.i.d. family $(Z_1,\ldots,Z_N)$ of m-dimensional random vectors, $N(0,\mathrm{Id})$-distributed, the recursion

$$\hat{X}_0 := X_{\text{init}}, \tag{6.14a}$$

$$\forall_{i=0,\ldots,N-1} \quad \hat{X}_{t_{i+1}} := \hat{X}_{t_i} + a\big(t_i,\hat{X}_{t_i}\big)\,(t_{i+1}-t_i) + b\big(t_i,\hat{X}_{t_i}\big)\,Z_{i+1}\,\sqrt{t_{i+1}-t_i}, \tag{6.14b}$$

is called an *Euler scheme* for the SDE (6.1). It defines the discrete stochastic process $(\hat{X}_0,\ldots,\hat{X}_{t_N})$, supposed to approximate the solution to (6.1).

Remark 6.5. For a fixed time step size $h>0$, we have $t_i = ih$ and, writing $\hat{X}_i$ instead of $\hat{X}_{t_i}$, (6.14b) simplifies to

$$\forall_{i=0,\ldots,N-1} \quad \hat{X}_{i+1} := \hat{X}_i + a\big(ih,\hat{X}_i\big)\,h + b\big(ih,\hat{X}_i\big)\,Z_{i+1}\,\sqrt{h}. \tag{6.15}$$

—

As long as a and b are easy to evaluate, implementation of the Euler scheme is straightforward, cf. the recursions in Sections 4.1.2 and 4.1.5. We will now address the questions of how to improve on the Euler scheme and in what sense the process $(\hat{X}_0, \ldots, \hat{X}_{t_N})$ approximates the strong solution $(X_t)_{t\in[0,T]}$ to (6.1).

6.3 Refinement of the Euler Scheme

Before discussing quantities for gauging the accuracy of discretizations in Sec. 6.4 below, we present a heuristic derivation of a discretization scheme due to Milstein [Mil75]. For the sake of simplicity, we will restrict ourselves to autonomous SDE (where a, b do not explicitly depend on the time variable t) and to equidistant time steps h as in (6.15).

6.3.1 1-Dimensional Case

We consider $d = m = 1$, in particular, the solution $(X_t)_{t\in[0,T]}$ to (6.1) is $\mathbb{R}$-valued.

We start with the observation that the approximation of the drift term in (6.15) is $O(h)$, whereas the approximation of the diffusion term is $O(\sqrt{h})$ (of course, this is not entirely correct, since Z_{i+1} is not a function mapping real numbers to real numbers, but such things happen frequently in heuristic arguments). This yields the idea of trying to improve the diffusion approximation to make it $O(h)$ as well.

One obtains the Euler scheme (6.15) from (6.7b) by approximating the integrals using rectangle rules,

$$\int_t^{t+h} a(X_u)\,\mathrm{d}u \;\approx a(X_t)\,h, \tag{6.16a}$$

$$\int_t^{t+h} b(X_u)\,\mathrm{d}W_u \;\approx b(X_t)\,(W_{t+h} - W_t), \tag{6.16b}$$

approximating the integrands over $[t, t+h]$ by their respective values at the lower bound t. The idea is now to try to improve on the approximation in (6.16b). To this end, it is an exercise to assume b to be twice differentiable and to use Itô's formula (C.3) to show the process $(b(X_t))_{t\in[0,T]}$ satisfies the SDE

$$\mathrm{d}b(X_t) \;= \alpha_b(X_t)\,\mathrm{d}t \;+ \sigma_b(X_t)\,\mathrm{d}W_t\,, \tag{6.17}$$

where

$$\alpha_b(X_t) := b'(X_t)\,a(X_t) + \frac{1}{2}\,b''(X_t)\,b^2(X_t), \tag{6.18a}$$

$$\sigma_b(X_t) := b'(X_t)\,b(X_t). \tag{6.18b}$$

Now (6.17) is used together with the Euler scheme to approximate $b(X_u)$ for $t \le u \le t+h$:

$$b(X_u) \approx b(X_t) + \alpha_b(X_t)\,(u-t) + \sigma_b(X_t)\,(W_u - W_t). \tag{6.19}$$

Applying the same order analysis already used at the beginning of this section to (6.19), one observes $O(u-t)$ for the drift term and $O(\sqrt{u-t})$ for the diffusion term (in the same, not entirely correct, way as in the situation of the second paragraph of this section above). Dropping the higher order drift term, one obtains the approximation

$$\begin{aligned} \underset{u\in[t,t+h]}{\forall} \qquad b(X_u) &\approx b(X_t) + \sigma_b(X_t)\,(W_u - W_t) \\ &= b(X_t) + b'(X_t)\,b(X_t)\,(W_u - W_t). \end{aligned} \tag{6.20}$$

In (6.16b), the approximation $b(X_u) \equiv b(X_t)$ was used. Using (6.20) instead, provides

$$\begin{aligned} \int_t^{t+h} b(X_u)\,\mathrm{d}W_u \;&\approx \int_t^{t+h} \Big(b(X_t) + b'(X_t)\,b(X_t)\,(W_u - W_t)\Big)\,\mathrm{d}W_u \\ &= b(X_t)\,(W_{t+h} - W_t) + b'(X_t)\,b(X_t) \int_t^{t+h} (W_u - W_t)\,\mathrm{d}W_u\,. \end{aligned} \tag{6.21}$$

To use this in a refined Euler recursion scheme, we need to work on the last integral in (6.21) a bit further:

$$\begin{aligned} \int_t^{t+h} (W_u - W_t)\,\mathrm{d}W_u \;&= \int_t^{t+h} W_u\,\mathrm{d}W_u \;- W_t \int_t^{t+h} \mathrm{d}W_u \\ &= \int_0^{t+h} W_u\,\mathrm{d}W_u \;- \int_0^{t} W_u\,\mathrm{d}W_u \;- W_t\,\big(W_{t+h} - W_t\big) \\ &= Y_{t+h} - Y_t - W_t\,\big(W_{t+h} - W_t\big) \end{aligned} \tag{6.22}$$

with random variables

$$Y_t := \int_0^t W_u\,\mathrm{d}W_u\,. \tag{6.23}$$

Noting that the stochastic process $(Y_t)_{t\in[0,T]}$ satisfies the initial condition $Y_0 \equiv 0$ and the SDE

$$\mathrm{d}Y_t \;= W_t\,\mathrm{d}W_t\,, \tag{6.24}$$

Itô's formula (C.3) shows (exercise)

$$Y_t = \frac{W_t^2}{2} - \frac{t}{2}. \tag{6.25}$$

Now this expression is plugged back into (6.22):

$$\begin{aligned} \int_t^{t+h} (W_u - W_t)\,\mathrm{d}W_u \;&= \frac{W_{t+h}^2}{2} - \frac{t+h}{2} - \frac{W_t^2}{2} + \frac{t}{2} - W_t\,\big(W_{t+h} - W_t\big) \\ &= \frac{\big(W_{t+h} - W_t\big)^2}{2} - \frac{h}{2}. \end{aligned} \tag{6.26}$$

Using (6.26) to replace the integral in (6.21), we obtain the Milstein approximation (i.e. the refined Euler approximation)

$$\begin{aligned} X_{t+h} &= X_t + \int_t^{t+h} a(X_u)\,\mathrm{d}u \;+ \int_t^{t+h} b(X_u)\,\mathrm{d}W_u \\ &\approx X_t + a(X_t)\,h + b(X_t)\,(W_{t+h} - W_t) + \frac{1}{2}\,b'(X_t)\,b(X_t)\,\big((W_{t+h} - W_t)^2 - h\big). \end{aligned} \tag{6.27}$$

Using an i.i.d. family $(Z_1, \dots, Z_N)$ random variables, $N(0,1)$-distributed, such that $W_{t_i+h} - W_{t_i} = Z_{i+1}\sqrt{h}$, we obtain the following recursion, sometimes called the *Milstein scheme*:

$$\hat{X}_0 := X_{\text{init}}, \tag{6.28a}$$

$$\begin{aligned} \hat{X}_{i+1} := \hat{X}_i + a\big(\hat{X}_i\big)\, h + b\big(\hat{X}_i\big)\, Z_{i+1}\sqrt{h} \\ + \frac{1}{2}\, b'(\hat{X}_i)\, b(\hat{X}_i)\, (Z_{i+1}^2 - 1)\, h. \end{aligned} \qquad \underset{i=0,\dots,N-1}{\forall} \tag{6.28b}$$

Even though we assumed the autonomous situation in the above heuristic derivation of the Milstein scheme, it turns out (6.28) generalizes to the nonautonomous case in the canonical way (see [KP99, Ch. 10, (3.1)]):

$$\hat{X}_0 := X_{\text{init}}, \tag{6.29a}$$

$$\begin{aligned} \hat{X}_{i+1} := \hat{X}_i + a\big(t_i, \hat{X}_i\big)\, h + b\big(t_i, \hat{X}_i\big)\, Z_{i+1}\sqrt{h} \\ + \frac{1}{2}\, b'(t_i, \hat{X}_i)\, b(t_i, \hat{X}_i)\, (Z_{i+1}^2 - 1)\, h, \end{aligned} \qquad \underset{i=0,\dots,N-1}{\forall} \tag{6.29b}$$

where the prime b' now means the partial derivative with respect to the second variable.

As compared with the Euler scheme (6.15), the Milstein scheme contains an additional term that results in both the drift and diffusion term now being $O(h)$.

6.3.2 Multi-Dimensional Case

We now consider the general case of $\mathbb{R}^d$-valued $(X_t)_{t\in[0,T]}$ and $\mathbb{R}^m$-valued $(W_t)_{t\in[0,T]}$, $d, m \in \mathbb{N}$. For $k = 1, \dots, d$ and $l = 1, \dots, m$, let $(X_t)_k$, a_k, and b_{kl} denote the components of the functions X_t, a, and b, respectively. One proceeds analogous to the 1-dimensional case: For the components, one obtains from (6.7b):

$$(X_{t+h})_k = (X_t)_k + \int_t^{t+h} a_k(u, X_u)\,\mathrm{d}u \, + \sum_{l=1}^{m} \int_t^{t+h} b_{kl}(u, X_u)\,\mathrm{d}(W_l)_u\,. \tag{6.30}$$
$$\underset{\substack{t\in[0,T-h],\\ k=1,\dots,d}}{\forall}$$

To compare with the formulas from Sec. 6.3.1, we once again assume a, b to have no explicit time dependence. Then (6.16a) generalizes to

$$\int_t^{t+h} a_k(X_u)\,\mathrm{d}u \, \approx a_k(X_t)\, h \tag{6.31}$$
$$\underset{k=1,\dots,d}{\forall}$$

and, due to Itô's formula in multiple dimensions [Gla04, Th. B.1.1], (6.21) generalizes to

$$\begin{aligned} &\int_t^{t+h} b_{kl}(X_u)\,\mathrm{d}(W_l)_u \\ &\approx b_{kl}(X_t)\,\Big((W_{t+h})_l - (W_t)_l\Big) \\ &\quad + \sum_{\alpha=1}^{d}\sum_{\beta=1}^{m} \frac{\partial b_{kl}}{\partial x_\alpha}(X_t)\, b_{\alpha\beta}(X_t) \int_t^{t+h} \Big((W_u)_\beta - (W_t)_\beta\Big)\,\mathrm{d}(W_l)_u\,. \end{aligned} \tag{6.32}$$
$$\underset{\substack{k=1,\dots,d,\\ l=1,\dots,m}}{\forall}$$

Thus, introducing the abbreviations

$$\underset{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}}{\forall} \qquad (\Delta W)_l^{(i)} := (W_{t_i+h})_l - (W_{t_i})_l, \tag{6.33a}$$

$$\underset{\substack{i=0,\dots,N-1,\\ \beta,l=1,\dots,m}}{\forall} \qquad I_{\beta l}^{(i)} := \int_{t_i}^{t_i+h} \Big((W_u)_\beta - (W_t)_\beta\Big)\,\mathrm{d}(W_l)_u\,, \tag{6.33b}$$

and reinstating possible explicit time depencencies of a and b, the multi-dimensional Milstein scheme can be written as

$$\hat{X}^{(0)} := X_{\text{init}}, \tag{6.34a}$$

$$\begin{aligned}\underset{\substack{i=0,\dots,N-1,\\ k=1,\dots,d}}{\forall} \qquad \hat{X}_k^{(i+1)} &:= \hat{X}_k^{(i)} + a_k\big(t_i,\hat{X}^{(i)}\big)\,h + \sum_{l=1}^{m} b_{kl}\big(t_i,\hat{X}^{(i)}\big)\,(\Delta W)_l^{(i)} \\ &\quad + \sum_{l=1}^{m}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{m} \frac{\partial b_{kl}}{\partial x_\alpha}\big(t_i,\hat{X}^{(i)}\big)\,b_{\alpha\beta}\big(t_i,\hat{X}^{(i)}\big)\,I_{\beta l}^{(i)}.\end{aligned} \tag{6.34b}$$

To make use of (6.34) in practise is, in general, much more difficult than in the 1-dimensional case, where the main problem is the approximation of the stochastic integrals $I_{\beta l}^{(i)}$ for $\beta \neq l$: While, for $\beta = l$, the argument from Sec. 6.3.1 (cf. (6.26)) yields

$$\underset{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}}{\forall} \quad I_{ll}^{(i)} = \int_{t_i}^{t_i+h} \Big((W_u)_l - (W_{t_i})_l\Big)\,\mathrm{d}(W_l)_u = \frac{\big((W_{t_i+h})_l - (W_{t_i})_l\big)^2}{2} - \frac{h}{2}, \tag{6.35}$$

there is no simple formula to evaluate the mixed integrals $I_{\beta l}^{(i)}$ for $\beta \neq l$.

In [KP99, p. 347], it is suggested to reformulate (6.34) in terms of Stratonovich integrals (rather than the Itô integrals $I_{\beta l}^{(i)}$) and they then propose a simulation scheme for the involved Stratonovich integrals, see [KP99, Ch. 10, (3.7)-(3.10)]. Alternatively, [Gla04, p. 344] provides references for methods aiming at sampling from the respective distributions of the $I_{\beta l}^{(i)}$, $\beta \neq l$. However, in practise, it is often possible (and advisable), to take advantage of the special features of the concrete problem at hand to simplify (6.34) such that sampling the $I_{\beta l}^{(i)}$ for $\beta \neq l$ can be avoided.

6.4 Convergence Order, Error Criteria

As in the previous section, we will restrict ourselves to equidistant time steps of size $h > 0$. The goal is to discuss convergence and error criteria for the Euler scheme (6.15) and for the Milstein schemes (6.28) and (6.34). In particular, we want to address the question if the Milstein schemes really constitutes an improvement over the Euler scheme in any rigorous sense.

So the goal is to compare discrete processes $(\hat{X}_0, \hat{X}_h, \dots, \hat{X}_{Nh})$, $N := \lfloor T/h \rfloor$, given by (6.15), (6.28), or (6.34), at least for sufficiently small h, with the continuous process $(X_t)_{t\in[0,T]}$, constituting the strong solution to (6.1).

Definition 6.6. A function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, $d \in \mathbb{N}$, is called *polynomially bounded* if, and only if, there exist constants $q, k > 0$ such that

$$f(x) \leq k\left(1 + \|x\|^q\right) \quad \text{for each } x \in \mathbb{R}^d. \tag{6.36}$$

For each $\beta \in \mathbb{N}_0$, let $C_{\mathrm{P}}^{\beta}(\mathbb{R}^d)$ denote the set of all $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that f has continuous partials up to order β and such that all these derivatives are polynomially bounded.

Definition 6.7. Consider a fixed $\mathbb{R}^d$-valued stochastic process $(X_t)_{t\in[0,T]}$ and, for each $h > 0$, a discrete $\mathbb{R}^d$-valued process $(\hat{X}_0, \hat{X}_h, \ldots, \hat{X}_{Nh})$. We will refer to the family of discrete processes as the *discretization*.

(a) The discretization is defined to have *strong order of convergence* $\beta > 0$ if, and only if,

$$\exists_{\substack{c>0,\\ \epsilon>0}} \quad \forall_{0<h<\epsilon} \quad E\big(\|\hat{X}_{Nh} - X_T\|\big) \leq ch^{\beta}. \tag{6.37}$$

(b) The discretization is defined to have *weak order of convergence* $\beta \in \mathbb{N}$ if, and only if,

$$\forall_{f\in C_{\mathrm{P}}^{2\beta+2}(\mathbb{R}^d)} \quad \exists_{\substack{c>0,\\ \epsilon>0}} \quad \forall_{0<h<\epsilon} \quad \Big|E\big(f(\hat{X}_{Nh})\big) - E\big(f(X_T)\big)\Big| \leq ch^{\beta}. \tag{6.38}$$

Remark 6.8. The terms *strong* and *weak* in Def. 6.7(a) and Def. 6.7(b), respectively, stem from (6.37), roughly, requiring the *functions* $\hat{X}_{Nh}$ and X_T to be close, whereas (6.38), roughly, requires the *distributions* of $\hat{X}_{Nh}$ and X_T to be close. However, this is somewhat deceiving: While (6.37) clearly implies (6.38) for $d = 1$, $\beta \in \mathbb{N}$ and $f = \mathrm{Id}$ (since, in that case, $\big|E(\hat{X}_{Nh}) - E(X_T)\big| \leq E(|\hat{X}_{Nh} - X_T|)$, in general, (6.37) does not even imply $E(\hat{X}_{Nh}^2) < \infty$. Thus, in general, conditions (6.37) and (6.38) are just different, and it does not always make sense to try to compare them.

Theorem 6.9. *Assume the hypotheses of Th.* 6.3. *The following statements are meant with respect to the strong solution* $(X_t)_{t\in[0,T]}$ *of* (6.1).

(a) *If the coefficient functions* a, b *satisfy the additional following growth condition (which they do trivially in the case of no explicit time dependence)*

$$\exists_{C\geq 0} \quad \forall_{\substack{s,t\in[0,T],\\ x\in\mathbb{R}^d}} \quad \big\|a(s,x) - a(t,x)\big\| + \big\|b(s,x) - b(t,x)\big\| \leq C\left(1 + \|x\|\right)|s-t|^{1/2}, \tag{6.39}$$

then the discretization given by the Euler scheme (6.15) *has strong order of convergence* $1/2$.

(b) *If* $E\big(\|X_{\mathrm{init}}\|^i\big) < \infty$ *for each* $i \in \mathbb{N}$ *and* a, b *have no explicit time dependence, satisfying* $a, b \in C_{\mathrm{P}}^3(\mathbb{R}^d)$, *then the discretization given by the Euler scheme* (6.15) *has weak order of convergence* 1.

Proof. For (a), see [KP99, Th. 10.2.2].

(b) is a special case of [KP99, Th. 14.5.2]. ■

The following Th. 6.11 regarding the convergence of the Milstein scheme requires conditions that extend conditions (b) and (c) of Th. 6.3 to derivatives of the coefficient functions a and b. Following [KP99], we introduce the following differential operators to state these conditions:

Notation 6.10. Define

$$L^0 := \frac{\partial}{\partial t} + \sum_{k=1}^{d} \underline{a}_k \frac{\partial}{\partial x_k}, \tag{6.40a}$$

$$\underset{l=1,\dots,m}{\forall} \quad L^l := \sum_{k=1}^{d} b_{kl} \frac{\partial}{\partial x_k}, \tag{6.40b}$$

$$\underline{a} := a - \frac{1}{2} \sum_{l=1}^{m} L^l b^l, \tag{6.40c}$$

where b^l denotes the lth column of the matrix b. In particular, the definition of $\underline{a}$ requires b to have first partials.

Theorem 6.11. *Under the hypotheses of Th.* 6.3 *and additional requirement that a has first partials, b has first and second partials, and there exist $K_2, K_3, K_4 \geq 0$ (using the same notation for the constants as in* [KP99, Th. 10.3.5]*) such that*

$$\underset{\substack{t\in[0,T],\\ x,y\in\mathbb{R}^d}}{\forall} \quad \big\|\underline{a}(t,x) - \underline{a}(t,y)\big\| \leq K_2\|x-y\|, \tag{6.41a}$$

$$\underset{\substack{t\in[0,T],\\ x,y\in\mathbb{R}^d,\\ \beta,l=1,\dots,m}}{\forall} \quad \big\|L^\beta b^l(t,x) - L^\beta b^l(t,y)\big\| \leq K_2\|x-y\|, \tag{6.41b}$$

$$\underset{\substack{t\in[0,T],\\ x\in\mathbb{R}^d,\\ \lambda=0,\dots,m}}{\forall} \quad \big\|\underline{a}(t,x)\big\| + \big\|L^\lambda \underline{a}(t,x)\big\| \leq K_3\big(1+\|x\|\big) \tag{6.41c}$$

$$\underset{\substack{t\in[0,T],\\ x\in\mathbb{R}^d,\\ \beta,l=1,\dots,m}}{\forall} \quad \big\|L^\beta b^l(t,x)\big\| \leq K_3\big(1+\|x\|\big) \tag{6.41d}$$

$$\underset{\substack{t\in[0,T],x\in\mathbb{R}^d,\\ \lambda=0,\dots,m\\ \beta,l=1,\dots,m}}{\forall} \quad \big\|L^\lambda L^\beta b^l(t,x)\big\| \leq K_3\big(1+\|x\|\big) \tag{6.41e}$$

$$\underset{\substack{s,t\in[0,T],\\ x\in\mathbb{R}^d}}{\forall} \quad \big\|\underline{a}(s,x) - \underline{a}(t,x)\big\| \leq K_4\big(1+\|x\|\big)\,|s-t|^{1/2}, \tag{6.41f}$$

$$\underset{\substack{s,t\in[0,T],\\ x\in\mathbb{R}^d}}{\forall} \quad \big\|b^l(s,x) - b^l(t,x)\big\| \leq K_4\big(1+\|x\|\big)\,|s-t|^{1/2}, \tag{6.41g}$$

$$\underset{\substack{t\in[0,T],\\ x,y\in\mathbb{R}^d,\\ \beta,l=1,\dots,m}}{\forall} \quad \big\|L^\beta b^l(t,x) - L^\beta b^l(t,y)\big\| \leq K_4\big(1+\|x\|\big)\,|s-t|^{1/2}, \tag{6.41h}$$

the discretization given by the Milstein scheme (6.34) *(or, as a special case, by the* 1*-dimensional Milstein scheme* (6.28)*) has strong order of convergence* 1*.*

Proof. See [KP99, Th. 10.3.5], which is proved as a special case of [KP99, Th. 10.6.3], where the proof of [KP99, Th. 10.6.3] uses Itô-Taylor expansions. ■

Remark 6.12. It is stated without reference in [Gla04, p. 347] that the discretization given by the Milstein scheme (6.28) also has weak order of convergence 1 (the same as the Euler scheme).

6.5 Second-Order Methods

We remain in the setting of equidistant time steps $0 < h < T$.

6.5.1 1-Dimensional Case

As before, we will also assume there is no explicit time dependence in a, b during the following heuristic derivation of the second-order scheme.

In Th. 6.11 , we used the differential operators from (6.40) to make notation more managable. In the following, we will, once again, introduce suitable differentiable operators to simplify notation.

Notation 6.13. Define

$$\mathcal{L}^0 := a \frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{2} b^2 \frac{\mathrm{d}^2}{\mathrm{d}x^2} \tag{6.42a}$$

$$\mathcal{L}^1 := b \frac{\mathrm{d}}{\mathrm{d}x}, \tag{6.42b}$$

such that, for each twice differentiable $f : \mathbb{R} \longrightarrow \mathbb{R}$, one has

$$\forall_{x \in \mathbb{R}} \quad \mathcal{L}^0 f(x) = a(x) f'(x) + \frac{1}{2} b^2(x) f''(x), \tag{6.43a}$$

$$\forall_{x \in \mathbb{R}} \quad \mathcal{L}^1 f(x) = b(x) f'(x). \tag{6.43b}$$

Remark 6.14. If the $\mathbb{R}$-valued stochastic process $(X_t)_{t \in [0,T]}$ satisfies the SDE (6.1a) in 1 dimension with a, b not explicitly time-dependent, then, for twice differentiable $f : \mathbb{R} \longrightarrow \mathbb{R}$, Itô's formula (C.3) yields

$$\begin{aligned} \mathrm{d}f(X_t) &= \left(a(X_t)\, f'(X_t) + b^2(X_t)\, \frac{f''(X_t)}{2} \right) \mathrm{d}t \; + b(X_t)\, f'(X_t)\, \mathrm{d}W_t \\ &\overset{\text{Not. 6.13}}{=} \mathcal{L}^0 f(X_t)\, \mathrm{d}t \; + \mathcal{L}^1 f(X_t)\, \mathrm{d}W_t \,. \end{aligned} \tag{6.44}$$

—

As for all previous schemes, we start out from

$$\forall_{\substack{t\in[0,T],\\ 0<h\le T-t}} \quad X_{t+h} = X_t + \int_t^{t+h} a(X_u)\,\mathrm{d}u \;+ \int_t^{t+h} b(X_u)\,\mathrm{d}W_u\,, \tag{6.45}$$

by approximating the integrals. We obtained the Euler scheme (6.15) by using (6.16) and the Milstein schemes of Sec. 6.3 by refining the approximation (6.16b) of the diffusion term. Now we will have to refine the approximation (6.16a) of the drift term as well.

Assuming a to be twice differentiable allows application of (6.44) with $f = a$, yielding

$$\mathrm{d}a(X_t) = \mathcal{L}^0 a(X_t)\,\mathrm{d}t \;+ \mathcal{L}^1 a(X_t)\,\mathrm{d}W_t \tag{6.46a}$$

and

$$\forall_{t,u\in[0,T]} \quad a(X_u) = a(X_t) + \int_t^u \mathcal{L}^0 a(X_s)\,\mathrm{d}s \;+ \int_t^u \mathcal{L}^1 a(X_s)\,\mathrm{d}W_s \tag{6.46b}$$

$$\approx a(X_t) + \mathcal{L}^0 a(X_t) \int_t^u \mathrm{d}s \;+ \mathcal{L}^1 a(X_t) \int_t^u \mathrm{d}W_s\,. \tag{6.46c}$$

Recalling, $0 = t_0 < t_1 < \cdots < t_N = T$, where $t_i = ih$ for each $i = 0, \ldots, N$, we use the approximation (6.46c) for the drift term in (6.45) (with $t = t_i$) to obtain

$$\begin{aligned} \forall_{i=0,\ldots,N-1} \quad \int_{t_i}^{t_i+h} a(X_u)\,\mathrm{d}u \;\approx\; & a(X_{t_i})\,h + \mathcal{L}^0 a(X_{t_i}) \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}u \\ & + \mathcal{L}^1 a(X_{t_i}) \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}W_s\;\mathrm{d}u\,. \end{aligned} \tag{6.47}$$

Introducing the abbreviations (consistent with (6.33b) above; (6.48c) and (6.48d) will be used in the approximation of the diffusion term below)

$$\forall_{i=0,\ldots,N-1} \quad I_{00}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}u\,, \tag{6.48a}$$

$$\forall_{i=0,\ldots,N-1} \quad I_{10}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}W_s\;\mathrm{d}u\,, \tag{6.48b}$$

$$\forall_{i=0,\ldots,N-1} \quad I_{01}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}W_u\,, \tag{6.48c}$$

$$\forall_{i=0,\ldots,N-1} \quad I_{11}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}W_s\;\mathrm{d}W_u\,, \tag{6.48d}$$

(6.47) reads

$$\forall_{i=0,\ldots,N-1} \quad \int_{t_i}^{t_i+h} a(X_u)\,\mathrm{d}u \;\approx a(X_{t_i})\,h + \mathcal{L}^0 a(X_{t_i})\,I_{00}^{(i)} + \mathcal{L}^1 a(X_{t_i})\,I_{10}^{(i)}. \tag{6.49}$$

Analogously, we now work to approximate the diffusion term in (6.45): Assuming b to be twice differentiable allows application of (6.44) with $f = b$, yielding

$$\mathrm{d}b(X_t) = \mathcal{L}^0 b(X_t)\,\mathrm{d}t \;+ \mathcal{L}^1 b(X_t)\,\mathrm{d}W_t \tag{6.50a}$$

and

$$\underset{t,u\in[0,T]}{\forall} \quad b(X_u) = b(X_t) + \int_t^u \mathcal{L}^0 b(X_s)\,\mathrm{d}s + \int_t^u \mathcal{L}^1 b(X_s)\,\mathrm{d}W_s \tag{6.50b}$$

$$\approx b(X_t) + \mathcal{L}^0 b(X_t) \int_t^u \mathrm{d}s + \mathcal{L}^1 b(X_t) \int_t^u \mathrm{d}W_s\,. \tag{6.50c}$$

As the reader will have expected, we now use the approximation (6.50c) for the diffusion term in (6.45) (with $t = t_i$) to obtain

$$\begin{aligned}\int_{t_i}^{t_i+h} b(X_u)\,\mathrm{d}W_u \;&\approx b(X_{t_i})\,(W_{t_i+h} - W_{t_i}) + \mathcal{L}^0 b(X_{t_i}) \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}W_u \\ &\underset{i=0,\dots,N-1}{\forall} \qquad + \mathcal{L}^1 b(X_{t_i}) \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}W_s\;\mathrm{d}W_u \\ &\overset{(6.48)}{=} b(X_{t_i})\,(W_{t_i+h} - W_{t_i}) + \mathcal{L}^0 b(X_{t_i})\,I_{01}^{(i)} + \mathcal{L}^1 b(X_{t_i})\,I_{11}^{(i)}.\end{aligned} \tag{6.51}$$

Using both (6.49) and (6.51) in (6.45), we arrive at the approximation

$$\begin{aligned} X_{t_i+h} &\approx X_{t_i} + a(X_{t_i})\,h + \mathcal{L}^0 a(X_{t_i})\,I_{00}^{(i)} + \mathcal{L}^1 a(X_{t_i})\,I_{10}^{(i)} \\ &\quad + b(X_{t_i})\,(W_{t_i+h} - W_{t_i}) + \mathcal{L}^0 b(X_{t_i})\,I_{01}^{(i)} + \mathcal{L}^1 b(X_{t_i})\,I_{11}^{(i)} \\ &\overset{\text{Not. 6.13}}{=} X_{t_i} + a(X_{t_i})\,h + b(X_{t_i})\,(W_{t_i+h} - W_{t_i}) \\ &\quad + \left(a(X_{t_i})a'(X_{t_i}) + \frac{1}{2}\,b^2(X_{t_i})\,a''(X_{t_i})\right) I_{00}^{(i)} \\ &\quad + \left(a(X_{t_i})b'(X_{t_i}) + \frac{1}{2}\,b^2(X_{t_i})\,b''(X_{t_i})\right) I_{01}^{(i)} \\ &\quad + b(X_{t_i})a'(X_{t_i})\,I_{10}^{(i)} + b(X_{t_i})b'(X_{t_i})\,I_{11}^{(i)}. \end{aligned} \qquad \underset{i=0,\dots,N-1}{\forall} \tag{6.52}$$

In order to make the recursion given by (6.52) implementable, one has to be able to simulate the double integrals $I_{kl}^{(i)}$. To this end, we first note

$$I_{00}^{(i)} = \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}u = \frac{h^2}{2}. \tag{6.53}$$

Next, we already know from (6.26) that

$$I_{11}^{(i)} = \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}W_s\;\mathrm{d}W_u = \int_{t_i}^{t_i+h} (W_u - W_{t_i})\,\mathrm{d}W_u \overset{(6.26)}{=} \frac{\left(W_{t_i+h} - W_{t_i}\right)^2}{2} - \frac{h}{2}. \tag{6.54}$$

A relation between $I_{01}^{(i)}$ and $I_{10}^{(i)}$ can be deduced using the stochastic integration by parts formula (C.4b): Clearly, (C.4b) implies

$$(t_i + h)\,W_{t_i+h} - t_i\,W_{t_i} = \int_{t_i}^{t_i+h} W_u\,\mathrm{d}u + \int_{t_i}^{t_i+h} u\,\mathrm{d}W_u\,. \tag{6.55}$$

Thus

$$\begin{aligned} I_{01}^{(i)} &= \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}s\ \mathrm{d}W_u = \int_{t_i}^{t_i+h} (u-t_i)\,\mathrm{d}W_u \\ &\overset{(6.55)}{=} (t_i+h)\,W_{t_i+h} - t_i\,W_{t_i} - \int_{t_i}^{t_i+h} W_u\,\mathrm{d}u - t_i\,W_{t_i+h} + t_i\,W_{t_i} \\ &= h\,W_{t_i+h} - \int_{t_i}^{t_i+h} W_u\,\mathrm{d}u = h\,(W_{t_i+h}-W_{t_i}) - \int_{t_i}^{t_i+h}(W_u - W_{t_i})\,\mathrm{d}u \\ &= h\,(W_{t_i+h}-W_{t_i}) - \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}W_s\ \mathrm{d}u \\ &\overset{(6.48\mathrm{b})}{=} h\,(W_{t_i+h}-W_t) - I_{10}^{(i)}. \end{aligned} \tag{6.56}$$

So it just remains to find a way to simulate the distribution of

$$I_{10}^{(i)} = \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}W_s\ \mathrm{d}u = \int_{t_i}^{t_i+h}(W_u - W_{t_i})\,\mathrm{d}u\,. \tag{6.57}$$

More precisely, we have to be able to simulate the distribution of $I_{10}^{(i)}$, given $W_{t_i} = x$, $x \in \mathbb{R}$. The trick is to notice that $(W_{t_i+h} - W_{t_i})|\{W_{t_i} = x\}$ and $I_{10}^{(i)}|\{W_{t_i} = x\}$ are jointly normal, albeit not independent, where the details are compiled in the following proposition:

Proposition 6.15. *Let $(W_t)_{t\geq 0}$ be a 1-dimensional standard Brownian motion with drift 0 and variance 1, $t_i \geq 0$, $h > 0$, and $x \in \mathbb{R}$. If $I_{10}^{(i)}$ is according to (6.57) and*

$$(\Delta W)^{(i)} := W_{t_i+h} - W_{t_i}, \tag{6.58}$$

then

$$\begin{pmatrix} (\Delta W)^{(i)} \\ I_{10}^{(i)} \end{pmatrix} \Big| \{W_{t_i} = x\} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} h & \frac{h^2}{2} \\ \frac{h^2}{2} & \frac{h^3}{3} \end{pmatrix} \right). \tag{6.59}$$

Proof. Introducing the abbreviations

$$\underset{t\geq t_i}{\forall} \quad W_t^x := W_t|\{W_{t_i} = x\}, \tag{6.60a}$$

$$X := (\Delta W)^{(i)}|\{W_{t_i} = x\} = W_{t_i+h}^x - x, \tag{6.60b}$$

$$Y := I_{10}^{(i)}|\{W_{t_i} = x\} = \int_{t_i}^{t_i+h} (W_u^x - x)\,\mathrm{d}u\,, \tag{6.60c}$$

we note

$$\left(\underset{t\geq t_i}{\forall} \quad W_t^x \sim N(x, t-t_i) \right) \quad \Rightarrow \quad X \sim N(0,h). \tag{6.61}$$

To obtain the distribution of Y, we use the translation invariance of Brownian motions to observe $Y \sim \int_0^h \big(W_u|\{W_0 = x\} - x\big)\,\mathrm{d}u$. Then the proof of Prop. 4.21 shows $Y = Y_h$,

where $(Y_t)_{t\geq 0}$ is a 1-dimensional Brownian motion with drift 0 and variance $(h-t)^2$. Thus, Y is normally distributed with

$$E(Y) = 0 \tag{6.62a}$$

$$V(Y) = \int_0^h (h-t)^2 \,\mathrm{d}t \;= \left[\frac{(t-h)^3}{3}\right]_0^h = \frac{h^3}{3} \tag{6.62b}$$

$$\Rightarrow \quad Y \sim N\left(0, \frac{h^3}{3}\right). \tag{6.62c}$$

Thus, to establish (6.59), it merely remains to show $\mathrm{Cov}(X,Y) = \frac{h^2}{2}$. Since $E(X) = E(Y) = 0$, we have

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E(XY) = \int_\Omega \left(\left(W^x_{t_i+h}(\omega) - x\right) \int_{t_i}^{t_i+h} \left(W^x_u(\omega) - x\right) \mathrm{d}u \right) \mathrm{d}\omega \\
&\overset{\text{Fubini}}{=} \int_{t_i}^{t_i+h} \int_\Omega \left(W^x_{t_i+h}(\omega) - x\right) \left(W^x_u(\omega) - x\right) \mathrm{d}\omega \; \mathrm{d}u \\
&= \int_{t_i}^{t_i+h} \mathrm{Cov}(W^x_{t_i+h} - x,\, W^x_u - x) \,\mathrm{d}u \\
&\overset{(4.6)}{=} \int_{t_i}^{t_i+h} \min\{t_i + h - t_i, u - t_i\} \,\mathrm{d}u \\
&= \int_{t_i}^{t_i+h} (u - t_i) \,\mathrm{d}u \;= \frac{t_i^2}{2} + 2t_i h + \frac{h^2}{2} - \frac{t_i^2}{2} - 2t_i h = \frac{h^2}{2},
\end{aligned} \tag{6.63}$$

where the translation invariance of Brownian motions was also used to apply (4.6), since (4.6) required a standard Brownian motion starting at $t = 0$, whereas $(W^x_t - x)_{t\geq t_i}$ constitutes a standard Brownian motion starting at $t = t_i$.

Combining (6.61), (6.62), and (6.63) completes the proof of (6.59). ■

Putting everything together by making use of the above formulas for the $I^{(i)}_{kl}$ and (6.59) in (6.52), we are now in a position to formulate the 1-dimensional second-order scheme (it is actually also due to Milstein): Let $(Z^{(1)}, \ldots, Z^{(N)})$ be an i.i.d. family of random vectors such that

$$\underset{i=1,\ldots,N}{\forall} \quad Z^{(i)} = \begin{pmatrix} Z^{(i)}_1 \\ Z^{(i)}_2 \end{pmatrix} \;\sim\; N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} h & h^2/2 \\ h^2/2 & h^3/3 \end{pmatrix} \right). \tag{6.64}$$

Then the 1-dimensional second-order scheme reads

$$\hat{X}^{(0)} := X_{\text{init}}, \tag{6.65a}$$

$$\begin{aligned}\hat{X}^{(i+1)} &:= \hat{X}^{(i)} + a(\hat{X}^{(i)})\,h + b(\hat{X}^{(i)})\,Z_1^{(i+1)} \\ &\quad + \left(a(\hat{X}^{(i)})a'(\hat{X}^{(i)}) + \frac{1}{2}\,b^2(\hat{X}^{(i)})\,a''(\hat{X}^{(i)})\right)\,\frac{h^2}{2} \\ &\quad + \left(a(\hat{X}^{(i)})b'(\hat{X}^{(i)}) + \frac{1}{2}\,b^2(\hat{X}^{(i)})\,b''(\hat{X}^{(i)})\right)\,(Z_1^{(i+1)}\,h - Z_2^{(i+1)}) \\ &\quad + b(\hat{X}^{(i)})a'(\hat{X}^{(i)})\,Z_2^{(i+1)} + \frac{1}{2}\,b(\hat{X}^{(i)})b'(\hat{X}^{(i)})\,\left((Z_1^{(i+1)})^2 - h\right).\end{aligned} \quad \forall_{i=0,\dots,N-1} \tag{6.65b}$$

Theorem 6.16. *Assume the hypotheses of Th.* 6.3. *The following statements are meant with respect to the strong solution* $(X_t)_{t\in[0,T]}$ *of* (6.1).

(a) *If* $E\big(|X_{\text{init}}|^i\big) < \infty$ *for each* $i \in \mathbb{N}$ *and* a, b *have no explicit time dependence, satisfying* $a, b \in C^6(\mathbb{R})$, *where all derivatives are uniformly bounded, then the discretization given by the* 1*-dimensional second-order scheme* (6.65) *has weak order of convergence* 2.

(b) *The assertion of* (a) *remains true if the discretization is given by the following simplified* 1*-dimensional second-order scheme (arising from replacing* $I_{10}^{(i)}$ *by* $(\Delta W)^{(i)}\,h/2$ *and using* $Z_{i+1}\sqrt{h}$ *to simulate* $(\Delta W)^{(i)}$*), using an i.i.d. family of random variables* $(Z_1, \dots, Z_N)$, *each* $Z_i \sim N(0,1)$*):*

$$\hat{X}^{(0)} := X_{\text{init}}, \tag{6.66a}$$

$$\begin{aligned}&\hat{X}^{(i+1)} \\ &:= \hat{X}^{(i)} + a(\hat{X}^{(i)})\,h + b(\hat{X}^{(i)})\,Z_{i+1}\sqrt{h} \\ &\quad + \left(a(\hat{X}^{(i)})a'(\hat{X}^{(i)}) + \frac{1}{2}\,b^2(\hat{X}^{(i)})\,a''(\hat{X}^{(i)})\right)\,\frac{h^2}{2} \\ &\quad + \left(a(\hat{X}^{(i)})b'(\hat{X}^{(i)}) + \frac{1}{2}\,b^2(\hat{X}^{(i)})\,b''(\hat{X}^{(i)}) + b(\hat{X}^{(i)})a'(\hat{X}^{(i)})\right)\,Z_{i+1}\frac{h^{\frac{3}{2}}}{2} \\ &\quad + \frac{1}{2}\,b(\hat{X}^{(i)})b'(\hat{X}^{(i)})\,(Z_{i+1}^2 - 1)\,h.\end{aligned} \quad \forall_{i=0,\dots,N-1} \tag{6.66b}$$

Proof. Both (a) and (b) are special cases of [KP99, Th. 14.2.4] which, in turn, is proved as a special case of [KP99, Th. 14.5.2]. ■

Example 6.17. For the SDE

$$\mathrm{d}X_t \,= \sin(X_t)\,\mathrm{d}t \,+ X_t^2\,\mathrm{d}W_t\,, \tag{6.67a}$$

we have $a(x) = \sin(x)$ and $b(x) = x^2$, i.e. (6.66) takes the form

$$\hat{X}^{(0)} := X_{\text{init}}, \tag{6.67b}$$

$$\begin{aligned}\hat{X}^{(i+1)} := \hat{X}^{(i)} &+ \sin(\hat{X}^{(i)})\, h + (\hat{X}^{(i)})^2\, Z_{i+1}\sqrt{h} \\ &+ \left(\sin(\hat{X}^{(i)})\cos(\hat{X}^{(i)}) - \frac{1}{2}\,(\hat{X}^{(i)})^4\,\sin(\hat{X}^{(i)})\right)\,\frac{h^2}{2} \\ &+ \left(2\hat{X}^{(i)}\sin(\hat{X}^{(i)}) + (\hat{X}^{(i)})^4 + (\hat{X}^{(i)})^2\cos(\hat{X}^{(i)})\right)\,Z_{i+1}\frac{h^{\frac{3}{2}}}{2} \\ &+ (\hat{X}^{(i)})^3\,(Z_{i+1}^2 - 1)\,h.\end{aligned} \qquad \underset{i=0,\dots,N-1}{\forall} \tag{6.67c}$$

6.5.2 Multi-Dimensional Case

To see that this case does not constitute a purely academic exercise, we mention two examples arising from mathematical finance modeling:

Example 6.18. (a) In the *stochastic volatility model* according to [Hes93] an asset price S and the corresponding volatility V are modeled as $\mathbb{R}$-valued stochastic processes $(S_t)_{t\in[0,T]}$ and $(V_t)_{t\in[0,T]}$, respectively, $T > 0$, satisfying the coupled system of SDE

$$\mathrm{d}S_t \;= rS_t\,\mathrm{d}t \;+ S_t\,\sqrt{V_t}\,\mathrm{d}(W_1)_t\,, \tag{6.68a}$$

$$\mathrm{d}V_t \;= \kappa\,(\theta - V_t)\,\mathrm{d}t \;+ \sqrt{V_t}\,\big(\sigma_1\,\mathrm{d}(W_1)_t \;+ \sigma_2\,\mathrm{d}(W_2)_t\,\big), \tag{6.68b}$$

where $r, \kappa, \theta, \sigma_1, \sigma_2 > 0$ and $((W_1)_t)_{t\in[0,T]}$, $((W_2)_t)_{t\in[0,T]}$ are independent 1-dimensional standard Brownian motions with drift 0 and variance 1.

(b) According to the *LIBOR*[1] *market model* described in [Gla04, Sec. 3.7.1] (actually somewhat simplified for our purposes here), the so-called forward interest rates $L_1, \dots, L_d$, $d \in \mathbb{N}$, are $\mathbb{R}$-valued processes satisfying the coupled system of d SDE

$$\mathrm{d}(L_k)_t \;= (L_k)_t\,\mu_k\big((L_1)_t, \dots, (L_d)_t\big)\,\mathrm{d}t \;+ (L_k)_t\,\sigma_k^{\mathrm{t}}\,\mathrm{d}W_t \qquad \underset{k=1,\dots,d}{\forall} \tag{6.69}$$

with given functions $\mu_k : \mathbb{R}^d \longrightarrow \mathbb{R}$, $\sigma_k \in \mathbb{R}^d$ (interpreted as column vectors), and a d-dimensional standard Brownian motion $(W_t)_{t\in[0,T]}$ with drift 0 and covariance matrix Id.

Our goal now is to find a multi-dimensional version of the second-order scheme (6.65), considering the general SDE (6.1) with $\mathbb{R}^d$-valued $(X_t)_{t\in[0,T]}$ and $\mathbb{R}^m$-valued $(W_t)_{t\in[0,T]}$, $d, m \in \mathbb{N}$. As before, for $k = 1, \dots, d$ and $l = 1, \dots, m$, let $(X_t)_k$, a_k, and b_{kl} denote the components of the functions X_t, a, and b, respectively, where we continue to assume that a and b are not explicitly time-dependent.

[1] London Interbank Offered Rate

As in Sec. 6.3.2, our starting point is

$$\underset{\substack{t\in[0,T-h],\\ k=1,\dots,d}}{\forall} \quad (X_{t+h})_k = (X_t)_k + \int_t^{t+h} a_k(X_u)\,\mathrm{d}u \;+ \sum_{l=1}^{m} \int_t^{t+h} b_{kl}(X_u)\,\mathrm{d}(W_l)_u\,. \tag{6.70}$$

The strategy is to proceed analogous to the 1-dimensional case of Sec. 6.5.1 and to use Itô's formula to obtain suitable approximations of the integrals in (6.70). However, we now need to employ the multi-dimensional version of Itô's formula, i.e. (C.9). In preparation, we generalize our differential operators $\mathcal{L}^0, \mathcal{L}^1$ of (6.42):

Notation 6.19. Define

$$\mathcal{L}^0 := \sum_{k=1}^{d} a_k\,\partial_{x_k} + \frac{1}{2}\sum_{k,\alpha=1}^{d}\sum_{l=1}^{m} b_{kl}b_{\alpha l}\,\partial_{x_k}\partial_{x_\alpha} \tag{6.71a}$$

$$\underset{l=1,\dots,m}{\forall} \quad \mathcal{L}^l := \sum_{k=1}^{d} b_{kl}\,\partial_{x_k}. \tag{6.71b}$$

Remark 6.20. If the $\mathbb{R}^d$-valued stochastic process $(X_t)_{t\in[0,T]}$ satisfies the SDE (6.1a) with a, b not explicitly time-dependent, then, for twice differentiable $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, and using

$$\underset{k,\alpha=1,\dots,d}{\forall} \quad \Sigma_{k\alpha} := \sum_{l=1}^{m} b_{kl}b_{\alpha l}, \tag{6.72}$$

Itô's formula (C.9) yields

$$\begin{aligned}
\mathrm{d}f(X_t) \quad &= \quad \left(\sum_{k=1}^{d} \partial_{x_k} f(X_t)\, a_k(X_t) + \frac{1}{2}\sum_{k,\alpha=1}^{d} \partial_{x_k}\partial_{x_\alpha} f(X_t)\,\Sigma_{k\alpha}(X_t)\right)\mathrm{d}t \\
&\qquad + \sum_{k=1}^{d} \partial_{x_k} f(X_t)\, b_{k\cdot}(X_t)\,\mathrm{d}W_t \\
&\overset{\text{Not. 6.19}}{=} \quad \mathcal{L}^0 f(X_t)\,\mathrm{d}t \;+ \sum_{l=1}^{m} \mathcal{L}^l f(X_t)\,\mathrm{d}(W_l)_t\,.
\end{aligned} \tag{6.73}$$

—

Assuming a_k to be twice differentiable allows application of (6.73) with $f = a_k$, yielding

$$\underset{k=1,\dots,d}{\forall} \quad \mathrm{d}a_k(X_t) = \mathcal{L}^0 a_k(X_t)\,\mathrm{d}t \;+ \sum_{l=1}^{m} \mathcal{L}^l a_k(X_t)\,\mathrm{d}(W_l)_t \tag{6.74a}$$

and

$$\underset{\substack{k=1,\dots,d,\\ t,u\in[0,T]}}{\forall} \quad a_k(X_u) = a_k(X_t) + \int_t^u \mathcal{L}^0 a_k(X_s)\,\mathrm{d}s \;+ \sum_{l=1}^{m} \int_t^u \mathcal{L}^l a_k(X_s)\,\mathrm{d}(W_l)_s \tag{6.74b}$$

$$\approx a_k(X_t) + \mathcal{L}^0 a_k(X_t)\int_t^u \mathrm{d}s \;+ \sum_{l=1}^{m} \mathcal{L}^l a_k(X_t)\int_t^u \mathrm{d}(W_l)_s\,. \tag{6.74c}$$

Analogously, assuming the b_{kl} to be twice differentiable allows application of (6.73) with $f = b_{kl}$, yielding

$$\forall_{\substack{k=1,\dots,d,\\ l=1,\dots,m}} \quad \mathrm{d}b_{kl}(X_t) = \mathcal{L}^0 b_{kl}(X_t)\,\mathrm{d}t + \sum_{\beta=1}^{m} \mathcal{L}^\beta b_{kl}(X_t)\,\mathrm{d}(W_\beta)_t \tag{6.75a}$$

and

$$\forall_{\substack{k=1,\dots,d,\\ l=1,\dots,m,\\ t,u\in[0,T]}} \quad b_{kl}(X_u) = b_{kl}(X_t) + \int_t^u \mathcal{L}^0 b_{kl}(X_s)\,\mathrm{d}s + \sum_{\beta=1}^{m} \int_t^u \mathcal{L}^\beta b_{kl}(X_s)\,\mathrm{d}(W_\beta)_s \tag{6.75b}$$

$$\approx b_{kl}(X_t) + \mathcal{L}^0 b_{kl}(X_t) \int_t^u \mathrm{d}s + \sum_{\beta=1}^{m} \mathcal{L}^\beta b_{kl}(X_t) \int_t^u \mathrm{d}(W_\beta)_s\,. \tag{6.75c}$$

In generalization of (6.48), we introduce for $0 = t_0 < t_1 < \cdots < t_N = T$, where $t_i = ih$ for each $i = 0, \dots, N$,

$$\forall_{i=0,\dots,N-1} \quad I_{00}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}u = \frac{h^2}{2}, \tag{6.76a}$$

$$\forall_{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}} \quad I_{l0}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}(W_l)_s\;\mathrm{d}u\,, \tag{6.76b}$$

$$\forall_{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}} \quad I_{0l}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}(W_l)_u\,, \tag{6.76c}$$

$$\forall_{\substack{i=0,\dots,N-1,\\ l,\beta=1,\dots,m}} \quad I_{l\beta}^{(i)} := \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}(W_l)_s\;\mathrm{d}(W_\beta)_u\,. \tag{6.76d}$$

Combining the approximations (6.74c) and (6.75c) (for $t = t_i$) with (6.76) yields the approximations

$$\begin{aligned} \int_{t_i}^{t_i+h} a_k(X_u)\,\mathrm{d}u \;&\approx a_k(X_{t_i})\,h + \mathcal{L}^0 a_k(X_{t_i}) \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}s\;\mathrm{d}u \\ &\quad + \sum_{l=1}^{m} \mathcal{L}^l a_k(X_{t_i}) \int_{t_i}^{t_i+h} \int_{t_i}^{u} \mathrm{d}(W_l)_s\;\mathrm{d}u \\ &= a_k(X_{t_i})\,h + \mathcal{L}^0 a_k(X_{t_i})\,I_{00}^{(i)} + \sum_{l=1}^{m} \mathcal{L}^l a_k(X_{t_i})\,I_{l0}^{(i)} \end{aligned} \tag{6.77}$$

$$\forall_{\substack{i=0,\dots,N-1,\\ k=1,\dots,d}}$$

and, with

$$\forall_{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}} \quad (\Delta W_l)^{(i)} := (W_l)_{t_i+h} - (W_l)_{t_i}, \tag{6.78}$$

the approximations

$$\begin{aligned}\int_{t_i}^{t_i+h} b_{kl}(X_u)\,\mathrm{d}(W_l)_u &\approx b_{kl}(X_{t_i})\,(\Delta W_l)^{(i)} + \mathcal{L}^0 b_{kl}(X_{t_i}) \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}s\ \mathrm{d}(W_l)_u \\ &\quad + \sum_{\beta=1}^{m} \mathcal{L}^\beta b_{kl}(X_{t_i}) \int_{t_i}^{t_i+h}\int_{t_i}^{u} \mathrm{d}(W_\beta)_s\ \mathrm{d}(W_l)_u \\ &\quad \underset{\substack{i=0,\dots,N-1,\\ k=1,\dots,d,\\ l=1,\dots,m}}{\forall} \\ &= b_{kl}(X_{t_i})\,(\Delta W_l)^{(i)} + \mathcal{L}^0 b_{kl}(X_{t_i})\, I_{0l}^{(i)} \\ &\quad + \sum_{\beta=1}^{m} \mathcal{L}^\beta b_{kl}(X_{t_i})\, I_{\beta l}^{(i)}.\end{aligned} \tag{6.79}$$

Plugging (6.77), (6.79) and $I_{00}^{(i)} = \frac{h^2}{2}$ into (6.70) (for $t = t_i$ and replacing X_{t_i} by $\hat{X}^{(i)}$), we obtain the following multi-dimensional second-order scheme:

$$\hat{X}^{(0)} := X_{\text{init}}, \tag{6.80a}$$

$$\begin{aligned}\hat{X}_k^{(i+1)} &:= \hat{X}_k^{(i)} + a_k(\hat{X}^{(i)})\,h + \mathcal{L}^0 a_k(\hat{X}^{(i)})\,\frac{h^2}{2} + \sum_{l=1}^{m} \mathcal{L}^l a_k(\hat{X}^{(i)})\, I_{l0}^{(i)} \\ &\quad + \sum_{l=1}^{m} \left(b_{kl}(\hat{X}^{(i)})\,(\Delta W_l)^{(i)} + \mathcal{L}^0 b_{kl}(\hat{X}^{(i)})\, I_{0l}^{(i)} + \sum_{\beta=1}^{m} \mathcal{L}^\beta b_{kl}(\hat{X}^{(i)})\, I_{\beta l}^{(i)} \right). \\ &\quad \underset{\substack{i=0,\dots,N-1,\\ k=1,\dots,d}}{\forall}\end{aligned} \tag{6.80b}$$

Remark 6.21. (a) In practise, one will have to use (6.71) to replace the differential operators in (6.80b) by the actual derivatives of the components of the coefficient functions a and b, resulting in polynomials in the components of the coefficient functions and their first- and second-order partials.

(b) A computation analogous to (6.56) (using stochastic integration by parts) yields

$$\underset{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}}{\forall} \quad I_{0l}^{(i)} = h\,(\Delta W_l)^{(i)} - I_{l0}^{(i)} \tag{6.81a}$$

and

$$\underset{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}}{\forall} \quad I_{ll}^{(i)} = \frac{\left((\Delta W_l)^{(i)}\right)^2 - h}{2} \tag{6.81b}$$

is known from (6.35). So, to make use of (6.80), it "just" remains to sample the $(\Delta W_l)^{(i)}$ together with the $I_{l0}^{(i)}$ and the $I_{l\beta}^{(i)}$ for $l, \beta = 1, \dots, m$ with $l \neq m$. Unfortunately, as already remarked at the end of Sec. 6.3.2, this is difficult in this form, but simplifications as described in Sec. 6.5.3 and Sec. 6.5.4 below do yield feasible ways to implement the multi-dimensional second order scheme.

Example 6.22. In order to formulate the second-order scheme recursion (6.80b) for the stochastic volatility model (6.68) of Ex. 6.18(a) we start by bringing (6.68) into the form (6.1a) using $(X_t)_1 := S_t > 0$ and $(X_t)_2 := V_t > 0$. Letting

$$a : (\mathbb{R}^+)^2 \longrightarrow \mathbb{R}^2, \qquad a\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} r\, x_1 \\ \kappa\,(\theta - x_2) \end{pmatrix}, \tag{6.82a}$$

$$b : (\mathbb{R}^+)^2 \longrightarrow \mathbb{R}^{2\times 2}, \qquad b\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} x_1\sqrt{x_2} & 0 \\ \sigma_1\sqrt{x_2} & \sigma_2\sqrt{x_2} \end{pmatrix}, \tag{6.82b}$$

we have

$$\begin{aligned} \mathrm{d}S_t &= \mathrm{d}(X_t)_1 = a_1\begin{pmatrix} S_t \\ V_t \end{pmatrix}\mathrm{d}t + b_{11}\begin{pmatrix} S_t \\ V_t \end{pmatrix}\mathrm{d}(W_1)_t + b_{12}\begin{pmatrix} S_t \\ V_t \end{pmatrix}\mathrm{d}(W_2)_t \\ &= rS_t\,\mathrm{d}t + S_t\sqrt{V_t}\,\mathrm{d}(W_1)_t, \end{aligned} \tag{6.82c}$$

$$\begin{aligned} \mathrm{d}V_t &= \mathrm{d}(X_t)_2 = a_2\begin{pmatrix} S_t \\ V_t \end{pmatrix}\mathrm{d}t + b_{21}\begin{pmatrix} S_t \\ V_t \end{pmatrix}\mathrm{d}(W_1)_t + b_{22}\begin{pmatrix} S_t \\ V_t \end{pmatrix}\mathrm{d}(W_2)_t \\ &= \kappa\,(\theta - V_t)\,\mathrm{d}t + \sqrt{V_t}\,\big(\sigma_1\,\mathrm{d}(W_1)_t + \sigma_2\,\mathrm{d}(W_2)_t\big), \end{aligned} \tag{6.82d}$$

i.e. (6.68) does, indeed, have the form (6.1a). We proceed to compute the relevant derivatives in preparation for the formulation of (6.80b):

$$\begin{aligned} \Sigma\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= b\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} b^{\mathrm{t}}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1\sqrt{x_2} & 0 \\ \sigma_1\sqrt{x_2} & \sigma_2\sqrt{x_2} \end{pmatrix}\begin{pmatrix} x_1\sqrt{x_2} & \sigma_1\sqrt{x_2} \\ 0 & \sigma_2\sqrt{x_2} \end{pmatrix} \\ &= \begin{pmatrix} x_1^2\, x_2 & \sigma_1\, x_1\, x_2 \\ \sigma_1\, x_1\, x_2 & (\sigma_1^2 + \sigma_2^2)\, x_2 \end{pmatrix}, \end{aligned} \tag{6.83a}$$

$$\begin{aligned} \mathcal{L}^0 &= \sum_{k=1}^{2} a_k\,\partial_{x_k} + \frac{1}{2}\sum_{k,\alpha=1}^{2}\Sigma_{k\alpha}\,\partial_{x_k}\partial_{x_\alpha} \\ &= r\,x_1\,\partial_{x_1} + \kappa\,(\theta - x_2)\,\partial_{x_2} + \frac{1}{2}\Big(x_1^2\,x_2\,\partial_{x_1}\partial_{x_1} + \sigma_1\,x_1\,x_2(\partial_{x_1}\partial_{x_2} + \partial_{x_2}\partial_{x_1}) \\ &\qquad + (\sigma_1^2 + \sigma_2^2)\,x_2\,\partial_{x_2}\partial_{x_2}\Big), \end{aligned} \tag{6.83b}$$

$$\mathcal{L}^1 = \sum_{k=1}^{2} b_{k1}\,\partial_{x_k} = x_1\sqrt{x_2}\,\partial_{x_1} + \sigma_1\sqrt{x_2}\,\partial_{x_2}, \tag{6.83c}$$

$$\mathcal{L}^2 = \sum_{k=1}^{2} b_{k2}\,\partial_{x_k} = \sigma_2\sqrt{x_2}\,\partial_{x_2}, \tag{6.83d}$$

$$\mathcal{L}^0 a_1\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = r^2\,x_1, \tag{6.83e}$$

$$\mathcal{L}^0 a_2\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -\kappa^2\,(\theta - x_2), \tag{6.83f}$$

$$\mathcal{L}^0 b_{11}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = r\,x_1\sqrt{x_2} + \frac{\kappa\,(\theta - x_2)\,x_1}{2\sqrt{x_2}} + \frac{1}{2}\left(\sigma_1\,x_1\sqrt{x_2} - \frac{(\sigma_1^2 + \sigma_2^2)\,x_1}{4\sqrt{x_2}}\right), \tag{6.83g}$$

$$\mathcal{L}^0 b_{12}\begin{pmatrix}x_1\\x_2\end{pmatrix} = 0, \tag{6.83h}$$

$$\mathcal{L}^0 b_{21}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \frac{\kappa\,\sigma_1\,(\theta - x_2)}{2\sqrt{x_2}} - \frac{\sigma_1\,(\sigma_1^2+\sigma_2^2)}{8\sqrt{x_2}}, \tag{6.83i}$$

$$\mathcal{L}^0 b_{22}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \frac{\kappa\,\sigma_2\,(\theta - x_2)}{2\sqrt{x_2}} - \frac{\sigma_2\,(\sigma_1^2+\sigma_2^2)}{8\sqrt{x_2}}, \tag{6.83j}$$

$$\mathcal{L}^1 a_1\begin{pmatrix}x_1\\x_2\end{pmatrix} = r\,x_1\,\sqrt{x_2}, \tag{6.83k}$$

$$\mathcal{L}^1 a_2\begin{pmatrix}x_1\\x_2\end{pmatrix} = -\kappa\,\sigma_1\,\sqrt{x_2}, \tag{6.83l}$$

$$\mathcal{L}^1 b_{11}\begin{pmatrix}x_1\\x_2\end{pmatrix} = x_1\,x_2 + \frac{x_1\,\sigma_1}{2}, \tag{6.83m}$$

$$\mathcal{L}^1 b_{12}\begin{pmatrix}x_1\\x_2\end{pmatrix} = 0, \tag{6.83n}$$

$$\mathcal{L}^1 b_{21}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \frac{\sigma_1^2}{2}, \tag{6.83o}$$

$$\mathcal{L}^1 b_{22}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \frac{\sigma_1\,\sigma_2}{2}, \tag{6.83p}$$

$$\mathcal{L}^2 a_1\begin{pmatrix}x_1\\x_2\end{pmatrix} = 0, \tag{6.83q}$$

$$\mathcal{L}^2 a_2\begin{pmatrix}x_1\\x_2\end{pmatrix} = -\kappa\,\sigma_2\,\sqrt{x_2}, \tag{6.83r}$$

$$\mathcal{L}^2 b_{11}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \frac{x_1\,\sigma_2}{2}, \tag{6.83s}$$

$$\mathcal{L}^2 b_{12}\begin{pmatrix}x_1\\x_2\end{pmatrix} = 0, \tag{6.83t}$$

$$\mathcal{L}^2 b_{21}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \frac{\sigma_1\,\sigma_2}{2}, \tag{6.83u}$$

$$\mathcal{L}^2 b_{22}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \frac{\sigma_2^2}{2}. \tag{6.83v}$$

We obtain the recursion by plugging the appropriate terms from above into (6.80b):

$$\begin{aligned}\hat{S}^{(i+1)} := \hat{S}^{(i)} &+ r\,\hat{S}^{(i)}\,h + \frac{r^2\,\hat{S}^{(i)}\,h^2}{2} + r\,\hat{S}^{(i)}\,\sqrt{\hat{V}^{(i)}}\,I_{10}^{(i)} + \hat{S}^{(i)}\,\sqrt{\hat{V}^{(i)}}\,(\Delta W_1)^{(i)}\\ &+ r\,\hat{S}^{(i)}\,\sqrt{\hat{V}^{(i)}} + \left(\frac{\kappa\,(\theta - \hat{V}^{(i)})\,\hat{S}^{(i)}}{2\sqrt{\hat{V}^{(i)}}} + \frac{1}{2}\left(\sigma_1\,\hat{S}^{(i)}\,\sqrt{\hat{V}^{(i)}} - \frac{(\sigma_1^2+\sigma_2^2)\,\hat{S}^{(i)}}{4\sqrt{\hat{V}^{(i)}}}\right)\right) I_{01}^{(i)}\\ &+ \left(\hat{S}^{(i)}\,\hat{V}^{(i)} + \frac{\hat{S}^{(i)}\,\sigma_1}{2}\right) I_{11}^{(i)} + \frac{\hat{S}^{(i)}\,\sigma_2}{2}\,I_{21}^{(i)},\end{aligned} \tag{6.84a}$$

$$\begin{aligned}\hat{V}^{(i+1)} &:= \hat{V}^{(i)} + \kappa\,(\theta - \hat{V}^{(i)})\,h - \kappa^2\,(\theta - \hat{V}^{(i)})\,\frac{h^2}{2} - \kappa\,\sqrt{\hat{V}^{(i)}}\,\Big(\sigma_1\,I_{10}^{(i)} + \sigma_2\,I_{20}^{(i)}\Big)\\ &\quad + \sqrt{\hat{V}^{(i)}}\,\big(\sigma_1\,(\Delta W_1)^{(i)} + \sigma_2\,(\Delta W_2)^{(i)}\big)\\ &\quad + \Big(4\kappa\,\Big(\theta - \hat{V}^{(i)}\Big) - (\sigma_1^2 + \sigma_2^2)\Big)\,\frac{\sigma_1\,I_{01}^{(i)} + \sigma_2\,I_{02}^{(i)}}{8\sqrt{\hat{V}^{(i)}}}\\ &\quad + \frac{1}{2}\Big(\sigma_1^2\,I_{11}^{(i)} + \sigma_1\sigma_2\,I_{21}^{(i)} + \sigma_1\sigma_2\,I_{12}^{(i)} + \sigma_2^2\,I_{22}^{(i)}\Big)\,. \qquad (6.84b)\end{aligned}$$

Example 6.23. Consider the LIBOR market model (6.69) of Ex. 6.18(b). To write (6.69) in the form (6.1a), we let $(X_t)_k := (L_t)_k$ for $k = 1, \ldots, d$,

$$a:\,\mathbb{R}^d \longrightarrow \mathbb{R}^d, \qquad a\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} := \begin{pmatrix} x_1\,\mu_1(x_1, \ldots, x_d) \\ \vdots \\ x_d\,\mu_d(x_1, \ldots, x_d) \end{pmatrix}, \qquad (6.85a)$$

$$b:\,\mathbb{R}^d \longrightarrow \mathbb{R}^{d\times d}, \qquad b\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} := \begin{pmatrix} x_1\,\sigma_{11} & \ldots & x_1\,\sigma_{1d} \\ \vdots & \ddots & \vdots \\ x_d\,\sigma_{d1} & \ldots & x_d\,\sigma_{dd} \end{pmatrix}, \qquad (6.85b)$$

we have

$$\begin{aligned}\underset{k=1,\ldots,d}{\forall}\quad \mathrm{d}(L_k)_t &= \mathrm{d}(X_t)_k = a_k\begin{pmatrix}(L_1)_t \\ \vdots \\ (L_d)_t\end{pmatrix}\,\mathrm{d}t + \sum_{l=1}^{d} b_{kl}\begin{pmatrix}(L_1)_t \\ \vdots \\ (L_d)_t\end{pmatrix}\,\mathrm{d}(W_l)_t \qquad (6.85c)\\ &= (L_k)_t\,\mu_k\big((L_1)_t, \ldots, (L_d)_t\big)\,\mathrm{d}t + (L_k)_t\,\sigma_k^{\mathrm{t}}\,\mathrm{d}W_t\,,\end{aligned}$$

i.e. we have succeeded in writing (6.69) in the form (6.1a).

We will formulate the second-order scheme recursion (6.80b) only for $d = 2$. In preparation, we compute the following quantities:

$$\begin{aligned}\Sigma\begin{pmatrix}x_1 \\ x_2\end{pmatrix} &= b\begin{pmatrix}x_1 \\ x_2\end{pmatrix} b^{\mathrm{t}}\begin{pmatrix}x_1 \\ x_2\end{pmatrix} = \begin{pmatrix}x_1\,\sigma_{11} & x_1\,\sigma_{12} \\ x_2\,\sigma_{21} & x_2\,\sigma_{22}\end{pmatrix}\begin{pmatrix}x_1\,\sigma_{11} & x_2\,\sigma_{21} \\ x_1\,\sigma_{12} & x_2\,\sigma_{22}\end{pmatrix}\\ &= \begin{pmatrix} x_1^2\,(\sigma_{11}^2 + \sigma_{12}^2) & x_1x_2\,(\sigma_{11}\sigma_{21} + \sigma_{12}\sigma_{22}) \\ x_1x_2\,(\sigma_{11}\sigma_{21} + \sigma_{12}\sigma_{22}) & x_2^2\,(\sigma_{21}^2 + \sigma_{22}^2)\end{pmatrix}, \qquad (6.86a)\\ \mathcal{L}^0 &= \sum_{k=1}^{2} a_k\,\partial_{x_k} + \frac{1}{2}\sum_{k,\alpha=1}^{2}\Sigma_{k\alpha}\,\partial_{x_k}\partial_{x_\alpha}\\ &= x_1\,\mu_1(x_1, x_2)\,\partial_{x_1} + x_2\,\mu_2(x_1, x_2)\,\partial_{x_2}\\ &\quad + \frac{1}{2}\Big(x_1^2\,(\sigma_{11}^2 + \sigma_{12}^2)\,\partial_{x_1}\partial_{x_1} + x_1x_2\,(\sigma_{11}\sigma_{21} + \sigma_{12}\sigma_{22})\,(\partial_{x_1}\partial_{x_2} + \partial_{x_2}\partial_{x_1})\\ &\qquad + x_2^2\,(\sigma_{21}^2 + \sigma_{22}^2)\,\partial_{x_2}\partial_{x_2}\Big), \qquad (6.86b)\\ \mathcal{L}^1 &= \sum_{k=1}^{2} b_{k1}\,\partial_{x_k} = x_1\,\sigma_{11}\,\partial_{x_1} + x_2\,\sigma_{21}\,\partial_{x_2}, \qquad (6.86c)\end{aligned}$$

$$\mathcal{L}^2 = \sum_{k=1}^{2} b_{k2}\,\partial_{x_k} = x_1\,\sigma_{12}\,\partial_{x_1} + x_2\,\sigma_{22}\,\partial_{x_2}, \tag{6.86d}$$

$$\mathcal{L}^0 a_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathcal{L}^0\big(x_1\,\mu_1(x_1,\ldots,x_d)\big) \quad \text{(further expansion omitted here)}, \tag{6.86e}$$

$$\mathcal{L}^0 a_2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathcal{L}^0\big(x_2\,\mu_2(x_1,\ldots,x_d)\big) \quad \text{(further expansion omitted here)}, \tag{6.86f}$$

$$\mathcal{L}^0 b_{11} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \sigma_{11}\,x_1\,\mu_1(x_1,x_2), \tag{6.86g}$$

$$\mathcal{L}^0 b_{12} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \sigma_{12}\,x_1\,\mu_1(x_1,x_2), \tag{6.86h}$$

$$\mathcal{L}^0 b_{21} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \sigma_{21}\,x_2\,\mu_2(x_1,x_2), \tag{6.86i}$$

$$\mathcal{L}^0 b_{22} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \sigma_{22}\,x_2\,\mu_2(x_1,x_2), \tag{6.86j}$$

$$\mathcal{L}^1 a_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1\,\sigma_{11}\,\big(\mu_1(x_1,x_2) + x_1\,\partial_{x_1}\mu_1(x_1,x_2)\big) + x_1x_2\,\sigma_{21}\,\partial_{x_2}\mu_1(x_1,x_2), \tag{6.86k}$$

$$\mathcal{L}^1 a_2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1x_2\,\sigma_{11}\,\partial_{x_1}\mu_2(x_1,x_2) + x_2\,\sigma_{21}\,\big(\mu_2(x_1,x_2) + x_2\,\partial_{x_2}\mu_2(x_1,x_2)\big), \tag{6.86l}$$

$$\mathcal{L}^1 b_{11} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1\,\sigma_{11}^2, \tag{6.86m}$$

$$\mathcal{L}^1 b_{12} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1\,\sigma_{11}\sigma_{12}, \tag{6.86n}$$

$$\mathcal{L}^1 b_{21} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_2\,\sigma_{21}^2, \tag{6.86o}$$

$$\mathcal{L}^1 b_{22} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_2\,\sigma_{21}\sigma_{22}, \tag{6.86p}$$

$$\mathcal{L}^2 a_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1\,\sigma_{12}\,\big(\mu_1(x_1,x_2) + x_1\,\partial_{x_1}\mu_1(x_1,x_2)\big) + x_1x_2\,\sigma_{22}\,\partial_{x_2}\mu_1(x_1,x_2), \tag{6.86q}$$

$$\mathcal{L}^2 a_2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1x_2\,\sigma_{12}\,\partial_{x_1}\mu_2(x_1,x_2) + x_2\,\sigma_{22}\,\big(\mu_2(x_1,x_2) + x_2\,\partial_{x_2}\mu_2(x_1,x_2)\big), \tag{6.86r}$$

$$\mathcal{L}^2 b_{11} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1\,\sigma_{11}\sigma_{12}, \tag{6.86s}$$

$$\mathcal{L}^2 b_{12} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1\,\sigma_{12}^2, \tag{6.86t}$$

$$\mathcal{L}^2 b_{21} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_2\,\sigma_{21}\sigma_{22}, \tag{6.86u}$$

$$\mathcal{L}^2 b_{22} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_2\,\sigma_{22}^2. \tag{6.86v}$$

We obtain the recursion by plugging the appropriate terms from above into (6.80b):

$$\begin{aligned}
\hat{L}_k^{(i+1)} &:= \hat{L}_k^{(i)} + \hat{L}_k^{(i)}\,\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big)\,h + \mathcal{L}^0 a_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big)\,\frac{h^2}{2} \\
&\quad + \Big(\hat{L}_{\tilde{k}}^{(i)}\,\hat{L}_k^{(i)}\,\sigma_{\tilde{k}1}\,\partial_{x_{\tilde{k}}}\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big) \\
&\qquad + \hat{L}_k^{(i)}\,\sigma_{k1}\,\Big(\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big) + \hat{L}_k^{(i)}\,\partial_{x_k}\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big)\Big)\Big)\,I_{10}^{(i)} \\
&\quad + \Big(\hat{L}_{\tilde{k}}^{(i)}\,\hat{L}_k^{(i)}\,\sigma_{\tilde{k}2}\,\partial_{x_{\tilde{k}}}\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big) \\
&\qquad + \hat{L}_k^{(i)}\,\sigma_{k2}\,\Big(\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big) + \hat{L}_k^{(i)}\,\partial_{x_k}\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big)\Big)\Big)\,I_{20}^{(i)} \\
&\quad + \hat{L}_k^{(i)}\,\big(\sigma_{k1}\,(\Delta W_1)^{(i)} + \sigma_{k2}\,(\Delta W_2)^{(i)}\big) + \Big(\sigma_{k1}\,I_{01}^{(i)} + \sigma_{k2}\,I_{02}^{(i)}\Big)\,\hat{L}_k^{(i)}\,\mu_k\big(\hat{L}_1^{(i)},\hat{L}_2^{(i)}\big) \\
&\quad + \hat{L}_k^{(i)}\,\sigma_{k1}^2\,I_{11}^{(i)} + \hat{L}_k^{(i)}\,\sigma_{k1}\sigma_{k2}\,\Big(I_{12}^{(i)} + I_{21}^{(i)}\Big) + \hat{L}_k^{(i)}\,\sigma_{k2}^2\,I_{22}^{(i)},
\end{aligned} \tag{6.87}$$

for $k = 1, 2$, where $\tilde{1} := 2$ and $\tilde{2} := 1$.

6.5.3 Commutativity Condition

As mentioned in Rem. 6.21(b), the presence of the mixed integrals $I_{l\beta}^{(i)}$, $l \neq m$, in (6.80) means that simplifications are usually required to make use of the scheme in practise.

Sometimes (certainly not always), the situation is such that the following commutativity condition (6.88) holds, which then facilitates a useful simplification of the scheme.

Definition 6.24. We say the (diffusion) coefficient function $b : \mathbb{R}^d \longrightarrow \mathbb{R}^{d\times m}$ satisfies the *commutativity condition* if, and only if, it has first partials and

$$\forall_{\substack{k=1,\ldots,d,\\ l,\beta=1,\ldots,m}} \quad \mathcal{L}^l b_{k\beta} = \mathcal{L}^\beta b_{kl}, \tag{6.88}$$

where the differential operators are according to (6.71b).

Remark 6.25. If the commutativity condition (6.88) holds, then

$$\forall_{\substack{i=0,\ldots,N-1,\\ k=1,\ldots,d}} \quad \sum_{l=1}^m \sum_{\beta=1}^m \mathcal{L}^\beta b_{kl}\,I_{\beta l}^{(i)} \overset{(6.88)}{=} \sum_{l=1}^m \mathcal{L}^l b_{kl}\,I_{ll}^{(i)} + \sum_{l=1}^m \sum_{\beta=l+1}^m \mathcal{L}^\beta b_{kl}\,\Big(I_{\beta l}^{(i)} + I_{l\beta}^{(i)}\Big). \tag{6.89}$$

Proposition 6.26. *If $(W_l)_{t\in[0,T]}$ and $(W_\beta)_{t\in[0,T]}$ denote components of an m-dimensional standard Brownian motion with drift 0 and covariance matrix* Id*, then*

$$\forall_{\substack{i=0,\ldots,N-1,\\ l,\beta=1,\ldots,m,\\ l\neq\beta}} \quad I_{l\beta}^{(i)} + I_{\beta l}^{(i)} = (\Delta W_l)^{(i)}\,(\Delta W_\beta)^{(i)}, \tag{6.90}$$

where we made use of the notation of (6.76d) *and* (6.78).

Proof. Exercise (hint: apply Itô's formula (C.9) with $Y_t := W_t$ und $f(t, x_1, \dots, x_m) := x_l \, x_\beta$, cf. [Gla04, p. 354]). ■

Using (6.89), (6.90), and (6.81) in (6.80), we obtain the following multidimensional second-order scheme, simplified via the commutativity condition:

$$\hat{X}^{(0)} := X_{\text{init}}, \tag{6.91a}$$

$$\begin{aligned}
\hat{X}_k^{(i+1)} := \hat{X}_k^{(i)} &+ a_k(\hat{X}^{(i)})\, h + \sum_{l=1}^{m} b_{kl}(\hat{X}^{(i)})\, (\Delta W_l)^{(i)} + \mathcal{L}^0 a_k(\hat{X}^{(i)})\, \frac{h^2}{2} \\
&+ \sum_{l=1}^{m} \left(\left(\mathcal{L}^l a_k(\hat{X}^{(i)}) - \mathcal{L}^0 b_{kl}(\hat{X}^{(i)}) \right) I_{l0}^{(i)} + \mathcal{L}^0 b_{kl}(\hat{X}^{(i)})\, h\, (\Delta W_l)^{(i)} \right) \\
\forall_{\substack{i=0,\dots,N-1,\\ k=1,\dots,d}} \quad &+ \sum_{l=1}^{m} \left(\mathcal{L}^l b_{kl}(\hat{X}^{(i)})\, \frac{\left((\Delta W_l)^{(i)}\right)^2 - h}{2} \right. \\
&\qquad \left. + \sum_{\beta=l+1}^{m} \mathcal{L}^\beta b_{kl}(\hat{X}^{(i)})\, (\Delta W_l)^{(i)}\, (\Delta W_\beta)^{(i)} \right).
\end{aligned} \tag{6.91b}$$

Remark 6.27. To actually implement (6.91), one applies Prop. 6.15 for each $l = 1, \dots, m$, yielding

$$\forall_{\substack{i=0,\dots,N-1,\\ l=1,\dots,m,\\ x\in\mathbb{R}}} \quad Z_l^{(i)} := \begin{pmatrix} (\Delta W_l)^{(i)} \\ I_{l0}^{(i)} \end{pmatrix} \Big| \{(W_l)_{t_i} = x\} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} h & \frac{h^2}{2} \\ \frac{h^2}{2} & \frac{h^3}{3} \end{pmatrix} \right), \tag{6.92}$$

where one can show that the independence of the $(W_l)_t$ for $l = 1, \dots, m$ implies the independence of the $Z_l^{(i)}$ for $l = 1, \dots, m$.

Theorem 6.28. *Assume the hypotheses of Th.* 6.3. *If* $E\big(\|X_{\text{init}}\|^i\big) < \infty$ *for each* $i \in \mathbb{N}$ *and* a, b *have no explicit time dependence, satisfying* $a, b \in C^6(\mathbb{R}^d)$, *where all derivatives are uniformly bounded, then, with respect to the strong solution* $(X_t)_{t\in[0,T]}$ *of* (6.1), *the discretization given by the multi-dimensional second-order scheme* (6.80) *has weak order of convergence* 2. *In particular, if, additionally, the commutativity condition* (6.88) *is satisfied, then the discretization given by the simplified multi-dimensional second-order scheme* (6.91) *has weak order of convergence* 2.

Proof. Since (6.80) is the same as [KP99, (14.2.6)], the statement of our theorem is included in the statement of [KP99, Th. 14.2.4] which, as remarked before, is proved as a special case of [KP99, Th. 14.5.2]. ■

Example 6.29. The commutativity condition is satisfied for the LIBOR market model of Ex. 6.18(b) and Ex. 6.23: For each $k = 1, \dots, d$, each $l, \beta = 1, \dots, m$, and each

$x \in \mathbb{R}^d$, we have

$$\mathcal{L}^l b_{k\beta}(x) = \sum_{k=1}^{d} b_{kl}(x)\, \partial_{x_k} b_{k\beta}(x) = \sum_{k=1}^{d} x_k\, \sigma_{kl}\, \partial_{x_k}(x_k\, \sigma_{k\beta}) = \sum_{k=1}^{d} x_k\, \sigma_{kl}\, \sigma_{k\beta},$$
$$\mathcal{L}^\beta b_{kl}(x) = \sum_{k=1}^{d} b_{k\beta}(x)\, \partial_{x_k} b_{kl}(x) = \sum_{k=1}^{d} x_k\, \sigma_{k\beta}\, \partial_{x_k}(x_k\, \sigma_{kl}) = \sum_{k=1}^{d} x_k\, \sigma_{kl}\, \sigma_{k\beta}.$$

6.5.4 Simplified Second-Order Scheme

For the simplified second-order scheme (6.91) of the previous section, we had to assume the commutativity condition (6.88). However, as it turns out, there exist simplicications of (6.80) that retain the weak order of convergence 2 even if (6.88) is not available.

These simplifications consist of the following replacements

$$\underset{\substack{i=0,\dots,N-1,\\ l=1,\dots,m}}{\forall} \quad I_{l0}^{(i)} \approx \frac{(\Delta W_l)^{(i)}\, h}{2}, \tag{6.93a}$$

$$\underset{\substack{i=0,\dots,N-1,\\ l,\beta=1,\dots,m,\\ l\neq\beta}}{\forall} \quad I_{l\beta}^{(i)} \approx \frac{(\Delta W_l)^{(i)}\, (\Delta W_\beta)^{(i)} - V_{l\beta}^{(i)}}{2}, \tag{6.93b}$$

where the random variables $V_{l\beta}$ satisfy

$$\underset{\substack{i=0,\dots,N-1,\\ l,\beta=1,\dots,m}}{\forall} \quad V_{l\beta}^{(i)} : \Omega \longrightarrow \{-h, h\}, \tag{6.94a}$$

$$\underset{\substack{i=0,\dots,N-1,\\ l,\beta=1,\dots,m,\\ l\neq\beta}}{\forall} \quad V_{l\beta}^{(i)} = -V_{\beta l}^{(i)}, \tag{6.94b}$$

$$\underset{\substack{i=0,\dots,N-1,\\ l,\beta=1,\dots,m,\\ l\neq\beta}}{\forall} \quad P\{V_{l\beta}^{(i)} = h\} = P\{V_{l\beta}^{(i)} = -h\} = \frac{1}{2}, \tag{6.94c}$$

$$\underset{\substack{i=0,\dots,N-1,\\ l,\beta=1,\dots,m}}{\forall} \quad V_{ll}^{(i)} \equiv h, \tag{6.94d}$$

$$\left(V_{l\beta}^{(i)}\right)_{l<\beta} \quad \text{is independent and independent from} \quad \left((\Delta W_l)^{(i)}\right)_{l=1,\dots,m}. \tag{6.94e}$$

Note that (6.93a) is analogous to the replacements that were used in the 1-dimensional case to obtain (6.66) from (6.52). For explanations and justifications regarding the use of (6.93), we refer to [Gla04, p. 355f] and [KP99, p. 467].

Using (6.93), (6.94d), and (6.81) in (6.80), we obtain the following simplified multi-

dimensional second-order scheme:

$$\hat{X}^{(0)} := X_{\text{init}}, \tag{6.95a}$$

$$\begin{aligned}\hat{X}_k^{(i+1)} := \hat{X}_k^{(i)} &+ a_k(\hat{X}^{(i)})\,h + \sum_{l=1}^{m} b_{kl}(\hat{X}^{(i)})\,(\Delta W_l)^{(i)} + \mathcal{L}^0 a_k(\hat{X}^{(i)})\,\frac{h^2}{2} \\ &+ \frac{1}{2}\sum_{l=1}^{m}\left(\mathcal{L}^l a_k(\hat{X}^{(i)}) + \mathcal{L}^0 b_{kl}(\hat{X}^{(i)})\right)(\Delta W_l)^{(i)}\,h \\ &+ \frac{1}{2}\sum_{l=1}^{m}\sum_{\beta=1}^{m}\mathcal{L}^\beta b_{kl}(\hat{X}^{(i)})\left((\Delta W_l)^{(i)}\,(\Delta W_\beta)^{(i)} - V_{l\beta}^{(i)}\right).\end{aligned} \tag{6.95b}$$

$$\underset{\substack{i=0,\dots,N-1,\\ k=1,\dots,d}}{\forall}$$

Theorem 6.30. *Assume the hypotheses of Th.* 6.3. *If* $E\big(\|X_{\text{init}}\|^i\big) < \infty$ *for each* $i \in \mathbb{N}$ *and* a, b *have no explicit time dependence, satisfying* $a, b \in C^6(\mathbb{R}^d)$*, where all derivatives are uniformly bounded, then, with respect to the strong solution* $(X_t)_{t\in[0,T]}$ *of* (6.1)*, the discretization given by the simplified multi-dimensional second-order scheme* (6.95) *has weak order of convergence* 2*.*

Proof. Since (6.95) is the same as [KP99, (14.2.7)], the above statement is, once more, included in the statement of [KP99, Th. 14.2.4]. ■

A Measure Theory

A.1 σ-Algebras

A.1.1 σ-Algebra, Measurable Space

Notation A.1. For each set Ω, let $\mathcal{P}(\Omega)$ denote its power set.

Definition A.2. Let Ω be a set. A collection of subsets $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is called a *σ-algebra* on Ω if, and only if, it satisfies the following three conditions:

(a) $\emptyset \in \mathcal{A}$.

(b) If $A \in \mathcal{A}$, then $X \setminus A \in \mathcal{A}$.

(c) If $(A_n)_{n\in\mathbb{N}}$ is a sequence of sets in $\mathcal{A}$, then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

If $\mathcal{A}$ is a σ-algebra on Ω, then the pair $(\Omega, \mathcal{A})$ is called a *measurable space*. The sets $A \in \mathcal{A}$ are called *$\mathcal{A}$-measurable* (or merely *measurable* if $\mathcal{A}$ is understood).

A.1.2 Inverse Image, Trace

Proposition A.3. *Consider sets Ω, Ω', and a map $f : \Omega \longrightarrow \Omega'$. If $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra on Ω, then $\mathcal{B} := \{B \subseteq \Omega' : f^{-1}(B) \in \mathcal{A}\}$ is a σ-algebra on Ω'.*

Proof. Since $\emptyset = f^{-1}(\emptyset) \in \mathcal{A}$, we have $\emptyset \in \mathcal{B}$. If $B \in \mathcal{B}$, then $\Omega' \setminus B \in \mathcal{B}$, since $f^{-1}(\Omega' \setminus B) = \Omega \setminus f^{-1}(B) \in \mathcal{A}$, as $\Omega \setminus f^{-1}(B) \in \mathcal{A}$ due to $\mathcal{A}$ being a σ-algebra. Finally, if $(B_n)_{n\in\mathbb{N}}$ is a sequence of sets in $\mathcal{B}$, then

$$f^{-1}\left(\bigcup_{n\in\mathbb{N}} B_n\right) = \bigcup_{n\in\mathbb{N}} f^{-1}(B_n) \in \mathcal{A}, \tag{A.1}$$

as $\mathcal{A}$ is a σ-algebra. Thus, $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{B}$, which completes the proof that $\mathcal{B}$ is a σ-algebra on Ω'. ■

Proposition A.4. *Consider sets Ω, Ω', and a map $f : \Omega \longrightarrow \Omega'$. If $\mathcal{A} \subseteq \mathcal{P}(\Omega')$ is a σ-algebra on Ω', then $f^{-1}(\mathcal{A}) = \{f^{-1}(A) : A \in \mathcal{A}\}$ is a σ-algebra on Ω.*

Proof. As $\emptyset = f^{-1}(\emptyset)$, $\emptyset \in f^{-1}(\mathcal{A})$. If $A \in \mathcal{A}$, then $\Omega \setminus f^{-1}(A) = f^{-1}(\Omega' \setminus A) \in f^{-1}(\mathcal{A})$ since $\Omega' \setminus A \in \mathcal{A}$ due to $\mathcal{A}$ being a σ-algebra. Finally, if $(A_n)_{n\in\mathbb{N}}$ is a sequence of sets in $\mathcal{A}$, then, as $\mathcal{A}$ is a σ-algebra, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$. Then

$$\bigcup_{n\in\mathbb{N}} f^{-1}(A_n) = f^{-1}\left(\bigcup_{n\in\mathbb{N}} A_n\right) \in f^{-1}(\mathcal{A}), \tag{A.2}$$

completing the proof that $f^{-1}(\mathcal{A})$ is a σ-algebra on Ω. ■

Definition A.5. Let Ω be a set, and let $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ be a collection of subsets. If $B \subseteq \Omega$, then let $\mathcal{E}|B := \{A \cap B : A \in \mathcal{E}\}$ denote the *trace* of $\mathcal{E}$ on B, also known as the restriction of $\mathcal{E}$ to B.

Proposition A.6. *Consider sets Ω, B, such that $B \subseteq \Omega$. If $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra on Ω, then the trace $\mathcal{A}|B$ is a σ-algebra on B.*

Proof. Consider the canonical inclusion map $f : B \longrightarrow \Omega$, $f(a) = a$. Note that, for each $\mathcal{A} \subseteq \mathcal{P}(\Omega)$, one has

$$f^{-1}(\mathcal{A}) = \{f^{-1}(A) : A \in \mathcal{A}\} = \{A \cap B : A \in \mathcal{A}\} = \mathcal{A}|B, \tag{A.3}$$

such that the proposition follows immediately from Prop. A.4. ■

A.1.3 Intersection, Generated σ-Algebra

Proposition A.7. *Let Ω be a set. Each intersection of σ-algebras on Ω is again a σ-algebra on Ω. More precisely, if $\emptyset \neq I$ is an index set and $(\mathcal{A}_i)_{i\in I}$ is a family of σ-algebras on Ω, then $\bigcap_{i\in I} \mathcal{A}_i$ is a σ-algebra on Ω.*

Proof. Let $\mathcal{A} := \bigcap_{i\in I} \mathcal{A}_i$. Since $\emptyset \in \mathcal{A}_i$ for each $i \in I$, we have $\emptyset \in \mathcal{A}$. If $A \in \mathcal{A}$, then $A \in \mathcal{A}_i$ for each $i \in I$, implying $\Omega \setminus A \in \mathcal{A}_i$ for each $i \in I$, implying $\Omega \setminus A \in \mathcal{A}$. If $(A_n)_{n\in\mathbb{N}}$ is a sequence of sets in $\mathcal{A}$, then $A_n \in \mathcal{A}_i$ for each $i \in I$ and each $n \in \mathbb{N}$. In consequence $A := \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}_i$ for each $i \in I$, implying $A \in \mathcal{A}$. ■

Definition and Remark A.8. Let Ω be a set. If $\mathcal{E}$ is a collection of subsets of Ω, i.e. $\mathcal{E} \subseteq \mathcal{P}(\Omega)$, then let $\sigma_\Omega(\mathcal{E})$ denote the intersection of all σ-algebras on Ω that are supersets of $\mathcal{E}$ (i.e. that contain all the sets in $\mathcal{E}$). According to Prop. A.7, $\sigma_\Omega(\mathcal{E})$ is a σ-algebra on Ω. Obviously, it is the smallest σ-algebra on Ω containing $\mathcal{E}$. It is, thus, called the σ-algebra *generated* by $\mathcal{E}$.

—

The following Th. A.9 can be seen as a generalization of Prop. A.4.

Theorem A.9. *The forming of generated σ-algebras commutes with the forming of inverse images: If Ω, Ω' are sets, $f : \Omega \longrightarrow \Omega'$, and $\mathcal{F} \subseteq \mathcal{P}(\Omega')$, then*

$$\sigma_\Omega\big(f^{-1}(\mathcal{F})\big) = f^{-1}\big(\sigma_{\Omega'}(\mathcal{F})\big). \tag{A.4}$$

Proof. Since $\mathcal{F} \subseteq \sigma_{\Omega'}(\mathcal{F})$, one has $f^{-1}(\mathcal{F}) \subseteq f^{-1}(\sigma_{\Omega'}(\mathcal{F}))$, which implies the $\subseteq$-part of (A.4), as $f^{-1}\big(\sigma_{\Omega'}(\mathcal{F})\big)$ is a σ-algebra by Prop. A.4. To prove the $\supseteq$-part of (A.4), define

$$\mathcal{B} := \Big\{B \subseteq \Omega' :\ f^{-1}(B) \in \sigma_\Omega\big(f^{-1}(\mathcal{F})\big)\Big\}. \tag{A.5}$$

Then $\mathcal{F} \subseteq \mathcal{B}$, $\mathcal{B}$ is a σ-algebra by Prop. A.3, implying $\sigma_{\Omega'}(\mathcal{F}) \subseteq \mathcal{B}$. This, in turn, yields $f^{-1}\big(\sigma_{\Omega'}(\mathcal{F})\big) \subseteq f^{-1}(\mathcal{B}) \subseteq \sigma_\Omega\big(f^{-1}(\mathcal{F})\big)$, which is precisely the $\supseteq$-part of (A.4), completing the proof. ■

Corollary A.10. *Let Ω be a set, let $\mathcal{A}$ be a σ-algebra on Ω, $B \subseteq \Omega$, and $\mathcal{E} \subseteq \mathcal{A}$. If $\mathcal{A} = \sigma_\Omega(\mathcal{E})$, then $\mathcal{A}|B = \sigma_B(\mathcal{E}|B)$.*

Proof. One merely has to apply Th. A.9 to the canonical inclusion map $f : B \longrightarrow \Omega$, $f(a) = a$:

$$\mathcal{A}|B \overset{(A.3)}{=} f^{-1}(\mathcal{A}) = f^{-1}\big(\sigma_\Omega(\mathcal{E})\big) \overset{(A.4)}{=} \sigma_B\big(f^{-1}(\mathcal{E})\big) \overset{(A.3)}{=} \sigma_B(\mathcal{E}|B), \tag{A.6}$$

establishing the case. ■

Notation A.11. Let Ω be a set, and let $\mathcal{E} \subseteq \mathcal{P}(\Omega)$. Define

$$\mathcal{E}^\bullet := \left\{\bigcup_{n\in\mathbb{N}} A_n :\ A_n \in \mathcal{E} \text{ or } (\Omega \setminus A_n) \in \mathcal{E} \text{ for each } n \in \mathbb{N}\right\}.$$

Theorem A.12. *Let Ω be a set and $\mathcal{E} \subseteq \mathcal{P}(\Omega)$. Moreover, let ω_1 denote the smallest uncountable ordinal. Let $\mathcal{E}_0 := \mathcal{E} \cup \{\emptyset\}$, and, using the notation from Not.* A.11*, for each $0 < \alpha \in \omega_1$, define*

$$\mathcal{E}_\alpha := \left(\bigcup_{\beta\in\alpha} \mathcal{E}_\beta\right)^\bullet. \tag{A.7}$$

It then holds that

$$\sigma_\Omega(\mathcal{E}) = \bigcup_{\alpha\in\omega_1} \mathcal{E}_\alpha. \tag{A.8}$$

Proof. See [Els07, Sec. I.4.1]. ■

A.1.4 Borel σ-Algebra

Definition A.13. Let (Ω, τ) be a topological space. Then $\sigma_\Omega(\tau)$, i.e. the σ-algebra generated by the open sets of Ω, is called the *Borel* σ-algebra on (Ω, τ) (or on Ω if the topology τ is understood).

Corollary A.14. *Let (Ω, τ) be a topological space and let $\mathcal{B}$ denote the Borel σ-algebra on (Ω, τ). If $B \subseteq \Omega$, then $\mathcal{B}|B$ is the Borel σ-algebra on B with respect to the relative topology on B, i.e. $\mathcal{B}|B = \sigma_B(\tau|B)$.*

Proof. Since $\mathcal{B} = \sigma_\Omega(\tau)$, the statement is a special case of Cor. A.10. ■

Notation A.15. For each $n \in \mathbb{N}$, let $\mathcal{B}^n$ denote the Borel σ-algebra on $\mathbb{R}^n$ (with respect to the usual norm topology on $\mathbb{R}^n$). By a slight abuse of notation, for each $A \subseteq \mathbb{R}^n$, we also write $(A, \mathcal{B}^n)$ to denote the measurable space consisting of A and $\mathcal{B}^n|A$.

Remark A.16. It is often useful to know that the Borel σ-algebra $\mathcal{B}^n$ on $\mathbb{R}^n$ is also generated by the closed sets in $\mathbb{R}^n$, $n \in \mathbb{N}$, or by the set of all closed (respectively, open, half-open with upper endpoint included, half-open with lower endpoint included) n-dimensional intervals in $\mathbb{R}^n$, and, in each case, even by the (countable!) set of all such intervals with rational endpoints.

A.2 Measure Spaces

A.2.1 Measures and Measure Spaces

Definition A.17. Let $(\Omega, \mathcal{A})$ be a measurable space.

(a) A map $\mu : \mathcal{A} \longrightarrow [0, \infty]$ is called a *measure* on $(\Omega, \mathcal{A})$ (or on Ω is $\mathcal{A}$ is understood) if, and only if, μ satisfies the following conditions (i) and (ii):

(i) $\mu(\emptyset) = 0$.

(ii) μ is σ-*additive*, i.e., if $(A_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{A}$ consisting of pairwise disjoint sets, then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n). \tag{A.9}$$

If μ is a measure on $(\Omega, \mathcal{A})$, then the triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*. In the context of measure spaces, the sets $A \in \mathcal{A}$ are sometimes called μ-*measurable* instead of $\mathcal{A}$-measurable.

(b) Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. The measure μ is called *finite* or *bounded* if, and only if, $\mu(\Omega) < \infty$; it is called *σ-finite* if, and only if, there exists a sequence $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{A}$ such that $\Omega = \bigcup_{n=1}^{\infty} A_i$ and $\mu(A_n) < \infty$ for each $n \in \mathbb{N}$.

Proposition A.18. *If $(\Omega, \mathcal{A}, \mu)$ is a measure space and $A \in \mathcal{A}$, then the restriction $\nu := \mu\restriction_{\mathcal{A}|A}$ is a measure on $(A, \mathcal{A}|A)$.*

Proof. According to Prop. A.6, $\mathcal{A}|A$ is a σ-algebra on A, and $A \in \mathcal{A}$ implies $\mathcal{A}|A \subseteq \mathcal{A}$ such that ν is actually defined for each $A' \in \mathcal{A}|A$. Moreover, $\nu(\emptyset) = \mu(\emptyset) = 0$, and, as ν is the restriction of μ, ν inherits the σ-additivity from μ, proving that ν is a measure. ■

A.2.2 Null Sets, Completion

Definition A.19. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Then $N \in \mathcal{A}$ is called a *μ-null set* (or merely a *null set* if μ is understood) if, and only if, $\mu(N) = 0$.

Definition A.20. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. The measure μ is called *complete* if, and only if, every subset of a μ-null set is μ-measurable, i.e. is itself a μ-null set.

Theorem A.21. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and let $\mathcal{N}$ denote the collection of all subsets of μ-null sets. Define*

$$\tilde{\mathcal{A}} := \{A \cup N : A \in \mathcal{A},\ N \in \mathcal{N}\}, \tag{A.10a}$$

$$\tilde{\mu} : \tilde{\mathcal{A}} \longrightarrow [0, \mu], \quad \tilde{\mu}(A \cup N) := \mu(A). \tag{A.10b}$$

Then $\tilde{\mathcal{A}}$ is a σ-algebra on Ω, $\tilde{\mu}$ is well-defined by (A.10b) *and constitutes a complete measure on Ω. Moreover, $\tilde{\mu}$ is the smallest complete measure extending μ in the sense that each complete measure extending μ to a σ-algebra on Ω containing $\mathcal{A}$ must be an extension of $\tilde{\mu}$.*

Proof. See, e.g., [Els07, Sec. II.6.3]. ■

Definition A.22. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Then the measure space $(\Omega, \tilde{\mathcal{A}}, \tilde{\mu})$ provided by (A.10) is called the *completion* of $(\Omega, \mathcal{A}, \mu)$ and $\tilde{\mu}$ is called the *completion* of μ.

Proposition A.23. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $\mathcal{N}$ be the collection of all subsets of μ-null sets, and $B \in \mathcal{A}$. Then the completion of the trace is the trace of the completion, i.e.*

$$\{(A \cap B) \cup (N \cap B) : A \in \mathcal{A},\ N \in \mathcal{N}\} = \{(A \cup N) \cap B : A \in \mathcal{A},\ N \in \mathcal{N}\}. \tag{A.11}$$

Proof. Since $(A \cap B) \cup (N \cap B) = (A \cup N) \cap B$, the equality in (A.11) is clear. ■

A.2.3 Uniqueness Theorem

Definition A.24. Let Ω be a set. A collection of subsets $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ is called a $\cap$-*stable* if, and only if, it satisfies

$$A, B \in \mathcal{C} \quad \Rightarrow \quad A \cap B \in \mathcal{C}. \tag{A.12}$$

Theorem A.25 (Uniqueness of Measures)**.** *Let $(\Omega, \mathcal{A})$ be a measurable space, let $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ be a $\cap$-stable generator of $\mathcal{A}$, and let $\mu, \nu : \mathcal{A} \longrightarrow [0, \infty]$ be measures. If μ, ν satisfy*

(i) $\mu\restriction_{\mathcal{C}} = \nu\restriction_{\mathcal{C}}$, *i.e.* $\mu(A) = \nu(A)$ *for each* $A \in \mathcal{C}$;

(ii) $\mu(A_n) = \nu(A_n) < \infty$ *for some sequence* $(A_n)_{n\in\mathbb{N}}$ *with* $\Omega = \bigcup_{n\in\mathbb{N}} A_n$;

then the measures are equal, $\mu = \nu$.

Proof. See, e.g., [Els07, Th. II.5.6]. ■

A.2.4 Lebesgue-Borel and Lebesgue Measure on $\mathbb{R}^n$

Definition and Remark A.26. Let $n \in \mathbb{N}$.

(a) Let

$$\mathcal{H}^n := \{\emptyset\} \cup \big\{ [a, b[\colon\ a, b \in \mathbb{R}^n,\ a < b \big\} \tag{A.13}$$

denote the set of all n-dimensional half-open intervals with left boundary included (plus the empty set). We already noted in Rem. A.16 that $\mathcal{H}^n$ is a generator for $\mathcal{B}^n$. Clearly, $\mathcal{H}^n$ is also $\cap$-stable. Thus, from Th. A.25, we obtain the existence of a unique measure $\beta_n : \mathcal{B}^n \longrightarrow [0, \infty]$, satisfying

$$\beta_n\big([a, b[\big) = \prod_{i=1}^{n} (b_i - a_i) \quad \text{for each } a, b \in \mathbb{R}^n \text{ with } a < b. \tag{A.14}$$

This unique measure is called the *Lebesgue-Borel measure* on $\mathbb{R}^n$ or n-dimensional Lebesgue-Borel measure (also just called n-dimensional *Borel measure* in the literature, but we will call it Lebesgue-Borel measure to distinguish it from general Borel measures and to emphasize its relation to the Lebesgue measure introduced in (b) below.

(b) The completion as given by Def. A.22 of the Lebesgue-Borel measure β_n is called the *Lebesgue measure* on $\mathbb{R}^n$ (or n-dimensional Lebesgue measure) and is denoted by λ_n. In particular, $\lambda_n\restriction_{\mathcal{B}^n} = \beta_n$ and one did not even need to use the symbol β_n at all. However, β_n is used where one wants to emphasize one is merely considering the smaller σ-algebra $\mathcal{B}^n$.

(c) A set $A \subseteq \mathbb{R}^n$ is simply called *measurable* if, and only if, it is λ_n-measurable.

A.3 Measurable Maps

A.3.1 Definition, Composition

Definition A.27. Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces.

(a) A map $f : \Omega \longrightarrow \Omega'$ is called $\mathcal{A}$-$\mathcal{A}'$-measurable if, and only if,

$$f^{-1}(B) \in \mathcal{A} \quad \text{for each } B \in \mathcal{A}'. \tag{A.15}$$

Proposition A.28. *Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces. If $\mathcal{C}' \subseteq \mathcal{P}(\Omega')$ is a generator of $\mathcal{A}'$, then a map $f : \Omega \longrightarrow \Omega'$ is $\mathcal{A}$-$\mathcal{A}'$-measurable if, and only if,*

$$f^{-1}(B) \in \mathcal{A} \quad \textit{for each } B \in \mathcal{C}'. \tag{A.16}$$

Proof. Since $\mathcal{C}' \subseteq \mathcal{A}'$, (A.15) implies (A.16). To prove the converse, define

$$\mathcal{Q}' := \big\{B \subseteq \Omega' : f^{-1}(B) \in \mathcal{A}\big\}, \tag{A.17}$$

and note that $\mathcal{Q}'$ is a σ-algebra on Ω'. Now, (A.16) implies $\mathcal{C}' \subseteq \mathcal{Q}'$. Thus, as $\mathcal{Q}'$ is a σ-algebra, we obtain $\mathcal{A}' = \sigma_{\Omega'}(\mathcal{C}') \subseteq \mathcal{Q}'$, completing the proof that f is $\mathcal{A}$-$\mathcal{A}'$-measurable. ■

Example A.29. (a) Constant maps are always measurable. More precisely, if $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ are measurable spaces, $f : \Omega \longrightarrow \Omega'$, $f \equiv c$, $c \in \Omega'$, then, for each $B \in \mathcal{A}'$, there are precisely two possibilities: $c \in \mathcal{A}'$, implying $f^{-1}(B) = \Omega \in \mathcal{A}$; $c \notin \mathcal{A}'$, implying $f^{-1}(B) = \emptyset \in \mathcal{A}$. Thus, f is $\mathcal{A}$-$\mathcal{A}'$-measurable.

(b) Continuous maps are always Borel measurable: Let (Ω, τ) and (Ω', τ') be topological spaces with corresponding Borel σ-algebras $\sigma(\tau)$ and $\sigma(\tau')$. If $f : \Omega \longrightarrow \Omega'$ is continuous, then $f^{-1}(U) \in \tau \subseteq \sigma(\tau)$ for each $U \in \tau'$ and Prop. A.28 implies f is $\sigma(\tau)$-$\sigma(\tau')$ measurable.

(c) If $M \subseteq \mathbb{R}^n$ and $f : M \longrightarrow \mathbb{R}^m$ is continuous ($n, m \in \mathbb{N}$), then f is Borel measurable, i.e. $\mathcal{B}^n$-$\mathcal{B}^m$-measurable, which is merely an important special case of (b) (cf. Not. A.15 and Cor. A.14).

Proposition A.30. *The composition of measurable maps is measurable – more precisely, if $(\Omega, \mathcal{A})$, $(\Omega', \mathcal{A}')$, and $(\Omega'', \mathcal{A}'')$ are measurable spaces, $f : \Omega \longrightarrow \Omega'$ is $\mathcal{A}$-$\mathcal{A}'$-measurable, and $g : \Omega' \longrightarrow \Omega''$ is $\mathcal{A}'$-$\mathcal{A}''$-measurable, then $g \circ f$ is $\mathcal{A}$-$\mathcal{A}''$-measurable.*

Proof. For each $B \in \mathcal{A}''$, we have $A := (g \circ f)^{-1}(B) = f^{-1}\big(g^{-1}(B)\big)$. Then $g^{-1}(B) \in \mathcal{A}'$, since g is $\mathcal{A}'$-$\mathcal{A}''$-measurable. In consequence, $A \in \mathcal{A}$, since f is $\mathcal{A}$-$\mathcal{A}'$-measurable, proving the $\mathcal{A}$-$\mathcal{A}''$ measurability of $g \circ f$. ■

A.3.2 Generated σ-Algebra, Pushforward Measure

Definition A.31. Let Ω be a set, $(\Omega_i, \mathcal{A}_i)_{i\in I}$ a family of measurable spaces, and $(f_i : \Omega \longrightarrow \Omega_i)_{i\in I}$ a family of maps, each f_i defined on Ω and mapping into Ω_i. Then

$$\sigma\left((f_i)_{i\in I}\right) := \sigma\left(\bigcup_{i\in I} f_i^{-1}(\mathcal{A}_i)\right) \tag{A.18}$$

is called the σ-algebra *generated* by the family of maps $(f_i)_{i\in I}$ and by the family of measurable spaces $(\Omega_i, \mathcal{A}_i)_{i\in I}$ (note that the dependence on the measurable spaces is suppressed in the notation in (A.18) – the measurable spaces are supposed to be understood from the context). If $I = \{i_1, \ldots, i_n\}$, $n \in \mathbb{N}$, is finite, we also use the notation

$$\sigma(f_{i_1}, \ldots, f_{i_n}) := \sigma\left((f_i)_{i\in I}\right). \tag{A.19}$$

Remark A.32. (a) Clearly, in the context of Def. A.31, $\sigma\left((f_i)_{i\in I}\right)$ is the smallest σ-algebra on Ω with respect to which all the maps f_i are measurable.

(b) If, in the context of Def. A.31, I has just one element, $I = \{i_0\}$, then, letting $f := f_{i_0}$, according to Prop. A.4, one has $\sigma(f) = f^{-1}(\mathcal{A}_{i_0})$. Moreover, if $\mathcal{A}$ is an arbitrary σ-algebra on Ω, then f is $\mathcal{A}$-$\mathcal{A}_{i_0}$-measurable if, and only if, $\sigma(f) \subseteq \mathcal{A}$.

Proposition A.33. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $(\Omega', \mathcal{A}')$ a measurable space. If $f : \Omega \longrightarrow \Omega'$ is $\mathcal{A}$-$\mathcal{A}'$-measurable, then*

$$\mu_f : \mathcal{A}' \longrightarrow [0, \infty], \quad \mu_f(B) := \mu\big(f^{-1}(B)\big), \tag{A.20}$$

defines a measure on $\mathcal{A}$.

Proof. $\mu_f(\emptyset) = \mu(\emptyset) = 0$ follows without difficulty, and, if $(B_n)_{n\in\mathbb{N}}$ is a sequence of pairwise disjoint sets in $\mathcal{A}'$, then $(f^{-1}(B_n))_{n\in\mathbb{N}}$ is a sequence of pairwise disjoint sets in $\mathcal{A}$, implying

$$\begin{aligned}\mu_f\left(\bigcup_{n=1}^{\infty} B_n\right) &= \mu\left(f^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right)\right) = \mu\left(\bigcup_{n=1}^{\infty} f^{-1}(B_n)\right) = \sum_{n=1}^{\infty} \mu\big(f^{-1}(B_n)\big) \\ &= \sum_{n=1}^{\infty} \mu_f(B_n),\end{aligned} \tag{A.21}$$

verifying σ-additivity of μ_f and completing the proof. ■

Definition A.34. We remain in the context of Prop. A.33. The measure μ_f is called a *pushforward* measure – the measure μ is pushed forward from $\mathcal{A}$ to $\mathcal{A}'$ by f. Alternatively to μ_f, one also finds the notation $f(\mu) := \mu_f$.

Proposition A.35. *The forming of push forward measures commutes with the composition of maps: If $(\Omega, \mathcal{A}, \mu)$ is a measure space, $(\Omega', \mathcal{A}')$ and $(\Omega'', \mathcal{A}'')$ are measurable spaces, $f : \Omega \longrightarrow \Omega'$ is $\mathcal{A}$-$\mathcal{A}'$-measurable, and $g : \Omega' \longrightarrow \Omega''$, then*

$$\mu_{g\circ f} = (\mu_f)_g \quad \text{or} \quad (g \circ f)(\mu) = g(f(\mu)). \tag{A.22}$$

Proof. Since (A.22) claims the equality of the two maps $\mu_{g\circ f}$ and $(\mu_f)_g$, we have to verify $\mu_{g\circ f}(C) = (\mu_f)_g(C)$ for each $C \in \mathcal{A}''$. To this end, for each $C \in \mathcal{A}''$, one calculates

$$\mu_{g\circ f}(C) = \mu\big((g\circ f)^{-1}(C)\big) = \mu\big(f^{-1}\big(g^{-1}(C)\big)\big) = \mu_f\big(g^{-1}(C)\big) = (\mu_f)_g(C), \tag{A.23}$$

thereby establishing the case. ■

A.3.3 Review: Order on, Arithmetic in, and Topology of $\overline{\mathbb{R}}$

The material of the present section should mostly be familiar from a previous class on Calculus or Advanced Calculus. Here, it is included for the reader's convenience and for the convenience of easy reference.

Notation A.36. By $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$, we denote the set of *extended real numbers*.

—

The extended real numbers are useful in many contexts, an important example being the Lebesgue integral of $\overline{\mathbb{R}}$-Valued measurable maps of the following Sec. A.4.1. We now review the respective definitions of order, arithmetic, and topology on $\overline{\mathbb{R}}$, which, as far as possible, constitute extensions of the respective definitions on $\mathbb{R}$.

Definition A.37. (a) The (total) order on $\mathbb{R}$ is extended to $\overline{\mathbb{R}}$ by setting $-\infty < a < \infty$ for each $a \in \mathbb{R}$. The absolute value function is extended from $\mathbb{R}$ to $\overline{\mathbb{R}}$ by defining $|\infty| := |-\infty| := \infty$.

(b) Addition, subtraction, and multiplication are extended from $\mathbb{R}$ to $\overline{\mathbb{R}}$ by defining:

$$\forall_{a\in\mathbb{R}} \quad a + (\pm\infty) := (\pm\infty) + a := \pm\infty, \tag{A.24a}$$

$$\forall_{a\in\mathbb{R}} \quad a - (\pm\infty) := -(\pm\infty) + a := \mp\infty, \tag{A.24b}$$

$$\infty + \infty := \infty, \quad -\infty + (-\infty) := -\infty, \quad -(\pm\infty) := \mp\infty, \tag{A.24c}$$

$$\forall_{a\in\overline{\mathbb{R}}} \quad a \cdot (\pm\infty) := (\pm\infty) \cdot a := \begin{cases} \pm\infty & \text{for } a \in]0, \infty], \\ \mp\infty & \text{for } a \in [-\infty, 0[, \end{cases} \tag{A.24d}$$

$$0 \cdot (\pm\infty) := (\pm\infty) \cdot 0 := 0, \tag{A.24e}$$

$$\infty - \infty := -\infty + \infty := 0. \tag{A.24f}$$

Remark A.38. The definitions of (A.24a) – (A.24d) are required to naturally extend properties and rules from $\mathbb{R}$ to $\overline{\mathbb{R}}$ as far as possible (for example, (A.24a) guarantees that, for each $a \in \mathbb{R}$, the map $x \mapsto a + x$ is continuous on $\overline{\mathbb{R}}$ with respect to the topology defined in Def. A.39 below). The definitions of (A.24e) are required in the context of integration theory (see Sec. A.4.1 below), where, e.g., the integral of an ∞-valued function over a set of measure 0 has value 0. The definitions of (A.24f) can be seen as arbitrary. *Caveat:* Some familiar rules of arithmetic do *not* hold on all of $\overline{\mathbb{R}}$: Addition is not associative on $\overline{\mathbb{R}}$ and distributivity does not hold in general (however, addition and multiplication are still commutative, multiplication is associative, and the restriction of addition to $]-\infty, \infty]$ as well as to $[-\infty, \infty[$ is associative).

Definition A.39. The set of *open intervals* $\overline{\mathcal{I}}$ in $\overline{\mathbb{R}}$ is defined to consist of $\overline{\mathbb{R}}$ plus all open intervals in $\mathbb{R}$ plus all intervals of the form $[-\infty, a[$ or $]a, \infty]$, $a \in \overline{\mathbb{R}}$:

$$\overline{\mathcal{I}} := \big\{\overline{\mathbb{R}}\big\} \cup \big\{]a, b[\colon\ a, b \in \overline{\mathbb{R}},\ a < b\big\} \cup \big\{[-\infty, a[\colon\ a \in \mathbb{R} \cup \{\infty\}\big\} \cup \big\{]a, \infty] \colon\ a \in \mathbb{R} \cup \{-\infty\}\big\}. \tag{A.25}$$

Then the (standard) topology on $\overline{\mathbb{R}}$ is defined by calling a set $O \subseteq \overline{\mathbb{R}}$ *open* if, and only if, each element $x \in O$ is contained in an open interval $I \in \overline{\mathcal{I}}$ that is contained in O, i.e. $x \in I \subseteq O$. In other words, $\overline{\mathcal{I}}$ is defined to be a *local base* (also known as a *neighborhood basis*) for the topology on $\overline{\mathbb{R}}$.

Remark A.40. $\overline{\mathbb{R}}$ with the topology defined in Def. A.39 constitutes a so-called *compactification* of $\mathbb{R}$, i.e. it is a compact topological space such that the topology on $\mathbb{R}$ is recovered as the relative topology when considering $\mathbb{R}$ as a subset of $\overline{\mathbb{R}}$.

Notation A.41. Let $\overline{\mathcal{B}}$ denote the Borel sets on $\overline{\mathbb{R}}$, i.e. the Borel sets with respect to the topology defined in Def. A.39.

Lemma A.42. *For the Borel sets on* $\overline{\mathbb{R}}$*, we have the following identities:*

$$\overline{\mathcal{B}} = \big\{B \cup E \colon\ B \in \mathcal{B}^1,\ E \subseteq \{-\infty, \infty\}\big\}, \tag{A.26a}$$

$$\overline{\mathcal{B}}|\mathbb{R} = \mathcal{B}^1. \tag{A.26b}$$

Proof. Let $\mathcal{A}$ denote the right-hand side of (A.26a). Since $\{-\infty\}$ and $\{\infty\}$ are closed sets in $\overline{\mathbb{R}}$, every $E \subseteq \{-\infty, \infty\}$ is a Borel set, implying $\mathcal{A} \subseteq \overline{\mathcal{B}}$. To verify the remaining inclusion, note $\sigma_{\overline{\mathbb{R}}}(\overline{\mathcal{I}}) = \overline{\mathcal{B}}$, $\overline{\mathcal{I}} \subseteq \mathcal{A}$, and $\mathcal{A}$ is a σ-algebra.

Since the topology on $\mathbb{R}$ is the relative topology inherited from $\overline{\mathbb{R}}$, (A.26b) follows from Cor. A.14. Alternatively, (A.26b) is immediate from (A.26a). ■

A.3.4 $\overline{\mathbb{R}}$-, $\mathbb{R}^n$-, and $\mathbb{C}^n$-Valued Measurable Maps

Definition A.43. (a) An $\overline{\mathbb{R}}$-valued map $f : \Omega \longrightarrow \overline{\mathbb{R}}$ is simply called *measurable* (with respect to some understood measurable space $(\Omega, \mathcal{A})$) if, and only if, f is $\mathcal{A}$-$\overline{\mathcal{B}}$ measurable (in particular, if f is $\mathbb{R}$-valued, then it is *measurable* if, and only if, it is $\mathcal{A}$-$\mathcal{B}^1$-measurable.

(b) Let $n \in \mathbb{N}$. An $\mathbb{R}^n$-valued map $f : \Omega \longrightarrow \mathbb{R}^n$ is simply called *measurable* (with respect to some understood measurable space $(\Omega, \mathcal{A})$) if, and only if, f is $\mathcal{A}$-$\mathcal{B}^n$ measurable.

(c) Let $n \in \mathbb{N}$. A $\mathbb{C}^n$-valued map $f : \Omega \longrightarrow \mathbb{C}^n \cong \mathbb{R}^{2n}$ is simply called *measurable* (with respect to some understood measurable space $(\Omega, \mathcal{A})$) if, and only if, f is $\mathcal{A}$-$\mathcal{B}^{2n}$ measurable.

Theorem A.44. *If* $(\Omega, \mathcal{A})$ *is a measurable space and* $f : \Omega \longrightarrow \overline{\mathbb{R}}$*, then the following statements* (i) – (v) *are equivalent:*

(i) *f is measurable.*

(ii) $f^{-1}]\alpha,\infty] \in \mathcal{A}$ *for each* $\alpha \in \mathbb{R}$.

(iii) $f^{-1}[\alpha,\infty] \in \mathcal{A}$ *for each* $\alpha \in \mathbb{R}$.

(iv) $f^{-1}[-\infty,\alpha[\in \mathcal{A}$ *for each* $\alpha \in \mathbb{R}$.

(v) $f^{-1}[-\infty,\alpha] \in \mathcal{A}$ *for each* $\alpha \in \mathbb{R}$.

The equivalences remain true if, in each statement, $\alpha \in \mathbb{R}$ *is replaced by* $\alpha \in \mathbb{Q}$.

Proof. Since

$$\begin{aligned}\overline{\mathcal{B}} &= \sigma\big\{]\alpha,\infty] : \alpha \in \mathbb{R}\big\} = \sigma\big\{]\alpha,\infty] : \alpha \in \mathbb{Q}\big\}\\ &= \sigma\big\{[\alpha,\infty] : \alpha \in \mathbb{R}\big\} = \sigma\big\{[\alpha,\infty] : \alpha \in \mathbb{Q}\big\}\\ &= \sigma\big\{[-\infty,\alpha[: \alpha \in \mathbb{R}\big\} = \sigma\big\{[-\infty,\alpha[: \alpha \in \mathbb{Q}\big\}\\ &= \sigma\big\{[-\infty,\alpha] : \alpha \in \mathbb{R}\big\} = \sigma\big\{[-\infty,\alpha] : \alpha \in \mathbb{Q}\big\},\end{aligned} \tag{A.27}$$

everything follows from Prop. A.28. ■

Theorem A.45. *Let* $n \in \mathbb{N}$ *and let* $(\Omega,\mathcal{A})$ *be a measurable space. A function* $f = (f_1,\dots,f_n) : \Omega \longrightarrow \mathbb{R}^n$ *is measurable if, and only if, each of the component functions* $f_1,\dots,f_n : \Omega \longrightarrow \mathbb{R}$ *is measurable. In particular,* $g : \Omega \longrightarrow \mathbb{C}^n$ *is measurable if, and only if, both* $\operatorname{Re} g$ *and* $\operatorname{Im} g$ *are measurable.*

Proof. See, e.g., [Els07, Th. III.4.5]. ■

Theorem A.46. *Let* $(\Omega,\mathcal{A})$ *be a measurable space and let* $(f_i)_{i\in\mathbb{N}}$ *be a sequence of measurable functions* $f_i : \Omega \longrightarrow \overline{\mathbb{R}}$. *Then* $\sup_{i\in\mathbb{N}} f_i$, $\inf_{i\in\mathbb{N}} f_i$, $\limsup_{i\to\infty} f_i$, $\liminf_{i\to\infty} f_i$, *and (if it exists in* $\overline{\mathbb{R}}$*)* $\lim_{i\to\infty} f_i$ *all are measurable. In particular, for each* $n \in \mathbb{N}$, $\max(f_1,\dots,f_n)$ *and* $\min(f_1,\dots,f_n)$ *are measurable.*

Proof. See, e.g., [Els07, Th. III.4.3]. ■

Theorem A.47. *Let* $(\Omega,\mathcal{A})$ *be a measurable space,* $f,g : \Omega \longrightarrow \overline{\mathbb{R}}$ *measurable functions, and* $\alpha,\beta \in \overline{\mathbb{R}}$. *Then* $\alpha f + \beta g$, fg, f/g *(with* $(f/g)(\omega)$ *set to some arbitrary fixed* $\gamma \in \overline{\mathbb{R}}$ *for* $g(\omega) = 0$*),* f^+, f^-, $|f|$ *all are measurable.*

Proof. For $\alpha f + \beta g$ and fg see [Els07, Th. III.4.7]. The proof given there also works for f/g. For the remaining cases note $f^+ = \max(f,0)$, $f^- = -\min(f,0)$, and $|f| = f^+ - f^-$. ■

Corollary A.48. *Let* $n \in \mathbb{N}$ *and let* $(\Omega,\mathcal{A})$ *be a measurable space.*

(a) *If* $f,g : \Omega \longrightarrow \mathbb{R}^n$ *are measurable functions, and* $\alpha,\beta \in \mathbb{R}$, *then the componentwise-defined functions* $\alpha f+\beta g$, fg, f/g *(with* $(f_i/g_i)(\omega)$ *set to some arbitrary fixed* $\gamma_i \in \mathbb{R}$ *for* $g_i(\omega) = 0$*),* f^+, f^-, $|f|$ *all are measurable.*

(b) (a) *remains true for* $f, g : \Omega \longrightarrow \mathbb{C}^n$ *measurable. In addition,* $\bar{f}$ *is also measurable.*

Proof. (a) follows when combining Th. A.47 with Th. A.45.
Since, for each $k = 1, \ldots, n$,

$$\begin{aligned} f_k g_k &= \operatorname{Re} f_k \operatorname{Re} g_k - \operatorname{Im} f_k \operatorname{Im} g_k + i(\operatorname{Im} f_k \operatorname{Re} g_k + \operatorname{Re} f_k \operatorname{Im} g_k), \\ f_k / g_k &= \frac{\operatorname{Re} f_k \operatorname{Re} g_k + \operatorname{Im} f_k \operatorname{Im} g_k}{(\operatorname{Re} g_k)^2 + (\operatorname{Im} g_k)^2} + i \, \frac{\operatorname{Im} f_k \operatorname{Re} g_k - \operatorname{Re} f_k \operatorname{Im} g_k}{(\operatorname{Re} g_k)^2 + (\operatorname{Im} g_k)^2}, \\ |f_k| &= \sqrt{(\operatorname{Re} f_k)^2 + (\operatorname{Im} f_k)^2}, \\ \bar{f}_k &= \operatorname{Re} f_k - 2i \operatorname{Im} f_k, \end{aligned}$$

(b) follows from (a) and Th. A.47. ■

Definition A.49. Let Ω be a set and $A \subseteq \Omega$. Then

$$\chi_A : \Omega \longrightarrow \mathbb{R}, \quad \chi_A(\omega) := \begin{cases} 1 & \text{for } \omega \in A, \\ 0 & \text{for } \omega \notin A, \end{cases} \tag{A.28}$$

is called the *characteristic* or *indicator function* of A.

Remark A.50. If $(\Omega, \mathcal{A})$ is a measurable space and $A \in \mathcal{A}$, then the characteristic function χ_A is measurable if, and only if, $A \in \mathcal{A}$ (since $\chi_A^{-1}(\{1\}) = A$).

Definition A.51. Let $(\Omega, \mathcal{A})$ be a measurable space. A function $f : \Omega \longrightarrow \mathbb{R}$ is called a *simple function* or *step function* if, and only if, it is a linear combination of measurable characteristic functions.

Remark A.52. A simple function is always measurable, since it is, by definition, a linear combination of measurable functions.

Theorem A.53. *Let* $(\Omega, \mathcal{A})$ *be a measurable space.*

(a) *A function* $f : \Omega \longrightarrow [0, \infty]$ *is measurable if, and only if, there exists an increasing sequence* $(\phi_i)_{i \in \mathbb{N}}$ *of simple functions such that* $f = \lim_{i \to \infty} f_i$.

(b) *Every bounded measurable* $\mathbb{R}$*-valued function* $f : \Omega \longrightarrow [0, \infty]$ *is the uniform limit of an increasing sequence* $(\phi_i)_{i \in \mathbb{N}}$ *of simple functions.*

(c) *Every measurable* $f : \Omega \longrightarrow]-\infty, \infty]$*, which is bounded from below is the pointwise limit of an increasing sequence* $(\phi_i)_{i \in \mathbb{N}}$ *of simple functions.*

(d) *Every measurable* $f : \Omega \longrightarrow \overline{\mathbb{R}}$ *is the pointwise limit of a (not necessarily increasing) sequence* $(\phi_i)_{i \in \mathbb{N}}$ *of simple functions.*

Proof. (a): See, e.g., [Els07, Th. III.4.13].

(b): The proof of [Els07, Th. III.4.13] also shows the uniform convergence for nonnegative, bounded, measurable f. For $\mathbb{R}$-valued bounded, measurable f, let $m \in \mathbb{R}$ be a lower bound (the nontrivial case is $m < 0$, i.e. $-m > 0$). Then we obtain an increasing sequence $(\phi_i)_{i\in\mathbb{N}}$ of simple functions uniformly converging to $f-m \geq 0$. Then $(\phi_i+m)_{i\in\mathbb{N}}$ is an increasing sequence of simple functions uniformly converging to f.

(c): As in (b), one applies (a) to $f - m \geq 0$, where $m \in \mathbb{R}$ is a lower bound for f.

(d): One writes $f = f^+ - f^-$ and applies (a) to f^+ and f^-. ■

A.4 Integration

A.4.1 Lebesgue Integral of $\overline{\mathbb{R}}$-Valued Measurable Maps

We define the Lebesgue integral in the usual way, first for nonnegative simple functions in Def. A.54(a), then for nonnegative measurable functions in Def. A.54(b), and then for so-called integrable functions in Def. A.54(c). The definitions of the Lebesgue integral for nonnegative simple functions and for nonnegative measurable functions make use of representations, where Th. A.55 then states that the value of the integral does actually *not* depend on the representation of the integrated function.

Definition A.54. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space.

(a) Let $f : \Omega \longrightarrow \mathbb{R}_0^+$ be a simple function, where

$$f = \sum_{i=1}^{N} \alpha_i \, \chi_{A_i} \tag{A.29a}$$

with $N \in \mathbb{N}$; $\alpha_1, \ldots, \alpha_N \geq 0$; and $A_1, \ldots, A_N \in \mathcal{A}$. Then

$$\int_\Omega f \, \mathrm{d}\mu \;:=\; \int_\Omega f(x) \, \mathrm{d}\mu(x) \;:=\; \sum_{i=1}^{N} \alpha_i \, \mu(A_i) \quad \in \quad [0, \infty] \tag{A.29b}$$

is called the *Lebesgue integral* of f over Ω with respect to μ.

(b) Let $f : \Omega \longrightarrow [0, \infty]$ be measurable, where $(\phi_i)_{i\in\mathbb{N}}$ is an increasing sequence of simple functions such that $f = \lim_{i\to\infty} \phi_i$. Then

$$\int_\Omega f \, \mathrm{d}\mu \;:=\; \int_\Omega f(x) \, \mathrm{d}\mu(x) \;:=\; \lim_{i\to\infty} \int_\Omega \phi_i \, \mathrm{d}\mu \quad \in \quad [0, \infty] \tag{A.30}$$

is called the *Lebesgue integral* of f over Ω with respect to μ (the limit in (A.30) exists, since (A.29b) implies $\int_\Omega \phi \, \mathrm{d}\mu \;\geq\; \int_\Omega \psi \, \mathrm{d}\mu$ for simple functions $\phi \geq \psi$, see [Els07, IV.1.3(c)]).

(c) A function $f : \Omega \longrightarrow \overline{\mathbb{R}}$ such that $\int_\Omega f^+ \,\mathrm{d}\mu < \infty$ and $\int_\Omega f^- \,\mathrm{d}\mu < \infty$ is called *integrable*. For integrable functions f,

$$\int_\Omega f \,\mathrm{d}\mu := \int_\Omega f(x) \,\mathrm{d}\mu(x) := \int_\Omega f^+ \,\mathrm{d}\mu - \int_\Omega f^- \,\mathrm{d}\mu \quad \in \quad \mathbb{R} \tag{A.31}$$

is called the *Lebesgue integral* of f over Ω with respect to μ.

If $f : \Omega \longrightarrow]-\infty, \infty]$ is nonnegative measurable or integrable and $A \in \mathcal{A}$, then one also defines

$$\int_A f \,\mathrm{d}\mu := \int_A f(x) \,\mathrm{d}\mu(x) := \int_\Omega f \,\chi_A \,\mathrm{d}\mu \,. \tag{A.32}$$

Theorem A.55. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space.*

(a) *The value of the integral in* (A.29b) *does not depend on the representation of the simple function f: If $M, N \in \mathbb{N}$; $\alpha_1, \ldots, \alpha_N \geq 0$; $\beta_1, \ldots, \beta_M \geq 0$; $A_1, \ldots, A_N \in \mathcal{A}$; and $B_1, \ldots, B_M \in \mathcal{A}$ are such that*

$$f = \sum_{i=1}^{N} \alpha_i \,\chi_{A_i} = \sum_{i=1}^{M} \beta_i \,\chi_{B_i}, \tag{A.33a}$$

then

$$\sum_{i=1}^{N} \alpha_i \,\mu(A_i) = \sum_{i=1}^{M} \beta_i \,\mu(B_i). \tag{A.33b}$$

(b) *The value of the integral in* (A.30) *does not depend on the representation of the nonnegative measurable function f: If $(\phi_i)_{i\in\mathbb{N}}$ and $(\psi_i)_{i\in\mathbb{N}}$ both are increasing sequences of simple functions, then*

$$f = \lim_{i\to\infty} \phi_i = \lim_{i\to\infty} \psi_i \quad \Rightarrow \quad \lim_{i\to\infty} \int_\Omega \phi_i \,\mathrm{d}\mu = \lim_{i\to\infty} \int_\Omega \psi_i \,\mathrm{d}\mu \,. \tag{A.34}$$

In particular, (A.29b) *and* (A.30) *are consistently defined.*

Proof. (a): See, e.g., [Els07, Lem. IV.1.1].

(b): See, e.g., [Els07, Th. IV.2.1 and Cor. IV.2.2]. ■

Theorem A.56. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $f, g : \Omega \longrightarrow \overline{\mathbb{R}}$ be both nonnegative measurable or both integrable, let $A, B \in \mathcal{A}$, and let $\alpha, \beta \in \mathbb{R}$ ($\alpha, \beta = \infty$ is allowed for f, g nonnegative measurable).*

(a) *The Lebesgue integral is linear:*

$$\int_A (\alpha f + \beta g) \,\mathrm{d}\mu = \alpha \int_A f \,\mathrm{d}\mu + \beta \int_A g \,\mathrm{d}\mu \,. \tag{A.35a}$$

(b) *If A, B are disjoint, then*

$$\int_{A\cup B} f\,\mathrm{d}\mu \,=\, \int_A f\,\mathrm{d}\mu \,+\, \int_B f\,\mathrm{d}\mu\,. \tag{A.35b}$$

(c) *The Lebesgue integral is isotone:*

$$f \le g \quad\Rightarrow\quad \int_A f\,\mathrm{d}\mu \,\le\, \int_A g\,\mathrm{d}\mu\,. \tag{A.35c}$$

(d) *The Lebesgue integral satisfies the triangle inequality:*

$$\left|\int_A f\,\mathrm{d}\mu\right| \le \int_A |f|\,\mathrm{d}\mu\,. \tag{A.35d}$$

(e) *Mean Value Theorem for Integration: If there exist numbers $m, M \in \mathbb{R}$ such that $m \le f \le M$ on A, then*

$$m\,\mu(A) \le \int_A f\,\mathrm{d}\mu \,\le M\,\mu(A). \tag{A.35e}$$

The theorem's name comes from the fact that, for $0 < \mu(A) < \infty$, $\mu(A)^{-1}\int_A f\,\mathrm{d}\mu$ is sometimes referred to as the mean value *of f on A.*

(f) *If f is nonnegative measurable, then*

$$\int_\Omega f\,\mathrm{d}\mu \,= \sup\left\{\int_\Omega \phi\,\mathrm{d}\mu \,:\, \phi:\,\Omega\longrightarrow\mathbb{R}_0^+ \text{ simple, } \phi\le f\right\}. \tag{A.35f}$$

(g) *If f is nonnegative measurable, then*

$$\int_\Omega f\,\mathrm{d}\mu \,= 0 \quad\Leftrightarrow\quad \mu\big(\{f>0\}\big) = 0. \tag{A.35g}$$

Proof. (a): For the nonnegative measurable case, see e.g., [Els07, Lem. IV.2.4(a)]. Then the integrable case follows by writing $f = f^+ - f^-$ and $g = g^+ - g^-$.

(b): Since $f\chi_{A\cup B} = f\chi_A + f\chi_B$, (A.35b) is immediate from (a).

(c): For the nonnegative measurable case, see e.g., [Els07, Lem. IV.2.4(b)]. Then the integrable case follows from (a) by applying the nonnegative measurable case to $g - f \ge 0$.

(d) follows from (c), since $f \le |f|$ and $-f \le |f|$.

(e) is also an easy consequence of (c).

(f): The $\le$-part of (A.35f) follows from (A.35c), whereas the $\ge$-part follows from (A.30).

(g): See, e.g., [Els07, Th. IV.2.6]. ■

Theorem A.57 (Monotone Convergence)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. For every increasing sequence $(f_i)_{i\in\mathbb{N}}$ of nonnegative measurable functions $f_i : \Omega \longrightarrow [0,\infty]$, the following holds true:*

$$\int_\Omega \left(\lim_{i\to\infty} f_i\right) \mathrm{d}\mu = \lim_{i\to\infty} \int_\Omega f_i \,\mathrm{d}\mu\,. \tag{A.36}$$

Proof. See, e.g., [Els07, Th. IV.2.7]. ■

A.4.2 Lebesgue Integral of $\mathbb{R}^n$- and $\mathbb{C}^n$-Valued Measurable Maps

Definition A.58. Let $n \in \mathbb{N}$. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $f = (f_1, \ldots, f_n) : \Omega \longrightarrow \mathbb{C}^n$. We call f *integrable* if, and only if, for each $k = 1, \ldots, n$, $\operatorname{Re} f_k$ and $\operatorname{Im} f_k$ are both integrable in the sense of Def. A.54(c). Then, define the *Lebesgue integral* of f over Ω with respect to μ componentwise:

$$\int_\Omega f \,\mathrm{d}\mu := \left(\int_\Omega \operatorname{Re} f_1 + i \int_\Omega \operatorname{Im} f_1 \,\mathrm{d}\mu\,, \ldots, \int_\Omega \operatorname{Re} f_n + i \int_\Omega \operatorname{Im} f_n \,\mathrm{d}\mu\right). \tag{A.37}$$

As before, one also defines

$$\forall_{A\in\mathcal{A}} \quad \int_A f \,\mathrm{d}\mu := \int_\Omega f\chi_A \,\mathrm{d}\mu\,. \tag{A.38}$$

Remark A.59. Clearly, (A.35a) and (A.35b) remain valid for integrable $f, g : \Omega \longrightarrow \mathbb{C}^n$ and $\alpha, \beta \in \mathbb{C}$.

A.4.3 L^p-Spaces

Notation A.60. (a) We use the symbol $\mathbb{K}$ to denote either $\mathbb{R}$ or $\mathbb{C}$. Thus, if $\mathbb{K}$ occurs in a statement or definition, then the statement or definition is meant to be valid for $\mathbb{K}$ replaced by $\mathbb{R}$ and for $\mathbb{K}$ replaced by $\mathbb{C}$.

(b) Analogous to (a), we use the symbol $\hat{\mathbb{K}}$ to denote either $\overline{\mathbb{R}}$ or $\mathbb{C}$.

Notation A.61. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Using the convention

$$\forall_{p\in\mathbb{R}^+} \quad \infty^p := \infty, \tag{A.39}$$

we define for every measurable $f : \Omega \longrightarrow \hat{\mathbb{K}}$:

$$\forall_{p\in[1,\infty[} \quad N_p(f) := \left(\int_\Omega |f|^p \,\mathrm{d}\mu\right)^{1/p} \quad \in \quad [0,\infty] \tag{A.40a}$$

and

$$N_\infty(f) := \inf\left\{\sup\left\{|f(\omega)| : \omega \in \Omega \setminus N\right\} : N \in \mathcal{A},\, \mu(N) = 0\right\} \quad \in \quad [0,\infty], \tag{A.40b}$$

where the number $N_\infty(f)$ is also known as the *essential supremum* of f.

Theorem A.62 (Hölder Inequality)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$. If $f, g : \Omega \longrightarrow \hat{\mathbb{K}}$ are measurable, then*

$$N_1(fg) \leq N_p(f)\, N_q(g), \tag{A.41}$$

where (A.41) *is known as* Hölder inequality*.*

Proof. See, e.g., [Els07, Th. VI.1.5]. ■

Definition A.63. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $p \in [1, \infty]$.

(a) Let $\mathcal{L}^p := \mathcal{L}^p_{\mathbb{K}} := \mathcal{L}^p_{\mathbb{K}}(\mu) := \mathcal{L}^p_{\mathbb{K}}(\Omega, \mathcal{A}, \mu)$ denote the set of all measurable functions $f : \Omega \longrightarrow \mathbb{K}$ such that $N_p(f) < \infty$. It is common to introduce the notation

$$\underset{f \in \mathcal{L}^p}{\forall} \quad \|f\|_p := N_p(f). \tag{A.42}$$

(b) Let $\mathcal{N}$ denote the set of measurable $f : \Omega \longrightarrow \mathbb{K}$ that vanish μ-almost everywhere. Clearly, both $\mathcal{N}$ and each $\mathcal{L}^p$ are vector spaces over $\mathbb{K}$, where $\mathcal{N}$ is a subspace of each $\mathcal{L}^p$. Thus, it makes sense to define the quotient spaces

$$L^p := L^p_{\mathbb{K}} := L^p_{\mathbb{K}}(\mu) := L^p_{\mathbb{K}}(\Omega, \mathcal{A}, \mu) := \mathcal{L}^p_{\mathbb{K}}(\Omega, \mathcal{A}, \mu)/\mathcal{N}. \tag{A.43}$$

If $f, g \in \mathcal{L}^p$ represent the same element of L^p, i.e. $[f] = [g] \in L^p$, then $\|f\|_p = \|g\|_p$. Thus, it makes sense to define

$$\underset{f \in \mathcal{L}^p}{\forall} \quad \|[f]\|_p := \|f\|_p. \tag{A.44}$$

Remark A.64. In practise, it is very common not to properly distinguish between elements $[f] \in L^p$ and their representatives $f \in \mathcal{L}^p$, and, in most situations, one gets aways with it without getting into trouble. There are circumstances, however, such as traces on boundaries and integration in product spaces, where one has to use caution while being lax with $[f]$ and f.

Theorem A.65. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $p \in [1, \infty]$.*

(a) *$\|\cdot\|_p$ constitutes a seminorm on $\mathcal{L}^p_{\mathbb{K}}(\mu)$, i.e. it makes $\mathcal{L}^p_{\mathbb{K}}(\mu)$ into a seminormed vector space over $\mathbb{K}$. However, if there exists a nonempty μ-null set in $\mathcal{A}$, then $\|\cdot\|_p$ is* not *a norm on $\mathcal{L}^p_{\mathbb{K}}(\mu)$ and the resulting topology on $\mathcal{L}^p_{\mathbb{K}}(\mu)$ is* not *Hausdorff.*

(b) *$\|\cdot\|_p$ constitutes a norm on $L^p_{\mathbb{K}}(\mu)$, i.e. it makes $L^p_{\mathbb{K}}(\mu)$ into a normed vector space over $\mathbb{K}$.*

Proof. See, e.g., [Els07, Th. VI.2.2 and Th. VI.2.3]. ■

Theorem A.66 (Riesz-Fischer)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $p \in [1, \infty]$. All the spaces $\mathcal{L}^p_{\mathbb{K}}(\mu)$ and $L^p_{\mathbb{K}}(\mu)$ are complete. In particular, all $L^p_{\mathbb{K}}(\mu)$ are Banach spaces.*

Proof. See, e.g., [Els07, Th. VI.2.5 and Cor. VI.2.6]. ■

A.4.4 Measures with Density

Proposition A.67. *If $(\Omega, \mathcal{A}, \mu)$ is a measure space and $f : \Omega \longrightarrow [0, \infty]$ is measurable, then*

$$f\mu : \mathcal{A} \longrightarrow [0, \infty], \quad (f\mu)(A) := \int_A f \, \mathrm{d}\mu \,, \tag{A.45}$$

defines a measure on $(\Omega, \mathcal{A})$.

Proof. We have $(f\mu)(\emptyset) = \int_\emptyset f \, \mathrm{d}\mu = 0$ and, if $(A_i)_{i\in\mathbb{N}}$ is a sequence in $\mathcal{A}$ consisting of pairwise disjoint sets, then

$$\begin{aligned}(f\mu)\left(\bigcup_{i=1}^{\infty} A_i\right) &= \int_\Omega \left(f \sum_{i=1}^{\infty} \chi_{A_i}\right) \mathrm{d}\mu \overset{(A.36)}{=} \sum_{i=1}^{\infty} \int_\Omega f\, \chi_{A_i} \, \mathrm{d}\mu \\ &= \sum_{i=1}^{\infty} \int_{A_i} f \, \mathrm{d}\mu = \sum_{i=1}^{\infty} (f\mu)(A_i),\end{aligned} \tag{A.46}$$

thereby establishing the case. ■

Definition A.68. If μ, ν are measures on the measurable space $(\Omega, \mathcal{A})$, then a measurable function $f : \Omega \longrightarrow [0, \infty]$ is called a *density* of ν with respect to μ if, and only if, $\nu = f\mu$ with $f\mu$ as in (A.45).

—

Clearly, in general, given a measure space $(\Omega, \mathcal{A}, \mu)$, not every measure ν on $(\Omega, \mathcal{A})$ has a density with respect to μ (see Ex. A.71 below). The existence of a density is related to the following notion to absolute continuity of measures:

Definition A.69. If μ, ν are measures on the measurable space $(\Omega, \mathcal{A})$, then ν is called *absolutely continuous* with respect to μ (or sometimes just μ-continuous), denoted $\nu \ll \mu$, if, and only if, every μ-null set is a ν-null set, i.e.

$$\nu \ll \mu \quad :\Leftrightarrow \quad \underset{A\in\mathcal{A}}{\forall} \quad \Big(\mu(A) = 0 \;\Rightarrow\; \nu(A) = 0 \Big). \tag{A.47}$$

Lemma A.70. *If μ, ν are measures on the measurable space $(\Omega, \mathcal{A})$ and $\nu = f\mu$ with a measurable function $f : \Omega \longrightarrow [0, \infty]$ (i.e. ν has a density with respect to μ), then $\nu \ll \mu$.*

Proof. If $A \in \mathcal{A}$ and $\mu(A) = 0$, then $\nu(A) = \int_A f \, \mathrm{d}\mu = 0$ as claimed. ■

Example A.71. If $(\Omega, \mathcal{A})$ is a measurable space and $\omega \in \Omega$, then

$$\delta_\omega : \mathcal{A} \longrightarrow [0, \infty], \quad \delta_\omega(A) := \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A, \end{cases} \tag{A.48}$$

defines a measure on $(\Omega, \mathcal{A})$ (as is readily verified), the so-called *Dirac measure* concentrated in ω. Nonnegative countable linear combinations of Dirac measures are called

discrete: A measure μ on $(\Omega, \mathcal{A})$ is called *discrete* if, and only if, there exist sequences $(\omega_i)_{i\in\mathbb{N}}$ in Ω and $(a_i)_{i\in\mathbb{N}}$ in $[0,\infty]$ such that

$$\mu : \mathcal{A} \longrightarrow [0,\infty], \quad \mu(A) = \sum_{i\in\mathbb{N}} a_i\, \delta_{\omega_i}(A). \tag{A.49}$$

If $(\Omega, \mathcal{A}) = (\mathbb{R}^n, \mathbb{B}^n)$, $n \in \mathbb{N}$, and $\mu \neq 0$ is a discrete measure as above, then

$$\begin{aligned} \mu\big(\{\omega_i : i \in \mathbb{N}\}\big) &= \sum_{i\in\mathbb{N}} \mu\big(\{\omega_i\}\big) = \sum_{i\in\mathbb{N}} a_i \delta_{\omega_i}\big(\{\omega_i\}\big) \\ &= \sum_{i\in\mathbb{N}} a_i \delta_{\omega_i}(\mathbb{R}^n) = \mu(\mathbb{R}^n) \neq 0 = \beta_n\big(\{\omega_i : i \in \mathbb{N}\}\big), \end{aligned} \tag{A.50}$$

i.e., by Lem. A.70, no (nontrivial) discrete measure can be absolutely continuous with respect to β_n (or λ_n). However, there are also nondiscrete measures on $(\mathbb{R}^n, \mathbb{B}^n)$ that do not have a density with respect to β_n (or λ_n) – this is related to the fact that there exist uncountable sets $N \in \mathcal{B}^n$ with $\beta_n(N) = 0$ (even for $n = 1$, see [RF10, Prop. 2.19]).

—

For σ-finite measures, the converse of Lem. A.70 holds as well:

Theorem A.72 (Radon-Nikodym)**.** *Let μ, ν be given measures on the measurable space $(\Omega, \mathcal{A})$. If μ is σ-finite (see Def.* A.17(b)*) and $\nu \ll \mu$, then ν has a density f with respect to μ, i.e. there exists a measurable $f : \Omega \longrightarrow [0,\infty]$ such that $\nu = f\mu$, i.e. $\nu(A) = \int_A f \,\mathrm{d}\mu$ for each $A \in \mathcal{A}$.*

Proof. See, e.g., [Bau92, Th. 17.10]. ■

Theorem A.73. *Densities with respect to a σ-finite measure μ are unique μ-almost everywhere: If μ, ν are measures on the measurable space $(\Omega, \mathcal{A})$, μ is σ-finite, and there are measurable functions $f, g : \Omega \longrightarrow [0,\infty]$, then $f = g$ μ-almost everywhere. Moreover, ν is σ-finite if, and only if, f is $\mathbb{R}$-valued μ-almost everywhere (i.e. if, and only if, $\mu(\{f = \infty\}) = 0$).*

Proof. See, e.g., [Bau92, Th. 17.11]. ■

Definition A.74. If μ, ν are measures on the measurable space $(\Omega, \mathcal{A})$, μ is σ-finite, and $\nu \ll \mu$, then the unique density $f : \Omega \longrightarrow [0,\infty]$ that ν has with respect to μ according to Ths. A.72 and A.73 is called the *Radon-Nikodym derivative* of ν with respect to μ; in consequence, it is sometimes denoted by

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} := f. \tag{A.51}$$

A.5 Product Spaces

A.5.1 Product σ-Algebras

Definition A.75. (a) Given a family of sets $(\Omega_i)_{i\in I}$, the *Cartesian product* of the Ω_i is the set of functions

$$\Omega := \prod_{i\in I} \Omega_i := \left\{ \left(f : I \to \bigcup_{j\in I} \Omega_j \right) : \underset{i\in I}{\forall} f(i) \in \Omega_i \right\}, \tag{A.52}$$

and the maps defined by

$$\underset{i\in I}{\forall} \quad \pi_i : \Omega \longrightarrow \Omega_i, \quad \pi_i(f) := f(i), \tag{A.53}$$

are called the corresponding *projections*.

(b) Given a family of measurable spaces $(\Omega_i, \mathcal{A}_i)_{i\in I}$, and $\Omega := \prod_{i\in I} \Omega_i$, define the *product* σ-algebra

$$\mathcal{A} := \bigotimes_{i\in I} \mathcal{A}_i := \sigma((\pi_i)_{i\in I}), \tag{A.54}$$

i.e. $\mathcal{A}$ is the smallest σ-algebra on Ω with respect to which all projections π_i, $i \in I$, are measurable.

Proposition A.76. *If $(\Omega_i, \mathcal{A}_i)_{i\in I}$ is a family of measurable spaces and $(\mathcal{E}_i)_{i\in I}$ is a family of generators (i.e. $\mathcal{E}_i \subseteq \mathcal{A}_i$ and $\sigma(\mathcal{E}_i) = \mathcal{A}_i$ for each $i \in I$), then*

$$\mathcal{E} := \left\{ E \times \prod_{i\in I\setminus\{j\}} \Omega_i : \, j \in I, \; E \in \mathcal{E}_j \right\} \tag{A.55}$$

is a generator of the product σ-algebra $\mathcal{A} := \bigotimes_{i\in I} \mathcal{A}_i$.

Proof. We have

$$\underset{j\in I}{\forall} \quad \underset{E\in\mathcal{E}_j}{\forall} \quad \pi_j^{-1}(E) = E \times \prod_{i\in I\setminus\{j\}} \Omega_i, \tag{A.56}$$

showing $\sigma(\mathcal{E}) \subseteq \mathcal{A}$. For the opposite inclusion, we note that (A.56) implies

$$\mathcal{E} = \bigcup_{j\in J} \pi_j^{-1}(\mathcal{E}_j), \tag{A.57}$$

in particular,

$$\underset{j\in I}{\forall} \quad \pi_j^{-1}(\mathcal{E}_j) \subseteq \mathcal{E} \subseteq \sigma(\mathcal{E}). \tag{A.58}$$

Thus,

$$\underset{j\in I}{\forall} \quad \sigma(\pi_j) \overset{\text{Prop. A.4}}{=} \pi_j^{-1}(\mathcal{A}_j) = \pi_j^{-1}(\sigma(\mathcal{E}_j)) \overset{\text{Th. A.9}}{=} \sigma\big(\pi_j^{-1}(\mathcal{E}_j)\big) \overset{(A.58)}{\subseteq} \sigma(\mathcal{E}), \tag{A.59}$$

implying $\mathcal{A} \subseteq \sigma(\mathcal{E})$. ∎

Proposition A.77. *If $(\Omega_i, \mathcal{A}_i)_{i\in I}$ is a* finite *family of measurable spaces (i.e. $\#I = n \in \mathbb{N}$) and $(\mathcal{E}_i)_{i\in I}$ is a family of generators (i.e. $\mathcal{E}_i \subseteq \mathcal{A}_i$ and $\sigma(\mathcal{E}_i) = \mathcal{A}_i$ for each $i \in I$) such that*

$$\forall_{i\in I} \quad \exists_{(E_{i,k})_{k\in\mathbb{N}} \text{ in } \mathcal{E}_i} \quad \Omega_i = \bigcup_{k\in\mathbb{N}} E_{i,k}, \tag{A.60}$$

then

$$\mathcal{E} := \left\{ \prod_{i\in I} E_i : \forall_{i\in I} \quad E_i \in \mathcal{E}_i \right\} \tag{A.61}$$

is a generator of the product σ-algebra $\mathcal{A} := \bigotimes_{i\in I} \mathcal{A}_i$.

Proof. See, e.g., [Els07, p. 113]. ■

Proposition A.78. *Forming product σ-algebras is associative: If $(\Omega_i, \mathcal{A}_i)_{i\in I}$ is a family of measurable spaces, K is a nonempty index set, and $(I_\kappa)_{\kappa\in K}$ is a family of nonempty subsets of I such that $I = \bigcup_{\kappa\in K} I_\kappa$, then, in the sense of the canonical identification between $\prod_{\kappa\in K} \left(\prod_{i\in I_\kappa} \Omega_i \right)$ and $\prod_{i\in I} \Omega_i$, the following holds:*

$$\bigotimes_{\kappa\in K} \left(\bigotimes_{i\in I_\kappa} \mathcal{A}_i \right) = \bigotimes_{i\in I} \mathcal{A}_i. \tag{A.62}$$

Proof. See, e.g., [Els07, Ex. III.5.5(a)]. ■

A.5.2 Product Borel σ-Algebras

A natural and important question is if forming products of Borel σ-algebras commutes with forming the product topology and then taking the resulting Borel σ-algebra. The short answer is that, unfortunately, in general the above constructions do *not(!)* commute in general, but that everything works fine (they do commute) for $\mathbb{R}^n$.

We start by recalling the definition of the product topology, which is completely analogous to the definition of the product σ-algebra in Def. A.75(b) above:

Definition A.79. Given a family of topological spaces $(\Omega_i, \tau_i)_{i\in I}$, and $\Omega := \prod_{i\in I} \Omega_i$, define the *product* topology

$$\tau := \bigotimes_{i\in I} \tau_i \tag{A.63}$$

to be the smallest topology on Ω with respect to which all projections π_i, $i \in I$, are continuous. Then, clearly, τ is generated by the subbase of open sets

$$\mathcal{S} := \left\{ O \times \prod_{i\in I\setminus\{j\}} \Omega_i : j \in I,\ O \in \tau_j \right\}. \tag{A.64}$$

If Ω and τ are as above, then one says that (Ω, τ) is the *topological product* of the $(\Omega_i, \tau_i)_{i\in I}$.

Theorem A.80. *Given a family of topological spaces $(\Omega_i, \tau_i)_{i\in I}$ with topological product (Ω, τ), one always has*

$$\sigma(\tau) \supseteq \bigotimes_{i\in I} \sigma(\tau_i), \tag{A.65}$$

i.e. the Borel σ-algebra of the product topology always contains the product of the Borel σ-algebras.

Proof. By definition, the product of the Borel σ-algebras, i.e. the σ-algebra on the right-hand side of (A.65) is the smallest σ-algebra with respect to which all the projections are measurable. However, since all projections are τ-continuous, all projections are $\sigma(\tau)$-measurable, thereby proving (A.65). ■

Caveat A.81. Without additional hypotheses, such as the ones given in Th. A.82 below, equality can not be expected in (A.65): An entire class of examples, where equality fails even for just two factors is given by [Els07, Rem. III.5.16], whereas [Els07, Exercise III.5.3] provides a concrete example.

Theorem A.82. *If $(\Omega_i, \tau_i)_{i\in\mathbb{N}}$ is a* countable *family of topological spaces such that each topology τ_i has a* countable *base, and if (Ω, τ) denotes the topological product, then*

$$\sigma(\tau) = \bigotimes_{i\in I} \sigma(\tau_i), \tag{A.66}$$

i.e. the Borel σ-algebra of the product topology is the same as the product of the Borel σ-algebras.

Proof. See, e.g., [Els07, Th. III.5.10]. ■

Corollary A.83. *Let $m, n \in \mathbb{N}$. For the Borel σ-algebras $\mathcal{B}^m$ and $\mathcal{B}^n$ of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, we have*

$$\mathcal{B}^{m+n} = \mathcal{B}^m \otimes \mathcal{B}^n, \quad \mathcal{B}^n = \otimes_{i=1}^n \mathcal{B}^1. \tag{A.67a}$$

Proof. To apply Th. A.82, one merely has to note that the (usual) topology on $\mathbb{R}$ has a countable base (e.g. given by the collection of all open intervals with rational endpoints). ■

A.5.3 Product Measure Spaces

Definition A.84. Let $n \in \mathbb{N}$ and let $(\Omega_i, \mathcal{A}_i, \mu_i)_{i=1}^n$ be a finite family of measure spaces, $\Omega := \prod_{i=1}^n \Omega_i$, $\mathcal{A} := \otimes_{i=1}^n \mathcal{A}_i$. Then a measure μ on $(\Omega, \mathcal{A})$ is called a *product measure* if, and only if,

$$\forall_{(A_1,\ldots,A_n)\in\prod_{i=1}^n \mathcal{A}_i} \quad \mu\left(\prod_{i=1}^n A_i\right) = \prod_{i=1}^n \mu_i(A_i). \tag{A.68}$$

Theorem A.85. *Let $n \in \mathbb{N}$ and let $(\Omega_i, \mathcal{A}_i, \mu_i)_{i=1}^n$ be a finite family of measure spaces, $\Omega := \prod_{i=1}^n \Omega_i$, $\mathcal{A} := \otimes_{i=1}^n \mathcal{A}_i$.*

(a) *There always exists at least one product measure on* $(\Omega, \mathcal{A})$.

(b) *If each measure* μ_i *is* σ*-finite, then there exists a* unique *product measure* μ *on* $(\Omega, \mathcal{A})$*, denoted by* $\otimes_{i=1}^n \mu_i := \mu$*. Moreover,* $\otimes_{i=1}^n \mu_i$ *is itself* σ*-finite and, defining*

$$\bigvee_{M \subseteq \Omega} \quad \bigvee_{x_n \in \Omega_n} \quad M_{x_n} := \{(x_1, \ldots, x_{n-1}) : (x_1, \ldots, x_n) \in M\}, \tag{A.69a}$$

one has

$$\bigvee_{M \in \mathcal{A}} \quad x_n \mapsto \left(\otimes_{i=1}^{n-1} \mu_i\right)(M_{x_n}) \quad \text{is measurable} \tag{A.69b}$$

and

$$\bigvee_{M \in \mathcal{A}} \quad \left(\otimes_{i=1}^{n} \mu_i\right)(M) = \int_{\Omega_n} \left(\otimes_{i=1}^{n-1} \mu_i\right)(M_{x_n}) \, \mathrm{d}\mu_n(x_n). \tag{A.69c}$$

In terms of the canonical identification $(\prod_{i=1}^{n-1} \Omega_i) \times \Omega_n \cong \prod_{i=1}^{n} \Omega_i$*, one has*

$$(\otimes_{i=1}^{n-1} \mu_i) \otimes \mu_n = \otimes_{i=1}^{n} \mu_i. \tag{A.69d}$$

Proof. For $n = 2$, see, e.g., [Els07, Th. V.1.2/1.3]. The general case then follows by induction (cf. [Els07, Th. V.1.12]). ■

Caveat A.86. In general, the conclusion of Th. A.85(b) does not hold if the μ_i are not all σ-finite: Several different product measures can exist (see [Els07, Ex. V.1.4] for an example with $n = 2$) and the map in (A.69b) can be *nonmeasurable* so that (A.69c) does not even make sense (see [Beh87, p. 96] for an example with $n = 2$).

A.5.4 Theorems of Tonelli and Fubini

Theorem A.87. *Let* $(\Omega_1, \mathcal{A}, \mu)$ *and* $(\Omega_2, \mathcal{B}, \nu)$ *be* σ*-finite measure spaces.*

(a) Tonelli's Theorem*: For each nonnegative* $(\mathcal{A} \otimes \mathcal{B})$*-measurable function* $f : \Omega_1 \times \Omega_2 \longrightarrow [0, \infty]$*, the functions given by*

$$\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2) \, \mathrm{d}\nu(\omega_2) \ \in [0, \infty], \quad \omega_2 \mapsto \int_{\Omega_1} f(\omega_1, \omega_2) \, \mathrm{d}\mu(\omega_1) \ \in [0, \infty], \tag{A.70}$$

are $\mathcal{A}$*-measurable (resp.* $\mathcal{B}$*-measurable) and*

$$\begin{aligned} &\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) \, \mathrm{d}(\mu \otimes \nu)(\omega_1, \omega_2) \\ &= \int_{\Omega_2} \int_{\Omega_1} f(\omega_1, \omega_2) \, \mathrm{d}\mu(\omega_1) \, \mathrm{d}\nu(\omega_2) \\ &= \int_{\Omega_1} \int_{\Omega_2} f(\omega_1, \omega_2) \, \mathrm{d}\nu(\omega_2) \, \mathrm{d}\mu(\omega_1). \end{aligned} \tag{A.71}$$

(b) Fubini's Theorem*: For each $(\mu \otimes \nu)$-integrable function $f : \Omega_1 \times \Omega_2 \longrightarrow \hat{\mathbb{K}}$, the function $f(\omega_1, \cdot)$ is ν-integrable for μ-almost every $\omega_1 \in \Omega_1$ – in particular,*

$$A := \{\omega_1 \in \Omega_1 : f(\omega_1, \cdot) \text{ is not } \nu\text{-integrable}\} \in \mathcal{A}; \tag{A.72a}$$

the function $f(\cdot, \omega_2)$ is μ-integrable for ν-almost every $\omega_2 \in \Omega_2$ – in particular,

$$B := \{\omega_2 \in \Omega_2 : f(\cdot, \omega_2) \text{ is not } \mu\text{-integrable}\} \in \mathcal{B}; \tag{A.72b}$$

the functions given by

$$\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)\, \mathrm{d}\nu(\omega_2), \quad \omega_2 \mapsto \int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{d}\mu(\omega_1), \tag{A.73}$$

are μ-integrable over $X \setminus A$ (resp. ν-integrable over $Y \setminus B$) and

$$\begin{aligned} &\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\, \mathrm{d}(\mu \otimes \nu)(\omega_1, \omega_2) \\ &= \int_{\Omega_2 \setminus B} \int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{d}\mu(\omega_1)\, \mathrm{d}\nu(\omega_2) \\ &= \int_{\Omega_1 \setminus A} \int_{\Omega_2} f(\omega_1, \omega_2)\, \mathrm{d}\nu(\omega_2)\, \mathrm{d}\mu(\omega_1). \end{aligned} \tag{A.74}$$

(c) *If $f : \Omega_1 \times \Omega_2 \longrightarrow \hat{\mathbb{K}}$ is $(\mu \otimes \nu)$-measurable and one of the integrals*

$$\begin{gathered} \int_{\Omega_1 \times \Omega_2} |f|\, \mathrm{d}(\mu \otimes \nu), \\ \int_{\Omega_2} \int_{\Omega_1} |f(\omega_1, \omega_2)|\, \mathrm{d}\mu(\omega_1)\, \mathrm{d}\nu(\omega_2), \quad \int_{\Omega_1} \int_{\Omega_2} |f(\omega_1, \omega_2)|\, \mathrm{d}\nu(\omega_2)\, \mathrm{d}\mu(\omega_1), \end{gathered} \tag{A.75}$$

is finite, then all three integrals are finite and equal, f is $(\mu \otimes \nu)$-integrable and, in particular, all the assertions of (b) *hold.*

Proof. See, e.g., [Els07, Sec. V.§2]. ■

B Probability Theory

B.1 Basic Concepts and Terminology

B.1.1 Probability Space, Random Variables, Distribution

Definition B.1. A measure space $(\Omega, \mathcal{A}, P)$ is called a *probability space* if, and only if, $P(\Omega) = 1$. Then Ω is called the *sample space*, elements A of $\mathcal{A}$ are called *events*, P is called a *probability measure* or *probability distribution*, and $P(A)$ is called the *probability* of the event A.

Definition B.2. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(\Omega', \mathcal{A}')$ a measurable space.

(a) A function $X : \Omega \longrightarrow \Omega'$ is called a $(\Omega', \mathcal{A}')$-*random variable* if, and only if, X is $\mathcal{A}$-$\mathcal{A}'$-measurable. If $\mathcal{A}'$ is understood, then X is just called an Ω'-valued random variable, for $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}^1)$, $(\Omega', \mathcal{A}') = (\overline{\mathbb{R}}, \overline{\mathcal{B}})$, or $(\Omega', \mathcal{A}') = (\mathbb{C}, \mathcal{B}^2)$ just random variable; and for $(\Omega', \mathcal{A}') = (\mathbb{R}^n, \mathcal{B}^n)$ or $(\Omega', \mathcal{A}') = (\mathbb{C}^n, \mathcal{B}^{2n})$, $n \in \mathbb{N}$, an n-dimensional random variable or *random vector*.

(b) The *distribution* of a random variable $X : \Omega \longrightarrow \Omega'$ is the pushforward measure $P_X = X(P)$ on $(\Omega', \mathcal{A}')$ (cf. Def. A.34 and Prop. A.33), i.e.

$$P_X(B) = \big(X(P)\big)(B) = P\big(X^{-1}(B)\big) \quad \text{for each } B \in \mathcal{A}'. \tag{B.1}$$

If P_X is the distribution of X, then one also says that X is P_X-distributed and writes $X \sim P_X$. A family of random variables $(X_i)_{i \in I}$ is called *identically distributed* if, and only if, they all have the same distribution, i.e. if, and only if, $P_{X_i} = P_{X_j}$ for all $i, j \in I$. If $P_{X_i} = P_X$ for all $i \in I$, then one sometimes calls the X_i identically distributed *copies* of X.

B.1.2 Expected Value, Moments, Standard Variation, Variance

Definition B.3. Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $X : \Omega \longrightarrow \hat{\mathbb{K}}$ be a random variable.

(a) If $X \geq 0$ (in particular $\overline{\mathbb{R}}$-valued) or X is integrable, then

$$E(X) := \int_\Omega X \, \mathrm{d}P \in \hat{\mathbb{K}} \tag{B.2}$$

is called the *expected value* of X.

(b) Let $p \in [1, \infty[$, $\alpha \in \mathbb{K}$. For each $X \in L^p(P, \mathbb{K})$, we call

$$E(|X - \alpha|^p) \in \mathbb{R}_0^+$$

the pth *absolute moment* of X centered at α.

(c) For each $X \in L^1(P, \mathbb{K})$, we call

$$V(X) := E\big((X - E(X))^2\big) \in [0, \infty] \tag{B.3}$$

the *variance* of X and

$$\sigma(X) := \sqrt{V(X)} \in [0, \infty] \tag{B.4}$$

the *standard deviation* of X. It is also customary to write $\sigma^2(X)$ instead of $V(X)$.

Definition B.4. Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $X : \Omega \longrightarrow \mathbb{K}^n$, $n \in \mathbb{N}$, be a random vector. If $X \in L^1(P, \mathbb{K}^n)$, then

$$E(X) := \big(E(X_1), \ldots, E(X_n)\big) \tag{B.5}$$

is called the *expectation vector* of X.

Theorem B.5. *Let $(\Omega, \mathcal{A}, P)$ be a measure space, let $X : \Omega \longrightarrow \overline{\mathbb{R}}$ be measurable.*

(a) Markov's inequality *holds for each $p, \alpha \in \mathbb{R}^+$:*

$$P\big(\{|X| \geq \alpha\}\big) \leq \frac{1}{\alpha^p} \int_\Omega |X|^p \, \mathrm{d}P\,, \tag{B.6}$$

where, as is customary, $\{|X| \geq \alpha\}$ was written instead of $\{\omega \in \Omega : |X(\omega)| \geq \alpha\}$.

(b) *The* Chebyshev inequality *holds for each $\alpha \in \mathbb{R}^+$, provided that $(\Omega, \mathcal{A}, P)$ is a probability space and $X \in L^1(P)$:*

$$P\big(\{|X - E(X)| \geq \alpha\}\big) \leq \frac{1}{\alpha^2} V(X). \tag{B.7}$$

Proof. (a): For each $\alpha > 0$, we have $A_\alpha := \{|X| \geq \alpha\} \in \mathcal{A}$ and compute

$$\int_\Omega |X|^p \, \mathrm{d}P \;\geq \int_{A_\alpha} |X|^p \, \mathrm{d}P \;\geq \int_{A_\alpha} \alpha^p \, \mathrm{d}P \;= \alpha^p \, P(A_\alpha), \tag{B.8}$$

proving (a).

(b) immediately follows from (a) by applying Markov's inequality with $p = 2$ and X replaced by $X - E(X)$. ■

B.1.3 Independence

Definition B.6. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $I \neq \emptyset$ an index set. The family $(A_i)_{i \in I}$ of events from $\mathcal{A}$ is called *independent* if, and only if, for each nonempty finite subset of I with distinct elements $i_1, \ldots, i_n$:

$$P(A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_{i_1}) \cdots P(A_{i_n}). \tag{B.9}$$

The events $(A_i)_{i \in I}$ are called *pairwise independent* if, and only if, (B.9) holds for each two-element subset of I with elements i_1, i_2.

Example B.7. Simple examples show that, in general, pairwise independence does not imply independence. The following standard example arises from modeling rolling a fair die independently for two consecutive times: Let A_1 (resp. A_2) be the event that the first (resp. the second) rolling resulted in an odd number, and let A_3 be the event that the sum of both rollings was odd. Then the events are pairwise independent, but not independent: Let $(\Omega, \mathcal{A}, P)$ be the probability space, where $\Omega := \{1, 2, 3, 4, 5, 6\}^2$, $\mathcal{A} := \mathcal{P}(\Omega)$, $P\{(i, j)\} := 1/36$ for each $(i, j) \in \Omega$,

$$A_1 := \{(i, j) \in \Omega : i \text{ is odd}\}, \tag{B.10a}$$
$$A_2 := \{(i, j) \in \Omega : j \text{ is odd}\}, \tag{B.10b}$$
$$A_3 := \{(i, j) \in \Omega : i + j \text{ is odd}\}. \tag{B.10c}$$

Since

$$P(A_1 \cap A_2) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_1) \cdot P(A_2), \tag{B.11a}$$

$$P(A_1 \cap A_3) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_1) \cdot P(A_3), \tag{B.11b}$$

$$P(A_2 \cap A_3) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_2) \cdot P(A_3), \tag{B.11c}$$

A_1, A_2, A_3 are pairwise independent. However, since

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(A_1) \cdot P(A_2) \cdot P(A_3), \tag{B.12}$$

A_1, A_2, A_3 are not independent.

—

For mostly technical reasons, it turns out that the following generalization of Def. B.6 is useful:

Definition B.8. Let $(\Omega, \mathcal{A}, P)$ be a probability space and I an index set. The family $(\mathcal{E}_i)_{i\in I}$ of sets $\emptyset \neq \mathcal{E}_i \subseteq \mathcal{A}$ is called *independent* if, and only if, for each nonempty finite subset of I with distinct elements $i_1, \ldots, i_n$ and each possible choice $A_{i_\nu} \in \mathcal{E}_{i_\nu}$, $\nu \in \{1, \ldots, n\}$, the equality (B.9) is valid.

Definition B.9. Let $(\Omega, \mathcal{A}, P)$ be a probability space, I an index set, and, for each $i \in I$, $(\Omega_i, \mathcal{A}_i)$ measurable spaces, $X_i : \Omega \longrightarrow \Omega_i$ random variables. Then the family $(X_i)_{i\in I}$ is called *independent* if, and only if, the family $(\sigma(X_i))_{i\in I}$ of generated σ-algebras is independent (recall from Rem. A.32(b) that $\sigma(X_i) = X_i^{-1}(\mathcal{A}_i)$).

—

It is tremendously useful that independence is always preserved under compositions:

Theorem B.10. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, I an index set; for each $i \in I$, let $(\Omega_i, \mathcal{A}_i)$ and $(\Omega_i', \mathcal{A}_i')$ be measurable spaces, $X_i : \Omega \longrightarrow \Omega_i$ random variables, and $Y_i : \Omega_i \longrightarrow \Omega_i'$ measurable maps. Then the independence of the family $(X_i)_{i\in I}$ implies the independence of the family $(Y_i \circ X_i)_{i\in I}$.*

Proof. For each $i \in I$ and each $B \in \mathcal{A}_i'$, one has $(Y_i \circ X_i)^{-1}(B) = X_i^{-1}\big(Y_i^{-1}(B)\big)$, implying $\sigma(Y_i \circ X_i) = (Y_i \circ X_i)^{-1}(\mathcal{A}_i') \subseteq X_i^{-1}(\mathcal{A}_i) = \sigma(X_i)$. Thus, if $\big(\sigma(X_i)\big)_{i\in I}$ is independent, then so is $\big(\sigma(Y_i \circ X_i)\big)_{i\in I}$. ■

Theorem B.11. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\Omega_1, \mathcal{A}_1), \ldots, (\Omega_n, \mathcal{A}_n)$ measurable spaces, $n \in \mathbb{N}$, and $X_i : \Omega \longrightarrow \Omega_i$ random variables, $i = 1, \ldots, n$. Then the finite family $X_1, \ldots, X_n$ is independent if, and only if,*

$$P\{X_1 \in A_1, \ldots, X_n \in A_n\} = \prod_{i=1}^{n} P\{X_i \in A_i\} \tag{B.13}$$

for each possible choice $A_i \in \mathcal{A}_i$. The statement remains valid if the last part is replaced by "for each possible choice $A_i \in \mathcal{Q}_i$", where, for each $i = 1, \ldots, n$, $\mathcal{Q}_i \subseteq \mathcal{P}(\Omega_i)$ is a $\cap$-stable generator of $\mathcal{A}_i$ (i.e. $\mathcal{A}_i = \sigma_{\Omega_i}(\mathcal{Q}_i)$ and $\mathcal{Q}_i$ is $\cap$-stable (cf. Def. A.24*)).*

Proof. See, e.g., [Bau02, Th. 7.2]. ■

Theorem B.12. *Let $(\Omega, \mathcal{A}, P)$ be a probability space. If $X_1, \ldots, X_n$, $n \in \mathbb{N}$, are $\mathbb{K}$-valued, independent random variables on Ω such that all $X_i \geq 0$ (in particular, $\mathbb{R}$-valued) or all X_i are integrable, then*

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i). \tag{B.14}$$

In particular, if all X_i are integrable, then so is the product $\prod_{i=1}^{n} X_i$.

Proof. See [Bau02, Th. 8.1] for the $\mathbb{R}$-valued case and [Bau02, p. 185] for the $\mathbb{C}$-valued case. ■

Definition B.13. Let $(\Omega, \mathcal{A}, P)$ be a probability space. If X, Y are $\mathbb{K}$-valued, integrable random variables on Ω such that XY is also integrable, then define the number

$$\mathrm{Cov}(X, Y) := E\Big((X - E(X))(Y - E(Y))\Big) = E(XY) - E(X)E(Y), \tag{B.15}$$

called the *covariance* of X and Y. Moreover, X and Y are called *uncorrelated* if, and only if, $\mathrm{Cov}(X, Y) = 0$.

Remark B.14. According to Th. B.12, if X and Y are independent, then they are also uncorrelated. However, simple examples, such as the following Ex. B.15, show that the converse is not true.

Example B.15. Let $(\Omega, \mathcal{A}, P)$ be the probability space, where $\Omega := \{1, 2, 3\}$, $\mathcal{A} := \mathcal{P}(\Omega)$, and $P\{i\} = 1/3$ for $i = 1, 2, 3$. Moreover, define

$$X : \Omega \longrightarrow \mathbb{R}, \quad X(i) := \begin{cases} 1 & \text{for } i = 1, \\ 0 & \text{for } i = 2, \\ -1 & \text{for } i = 3, \end{cases} \tag{B.16}$$

$$Y : \Omega \longrightarrow \mathbb{R}, \quad Y(i) := \begin{cases} 0 & \text{for } i = 1, \\ 1 & \text{for } i = 2, \\ 0 & \text{for } i = 3. \end{cases} \tag{B.17}$$

Since

$$E(XY) = 0 = E(X) = E(X)E(Y), \tag{B.18}$$

X, Y are uncorrelated. However, since

$$P\{X = 1, Y = 1\} = 0 \neq \frac{1}{3} \cdot \frac{1}{3} = P\{X = 1\} \cdot P\{Y = 1\} \tag{B.19}$$

and Th. B.11, X, Y are not independent.

B.1.4 Product Spaces

Definition B.16. Let $(\Omega, \mathcal{A}, P)$ be a probability space, I an index set, and, for each $i \in I$, $(\Omega_i, \mathcal{A}_i)$ measurable spaces, $X_i : \Omega \longrightarrow \Omega_i$ random variables. The X_i are called *identically distributed* if, and only if, they all have the same distribution, i.e. $P_{X_i} = P_{X_j}$ for all $i, j \in I$. If the family $(X_i)_{i\in I}$ is also independent, then the X_i are called *independent identically distributed (i.i.d.)*.

—

It is a remarkable and nontrivial result that i.i.d. families of every cardinality and of every distribution exist. This is related to the fact that one can form products of arbitrarily many probability spaces (see Cor. B.18 below).

Theorem B.17. *Let I be an index set and let $(\Omega_i, \mathcal{A}_i, P_i)_{i\in I}$ be a family of probability spaces. Moreover, let (cf. Def.* A.75*)*

$$(\Omega, \mathcal{A}), \quad \text{where} \quad \Omega := \prod_{i\in I} \Omega_i, \quad \mathcal{A} := \bigotimes_{i\in I} \mathcal{A}_i. \tag{B.20}$$

For each finite $J \subseteq I$, the projection

$$\pi_J : \Omega \longrightarrow \prod_{j\in J} \Omega_j, \quad \pi_J(\omega_i)_{i\in I} = (\omega_j)_{j\in J}, \tag{B.21}$$

is $\mathcal{A}$-$\otimes_{j\in J}\mathcal{A}_i$-measurable and there exists a unique measure P on $(\Omega, \mathcal{A})$ (called the product measure *of the P_i, also denoted $\otimes_{i\in I} P_i := P$), satisfying*

$$\underset{J\subseteq I:\, \#J<\infty}{\forall} \quad P_{\pi_J} = \otimes_{j\in J} P_j, \tag{B.22}$$

where $\otimes_{j\in J} P_j$ denotes the unique product measure of the $(P_j)_{j\in J}$ given by Th. A.85(b). *Moreover, P is a probability measure and*

$$\underset{\substack{J\subseteq I:\, \#J<\infty,\\ (A_j)_{j\in J}\in\prod_{j\in J}\mathcal{A}_j}}{\forall} \quad P\left(\prod_{j\in J} A_j \times \prod_{i\in I\setminus J} \Omega_i\right) = \prod_{j\in J} P_j(A_j). \tag{B.23}$$

Proof. See, e.g., [Bau02, Th. 9.2]. ■

Corollary B.18. *Let I be an index set. For each family $(\Omega_i, \mathcal{A}_i, P_i)_{i\in I}$ of probability spaces, there exists a probability space $(\Omega, \mathcal{A}, P)$ and an independent family $(X_i)_{i\in I}$ of random variables $X_i : \Omega \longrightarrow \Omega_i$ such that*

$$\underset{i\in I}{\forall} \quad P_{X_i} = P_i, \tag{B.24}$$

namely the product space with $(\Omega, \mathcal{A}, P)$ as defined in Th. B.17*, where the X_i are given by the projections $X_i = \pi_i : \Omega \longrightarrow \Omega_i$. In particular, choosing all $(\Omega_i, \mathcal{A}_i, P_i)$ to be the same probability space $(\Omega_1, \mathcal{A}_1, P_1)$ yields an i.i.d. family of random variables $(X_i)_{i\in I}$ with distribution P_1.*

Theorem B.19. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $I \neq \emptyset$ an index set, $(\Omega', \mathcal{A}')$ a measurable space, and $(X_i)_{i\in I}$ a family of independent random variables $X_i : \Omega \longrightarrow \Omega'$ satisfying*

$$\underset{i\in I}{\forall} \quad \underset{A_i\in\mathcal{A}'}{\exists} \quad 0 < P\{X_i \in A_i\} < 1. \tag{B.25}$$

Moreover, let $J \neq \emptyset$ be another index set, $(I_j)_{j\in J}$ a family of subsets $I_j \subseteq I$, where all I_j have the same cardinality $\#I_j = \kappa \neq \emptyset$, K a reference set with $\#K = \kappa$, and, for each $j \in J$, let $\phi_j : I_j \longrightarrow K$ be a bijection. If

$$\underset{j\in J}{\forall} \quad Y_j : \Omega \longrightarrow (\Omega')^K, \quad Y_j := (X_{\phi_j^{-1}(k)})_{k\in K}, \tag{B.26}$$

then the family $(Y_j)_{j\in J}$ is independent if, and only if, the I_j are pairwise disjoint.

Proof. Exercise. ■

B.1.5 Condition

Definition B.20. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $B \in \mathcal{A}$, and $P(B) > 0$.

(a) The map

$$P_B := P|B : \mathcal{A} \longrightarrow [0,1], \quad P_B(A) := (P|B)(A) := \frac{P(A\cap B)}{P(B)}, \tag{B.27}$$

is called the *conditional probability under the hypothesis B*.

(b) If $(\tilde\Omega, \tilde{\mathcal{A}}, \tilde P)$ is a probability space, $(\Omega', \mathcal{A}')$ is a measurable space and $X : \Omega \longrightarrow \Omega'$, $Y : \tilde\Omega \longrightarrow \Omega'$ are random variables, then Y is said to be distributed according to X under the condition B (denoted $Y \sim X|B$) if, and only if, $Y(\tilde P) = X(P_B)$, i.e. if, and only if,

$$\underset{A\in\mathcal{A}'}{\forall} \quad \tilde P\{Y \in A\} = \tilde P\big(Y^{-1}(A)\big) = P_B\big(X^{-1}(A)\big) = \frac{P\big(B \cap \{X \in A\}\big)}{P(B)}. \tag{B.28}$$

Proposition B.21. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $B \in \mathcal{A}$, and $P(B) > 0$.*

(a) *The conditional probability P_B as defined in* (B.27) *constitutes a probability measure on $(\Omega, \mathcal{A})$.*

(b) *The restriction $P_B : \mathcal{A}|B \longrightarrow [0,1]$ constitutes a probability measure on $(B, \mathcal{A}|B)$.*

(c) *Let $(\Omega', \mathcal{A}')$ be a measurable space and $X : \Omega \longrightarrow \Omega'$ a random variable. Then, considering the probability space $(B, \mathcal{A}|B, P_B)$,*

$$Y : B \longrightarrow \Omega', \quad Y := X\restriction_B, \tag{B.29}$$

is a random variable satisfying $Y \sim X|B$.

Proof. (a): Since $B \in \mathcal{A}$, we have $A \cap B \in \mathcal{A}$ for each $A \in \mathcal{A}$, i.e. P_B is well-defined. Moreover, $P_B(\emptyset) = 0$ and $P_B(\Omega) = 1$ are both immediate from (B.27). If $(A_n)_{n\in\mathbb{N}}$ is a sequence of pairwise disjoint sets in $\mathcal{A}$, then

$$P_B\left(\bigcup_{n=1}^{\infty} A_n\right) = \frac{1}{P(B)}\, P\left(B \cap \bigcup_{n=1}^{\infty} A_n\right) = \frac{1}{P(B)}\, P\left(\bigcup_{n=1}^{\infty} A_n \cap B\right)$$
$$= \frac{1}{P(B)} \sum_{n=1}^{\infty} P(A_n \cap B) = \sum_{n=1}^{\infty} P_B(A_n), \tag{B.30}$$

verifying the σ-additivity of P_B.

(b) is an immediate consequence of Prop. A.18 and $P_B(B) = 1$.

(c): If $A \in \mathcal{A}'$, then $Y^{-1}(A) = B \cap X^{-1}(A) \in \mathcal{A}|B$, since X is $\mathcal{A}$-measurable. Thus, Y is $\mathcal{A}|B$-measurable. To verify $Y \sim X|B$, we calculate

$$\underset{A\in\mathcal{A}'}{\forall} \quad P_B\{Y \in A\} = P_B\big(Y^{-1}(A)\big) = P_B\big(B \cap X^{-1}(A)\big) = \frac{P\big(B \cap \{X \in A\}\big)}{P(B)}, \tag{B.31}$$

which establishes the case. ■

B.1.6 Convergence

Definition B.22. Let $(\Omega, \mathcal{A}, P)$ be a measure space, let (Ω', τ) be a topological space, and let $X, X_n : \Omega \longrightarrow \Omega'$, $n \in \mathbb{N}$.

(a) The X_n converge to X *pointwise P-almost everywhere* if, and only if, there exists a P-null set $N \subseteq \Omega$ such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for each $\omega \in \Omega \setminus N$. If $(\Omega, \mathcal{A}, P)$ is a probability space, then one usually says that the X_n converge to X *almost surely* or *with probability* 1 (note that measurability of X_n, X is actually not needed here).

(b) Let $p \in [1, \infty[$ and assume $X_n, X \in L^p(P, \mathbb{K})$. Then the X_n converge to X *in the pth mean* or simply in $L^p(P)$ if, and only if,

$$\lim_{n\to\infty} \|X - X_n\|_p^p = \lim_{n\to\infty} \int_\Omega |X - X_n|^p \,\mathrm{d}P = 0. \tag{B.32}$$

If $(\Omega, \mathcal{A}, P)$ is a probability space, then (B.32) can be written as

$$\lim_{n\to\infty} E\big(|X - X_n|^p\big) = 0. \tag{B.33}$$

(c) Let $X_n, X : \Omega \longrightarrow \mathbb{K}$ be measurable. The X_n converge to X *in measure* if, and only if,

$$\lim_{n\to\infty} P\big(\{|X - X_n| \geq \alpha\} \cap A\big) = 0 \quad \text{for each } \alpha > 0 \text{ and } A \in \mathcal{A} \text{ with } P(A) < \infty. \tag{B.34}$$

If $(\Omega, \mathcal{A}, P)$ is a probability space, then (B.34) is equivalent to

$$\lim_{n\to\infty} P\big(\{|X - X_n| \geq \alpha\}\big) = 0 \quad \text{for each } \alpha > 0, \tag{B.35}$$

and one says that the X_n converge to X *in probability*.

Notation B.23. Let $C_{\mathrm{b}}(\mathbb{R}^n)$ denote the set of all continuous and bounded real-valued functions on $\mathbb{R}^n$, $n \in \mathbb{N}$.

Definition B.24. (a) Let $(\mathbb{R}^n, \mathcal{B}^n, P)$, $(\mathbb{R}^n, \mathcal{B}^n, P_i)$, $i \in \mathbb{N}$, be probability spaces, $n \in \mathbb{N}$. Then the P_i converge to P *weakly* (denoted $\lim_{i\to\infty} P_i = P$) if, and only if,

$$\lim_{i\to\infty} \int_{\mathbb{R}^n} f \,\mathrm{d}P_i = \int_{\mathbb{R}^n} f \,\mathrm{d}P \quad \text{for each } f \in C_{\mathrm{b}}(\mathbb{R}^n). \tag{B.36}$$

(b) Let $(\Omega, \mathcal{A}, P)$ be a probability space and X, X_i, $i \in \mathbb{N}$, $\mathbb{R}^n$-valued random variables, $n \in \mathbb{N}$. The X_i converge to X (or, more generally, to a probability measure μ on $\mathcal{B}^n$) *in distribution* if, and only if, the distributions P_{X_i} converge weakly to P_X (or, more generally, to μ).

Theorem B.25. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $X : \Omega \longrightarrow \mathbb{R}$, $X_i : \Omega \longrightarrow \mathbb{R}$, $i \in \mathbb{N}$, be random variables. If the X_i converge to X in probability, then they converge to X in distribution.*

Proof. See, e.g., [Bau02, Th. 5.1]. ■

B.1.7 Density and Distribution Functions

Definition B.26. Let $(\Omega, \mathcal{A}, P)$ be a probability space, let $(\Omega', \mathcal{A}', \mu)$ be a measure space, and let $f : \Omega' \longrightarrow [0, \infty]$ be measurable.

(a) If P' is a probability measure on $(\Omega', \mathcal{A}')$, then f is called a *probability density function (PDF)* for P' with respect to μ if, and only if, $P' = f\mu$, i.e. if, and only if, f is a density for P' with respect to μ in the sense of Def. A.68, i.e. if, and only if,

$$\forall_{B\in\mathcal{A}'} \quad P'(B) = \int_B f \,\mathrm{d}\mu\,. \tag{B.37a}$$

(b) If $X : \Omega \longrightarrow \Omega'$ is a random variable, then f is called a *probability density function (PDF)* of X and one says X is *distributed* according to f if, and only if, f is a PDF for the distribution of X in the sense of (a), i.e. if, and only if,

$$\forall_{B\in\mathcal{A}'} \quad P_X(B) = P\{X \in B\} = \int_B f \,\mathrm{d}\mu\,. \tag{B.37b}$$

Corollary B.27. *In the situation of Def.* B.26*, let the measure μ be σ-finite. Then the probability measure P' (resp. the random variable X) has a PDF with respect to μ if, and only if, P' (resp. P_X) is absolutely continuous with respect to μ (cf. Def.* A.69*). Moreover, the density is unique μ-almost everywhere.*

Proof. Existence is given by the Radon-Nikodym Th. A.72 (and Lem. A.70). Uniqueness is given by Th. A.73. ■

Definition B.28. Let P be a probability measure on $(\mathbb{R}, \mathcal{B}^1)$.

(a) The function

$$F_{P,\mathrm{r}} : \mathbb{R} \longrightarrow [0,1], \quad F_{P,\mathrm{r}}(x) := P]-\infty, x], \tag{B.38a}$$

is called the *right-continuous (r.c.) cumulative distribution function (CDF)* or just *(r.c.) distribution function* of P (cf. Th. B.29 below). Sometimes it is convenient to extend the r.c. CDF to $\overline{\mathbb{R}}$ by defining

$$F_{P,\mathrm{r}} : \overline{\mathbb{R}} \longrightarrow [0,1], \quad F_{P,\mathrm{r}}(x) := \begin{cases} 0 & \text{for } x = -\infty, \\ P]-\infty, x] & \text{for } x \in \mathbb{R}, \\ 1 & \text{for } x = \infty. \end{cases} \tag{B.38b}$$

(b) The function

$$F_{P,\mathrm{l}} : \mathbb{R} \longrightarrow [0,1], \quad F_{P,\mathrm{l}}(x) := P]-\infty, x[, \tag{B.39a}$$

is called the *left-continuous (l.c.) cumulative distribution function (CDF)* or just *(l.c.) distribution function* of P (cf. Th. B.29 below). Sometimes it is convenient to extend the l.c. CDF to $\overline{\mathbb{R}}$ by defining

$$F_{P,\mathrm{l}} : \overline{\mathbb{R}} \longrightarrow [0,1], \quad F_{P,\mathrm{l}}(x) := \begin{cases} 0 & \text{for } x = -\infty, \\ P]-\infty, x[& \text{for } x \in \mathbb{R}, \\ 1 & \text{for } x = \infty. \end{cases} \tag{B.39b}$$

Theorem B.29. *Define*

$$\mathcal{F}_\uparrow := \big\{(f : \mathbb{R} \longrightarrow [0,1]) : f \text{ increasing}, \lim_{x\to-\infty} f(x) = 0, \lim_{x\to\infty} f(x) = 1\big\}, \tag{B.40a}$$

$$\mathcal{R}_\uparrow := \big\{f \in \mathcal{F}_\uparrow : f \text{ right-continuous}\big\}, \tag{B.40b}$$

$$\mathcal{L}_\uparrow := \big\{f \in \mathcal{F}_\uparrow : f \text{ left-continuous}\big\}. \tag{B.40c}$$

If $\mathcal{P}$ denotes the set of all probability measures on $(\mathbb{R}, \mathcal{B}^1)$, then the maps

$$R : \mathcal{P} \longrightarrow \mathcal{R}_\uparrow, \quad R(P) := F_{P,\mathrm{r}}, \tag{B.41a}$$

$$L : \mathcal{P} \longrightarrow \mathcal{L}_\uparrow, \quad L(P) := F_{P,\mathrm{l}}, \tag{B.41b}$$

are both bijective.

Moreover, if $P \in \mathcal{P}$, then both $F_{P,\mathrm{r}}$ and $F_{P,\mathrm{l}}$ are continuous at $x \in \mathbb{R}$ if, and only if, $P(\{x\}) = 0$, i.e. if, and only if, x is not a so-called atom.

Proof. See [Bau92, Ths. 6.5,6.6] for the bijectivity of (B.41b); the bijectivity of (B.41a) can be proved completely analogously.

Let $x \in \mathbb{R}$ and $P(\{x\})$. Consider an increasing sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ such that $\lim_{n\to\infty} x_n = x$. Then

$$\begin{aligned}\lim_{n\to\infty} F_{P,\mathrm{r}}(x_n) &= \lim_{n\to\infty} P]-\infty, x_n] = P]-\infty, x[\\ &= P]-\infty, x] \quad \text{if, and only if, } P(\{x\}) = 0, \end{aligned} \tag{B.42a}$$

showing that $F_{P,\mathrm{r}}$ is continuous if, and only if, $P(\{x\}) = 0$. Similarly, if $(x_n)_{n\in\mathbb{N}}$ is a decreasing sequence in $\mathbb{R}$ such that $\lim_{n\to\infty} x_n = x$. Then

$$\begin{aligned}\lim_{n\to\infty} F_{P,\mathrm{l}}(x_n) &= \lim_{n\to\infty} P]-\infty, x_n[= \lim_{n\to\infty} \big(1 - P[x_n, \infty[\big) = 1 - P]x, \infty[\\ &= 1 - P[x, \infty[\quad \text{if, and only if, } P(\{x\}) = 0, \end{aligned} \tag{B.42b}$$

showing that $F_{P,\mathrm{l}}$ is continuous if, and only if, $P(\{x\}) = 0$. ■

B.2 Important Theorems

B.2.1 Laws of Large Numbers

Definition B.30. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A sequence $(X_i)_{i\in\mathbb{N}}$ of $\mathbb{R}$-valued, integrable random variables on Ω is said to satisfy the *weak* (resp. *strong*) *law of large numbers* if, and only if,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \big(X_i - E(X_i)\big) = 0 \tag{B.43}$$

in the sense of convergence in probability (resp. in the sense of convergence almost surely).

Theorem B.31 (Khintchine, Weak Law of Large Numbers)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space. If a sequence $(X_i)_{i\in\mathbb{N}}$ of $\mathbb{R}$-valued, integrable, and pairwise uncorrelated random variables on Ω satisfies*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) = 0, \tag{B.44}$$

then it also satisfies the weak law of large numbers.

Proof. See, e.g., [Bau02, Th. 10.2]. ■

Theorem B.32 (Etemadi, Kolmogorov, Strong Law of Large Numbers)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space. Each sequence $(X_i)_{i\in\mathbb{N}}$ of $\mathbb{R}$-valued, integrable, identically distributed, and pairwise independent random variables on Ω satisfies the strong law of large numbers (Kolmogorov had proved the theorem under the stronger hypothesis that the entire sequence is independent).*

Proof. See, e.g., [Bau02, Th. 12.1]. ■

B.2.2 The Central Limit Theorem

Notation B.33. For each $\alpha \in \mathbb{R}$, $\sigma > 0$, let g_{α,σ^2} denote the function

$$g_{\alpha,\sigma^2} : \mathbb{R} \longrightarrow \mathbb{R}^+, \quad g_{\alpha,\sigma^2}(x) := (2\pi\sigma^2)^{-\frac{1}{2}}\, e^{-\frac{(x-\alpha)^2}{2\,\sigma^2}}. \tag{B.45}$$

Remark B.34. Recalling

$$(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x \;= 1, \tag{B.46}$$

a simple change of variables shows

$$\int_{-\infty}^{\infty} g_{\alpha,\sigma^2}(x)\,\mathrm{d}x \;= 1 \quad \text{for each } \alpha \in \mathbb{R},\, \sigma > 0. \tag{B.47}$$

Definition and Remark B.35. For each $\alpha \in \mathbb{R}$, $\sigma > 0$, the measure on $\mathcal{B}^1$ defined by

$$N(\alpha, \sigma^2) := \nu_{\alpha,\sigma^2} := g_{\alpha,\sigma^2}\,\lambda_1 \tag{B.48}$$

is called the *normal* or the *Gaussian* distribution on $\mathbb{R}$, centered in α and with variance σ^2. One calls $N(0,1)$ the *standard* normal distribution. From Rem. B.34, we know each $N(\alpha,\sigma^2)$ defines a probability measure on $\mathcal{B}^1$. If $(\Omega, \mathcal{A}, P)$ is a probability space and $X : \Omega \longrightarrow \mathbb{R}$ a random variable such that $P_X = N(\alpha,\sigma^2)$, then one says X is $N(\alpha,\sigma^2)$-distributed. For $N(\alpha,\sigma^2)$-distributed X, one checks that

$$E(X) = \alpha, \tag{B.49a}$$
$$V(X) = \sigma^2. \tag{B.49b}$$

Theorem B.36 (Central Limit Theorem)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $(X_i)_{i\in\mathbb{N}}$ be a sequence in $L^2(P)$, consisting of independent and identically distributed (i.i.d.) $\mathbb{R}$-valued random variables on Ω with $\sigma := \sigma(X_i) > 0$.*

(a) *It holds that*

$$\lim_{n\to\infty} \left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} \big(X_i - E(X_i)\big) \right)(P) = N(0,1), \tag{B.50}$$

i.e. the random variables occurring on the left-hand side of (B.50) *converge in distribution to the standard normal distribution.*

(b) *If $F_1, F_2, \ldots$ is the sequence of distribution functions corresponding to the distributions of the random variables occurring on the left-hand side of* (B.50) *and if F is the distribution function of $N(0,1)$, then the F_n converge to F uniformly on $\mathbb{R}$.*

Proof. (a): See, e.g., [Bau02, Th. 27.1].

(b) follows by combining (a) with [Bau92, Th. 30.13]. ■

C Stochastic Calculus

C.1 Itô's Formula and Integration by Parts

C.1.1 1-Dimensional Case

Definition C.1. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A pair of $\mathbb{R}$-valued stochastic processes $(a_t, b_t)_{t\geq 0}$, $a_t, b_t : \Omega \longrightarrow \mathbb{R}$ for each $t \in \mathbb{R}_0^+$, is called *Itô-admissible* if, and only if, the paths $t \mapsto a_t(\omega)$ are locally integrable almost surely, and the paths $t \mapsto b_t(\omega)$ are locally square-integrable almost surely, i.e. if, and only if,

$$P\left\{\omega \in \Omega : \underset{T\in\mathbb{R}_0^+}{\forall} \quad \int_0^T |a_t(\omega)|\,\mathrm{d}t \,<\infty\right\} = 1, \tag{C.1a}$$

and

$$P\left\{\omega \in \Omega : \underset{T\in\mathbb{R}_0^+}{\forall} \quad \int_0^T |b_t(\omega)|^2\,\mathrm{d}t \,<\infty\right\} = 1. \tag{C.1b}$$

Theorem C.2 (Itô's Formula)**.** *Let $(a_t, b_t)_{t\geq 0}$ be Itô-admissible stochastic processes. Moreover, let $O \subseteq \mathbb{R}$ be open. If the O-valued stochastic process $(Y_t)_{t\geq 0}$ is a solution to the SDE*

$$\mathrm{d}Y_t \,= a_t\,\mathrm{d}t \,+ b_t\,\mathrm{d}W_t\,, \tag{C.2}$$

where $(W_t)_{t\geq 0}$ denotes a 1-dimensional standard Brownian motion with drift 0 and variance 1, and

$$f : \mathbb{R}_0^+ \times O \longrightarrow \mathbb{R}, \quad (t,x) \mapsto f(t,x),$$

has continuous first partials with respect to t and continuous second partials with respect to x, then $(\tilde{Y}_t)_{t\geq 0}$, where $\tilde{Y}_t := f(t, Y_t)$ for each $t \in \mathbb{R}_0^+$, is a solution to the SDE

$$\mathrm{d}\tilde{Y}_t = \left(\partial_t f(t,Y_t) + a_t\,\partial_x f(t,Y_t) + b_t^2\,\frac{\partial_{xx} f(t,Y_t)}{2}\right)\,\mathrm{d}t \,+ b_t\,\partial_x f(t,Y_t)\,\mathrm{d}W_t\,. \tag{C.3}$$

The relation (C.3) *is known as* Itô's formula*.*

Proof. See, e.g., [KP99, Th. 3.3.2]. ■

We can use Itô's formula to prove a simple version of the *integration by parts* formula for Itô integrals (much more general versions exist in the literature):

Theorem C.3 (Integration by Parts)**.** *Let $\alpha : \mathbb{R}_0^+ \longrightarrow \mathbb{R}$ be locally integrable and let $\sigma : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ be locally square-integrable. If $(X_t)_{t\geq 0}$ denotes a 1-dimensional Brownian motion with drift α and variance σ^2, then the integration by parts formula reads*

$$\mathrm{d}(t\,X_t) = X_t\,\mathrm{d}t \,+ t\,\mathrm{d}X_t \tag{C.4a}$$

or, written in the more explicit integral form,

$$t\,X_t = \int_0^t X_s\,\mathrm{d}s \;+ \int_0^t s\,\alpha(s)\,\mathrm{d}s \;+ \int_0^t s\,\sigma(s)\,\mathrm{d}W_s \tag{C.4b}$$

with $(W_t)_{t\geq 0}$ *as in Th.* C.2*.*

Proof. As $(X_t)_{t\geq 0}$ is a 1-dimensional Brownian motion with drift α and variance σ^2, according to Rem. 4.5, it satisfies the SDE

$$\mathrm{d}X_t = \alpha(t)\,\mathrm{d}t \;+ \sigma(t)\,\mathrm{d}W_t. \tag{C.5}$$

We apply Itô's formula (C.3) with $Y_t = X_t$,

$$f : \mathbb{R}_0^+ \times \mathbb{R} \longrightarrow \mathbb{R}, \quad f(t,x) := t\,x, \tag{C.6}$$

i.e. $\tilde{Y}_t = t\,X_t$, obtaining

$$\mathrm{d}\tilde{Y}_t = \big(X_t + \alpha(t)\,t\big)\,\mathrm{d}t \;+ \sigma(t)\,t\,\mathrm{d}W_t\,, \tag{C.7}$$

which is precisely (C.4). ■

C.1.2 Multi-Dimensional Case

In generalization of Th. C.2 and (C.3), one has:

Theorem C.4 (Itô's Formula)**.** *Let* $d, m \in \mathbb{N}$*. Let* $(a_t, b_t)_{t\geq 0}$ *be Itô-admissible stochastic processes, where* $(a_t)_{t\geq 0}$ *is* $\mathbb{R}^d$*-valued and* $(b_t)_{t\geq 0}$ *is* $\mathbb{R}^{d\times m}$*-valued (this is supposed to mean* $(a_t, b_t)_{t\geq 0}$ *satisfy* (C.1) *with* $|a_t(\omega)|$ *replaced by* $\|a_t(\omega)\|$ *and* $|b_t(\omega)|$ *replaced by* $\|b_t(\omega)\|$*, respectively, i.e.* $(a_t)_{t\geq 0}$ *has locally integrable paths and* $(b_t)_{t\geq 0}$ *has locally square-integrable paths). Moreover, let* $O \subseteq \mathbb{R}^d$ *be open. If the* O*-valued stochastic process* $(Y_t)_{t\geq 0}$ *is a solution to the SDE*

$$\mathrm{d}Y_t \;= a_t\,\mathrm{d}t \;+ b_t\,\mathrm{d}W_t\,, \tag{C.8}$$

where $(W_t)_{t\geq 0}$ *denotes an* m*-dimensional standard Brownian motion with drift* 0 *and covariance matrix* Id*, and*

$$f : \mathbb{R}_0^+ \times O \longrightarrow \mathbb{R}, \quad (t,x) \mapsto f(t,x),$$

has continuous first partials with respect to t *and continuous second partials with respect to* x*, then the (*1*-dimensional, i.e.* $\mathbb{R}$*-valued) process* $(\tilde{Y}_t)_{t\geq 0}$*, where* $\tilde{Y}_t := f(t, Y_t)$ *for each* $t \in \mathbb{R}_0^+$*, is a solution to the SDE*

$$\mathrm{d}\tilde{Y}_t = \partial_t f(t,Y_t) + \sum_{i=1}^d \partial_{x_i} f(t,Y_t)\,\mathrm{d}(Y_i)_t \;+ \frac{1}{2}\sum_{i,j=1}^d \partial_{x_i}\partial_{x_j} f(t,Y_t)\,\Sigma_{t,ij}\,\mathrm{d}t \tag{C.9a}$$

$$= \left(\partial_t f(t,Y_t) + \sum_{i=1}^d \partial_{x_i} f(t,Y_t)\,a_{t,i} + \frac{1}{2}\sum_{i,j=1}^d \partial_{x_i}\partial_{x_j} f(t,Y_t)\,\Sigma_{t,ij}\right)\mathrm{d}t$$

$$+ \sum_{i=1}^d \partial_{x_i} f(t,Y_t)\,b_{t,i\cdot}\,\mathrm{d}W_t\,, \tag{C.9b}$$

where $b_{t,i\cdot}$ denotes the ith row of b_t and

$$\underset{t\geq 0}{\forall} \quad \Sigma_t := b_t\, b_t^{\mathrm{t}} \in \mathbb{R}^{d\times d}. \tag{C.9c}$$

The relation (C.9) *is known as (the multi-dimensional version of)* Itô's formula.

Proof. Many textbooks, including [KP99], merely prove the 1-dimensional version of Itô's formula and state that the multi-dimensional version can be proved analogously. However, [HT94, Th. 4.46] does include a proof for the multi-dimensional version of Itô's formula formulated for semi-martingales, and our version constitutes a special case. ■

Remark C.5. Clearly, (C.9) reduces to (C.3) for $m = d = 1$.

References

[Bau92] Heinz Bauer. *Maß- und Integrationstheorie*, 2nd ed. Walter de Gruyter, Berlin, 1992 (German).

[Bau02] Heinz Bauer. *Wahrscheinlichkeitstheorie*, 5th ed. Walter de Gruyter, Berlin, 2002 (German).

[Beh87] E. Behrends. *Maß- und Integrationstheorie*. Springer-Verlag, Berlin, 1987 (German).

[BM58] G.E.P. Box and M.E. Muller. *A note on the generation of random normal deviates*. Annals of Mathematical Statistics **29** (1958), 610–611.

[CL97] R. Couture and P. L'Ecuyer. *Distribution Properties of Multiply-with-Carry Random Number Generators*. Mathematics of Computation **66** (1997), 591–607.

[Eat83] Morris L. Eaton. *Multivariate Statistics*. John Wiley & Sons, New York, 1983.

[Els07] Jürgen Elstrodt. *Maß- und Integrationstheorie*, 5th ed. Grundwissen Mathematik, Springer-Verlag, Berlin, 2007 (German).

[Geo09] Hans-Otto Georgii. *Stochastik*, 4th ed. Walter de Gruyter, Berlin, 2009 (German).

[Gla04] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Applications of Mathematics, Vol. 53, Springer Science + Business Media, New York, 2004.

[Hes93] H.I. Heston. *A closed-form solution for options with stochastic volatility with applications to bond and currency options*. Review of Financial Studies **6** (1993), 327–343.

[HT94] Wolfgang Hackenbroch and Anton Thalmaier. *Stochastische Analysis*. B. G. Teubner, Stuttgart, Germany, 1994 (German).

[Knu98] D.E. Knuth. *Seminumerical Algorithms*, 3rd ed. The Art of Computer Programming, Vol. 2, Addison-Wesley, Reading, MA, USA, 1998.

[Koe03] Max Koecher. *Lineare Algebra und analytische Geometrie*, 4th ed. Springer-Verlag, Berlin, 2003 (German), 1st corrected reprint.

[KP99] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Applications of Mathematics, Vol. 23, Springer Science + Business Media, Berlin, 1999, corrected 3rd printing.

[KS98] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*, 2nd ed. Graduate Texts in Mathematics, Vol. 113, Springer Science + Business Media, New York, 1998.

[Lev92] J.L. Leva. *A Fast Normal Random Number Generator*. ACM Tansactions on Mathematical Software **18** (1992), No. 4, 449–453.

[Mar68] G. Marsaglia. *Random Numbers Fall Mainly in the Planes*. Proceedings of the National Academy of Sciences **61** (1968), 25–28.

[Mar03a] G. Marsaglia. *Diehard Battery of Tests of Randomness v0.2 beta*. `http://www.cs.hku.hk/~diehard/`, 2003.

[Mar03b] G. Marsaglia. *Xorshift RNGs*. Journal of Statistical Software **8** (2003), No. 14, 1–6.

[MB64] G. Marsaglia and T.A. Bray. *A convenient method for generating normal variables*. SIAM Review **6** (1964), 260–264.

[Mil75] G.N. Milstein. *Approximate Integration of Stochastic Differential Equations*. Theory Probab. Appl. **19** (1975), 557–562.

[Øk03] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*, 6th ed. Springer-Verlag, Berlin, 2003.

[Phi23] P. Philip. *Numerical Mathematics I*. Lecture Notes, Ludwig-Maximilians-Universität, Germany, 2022/2023, AMS Open Math Notes Ref. # OMN:202204.111317, available in PDF format at `https://www.ams.org/open-math-notes/omn-view-listing?listingId=111317`.

[Pla10] Robert Plato. *Numerische Mathematik kompakt*, 4th ed. Vieweg Verlag, Wiesbaden, Germany, 2010 (German).

[PTVF07] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes. The Art of Scientific Computing*, 3rd ed. Cambridge University Press, New York, USA, 2007.

[RF10] Halsey Royden and Patrick Fitzpatrick. *Real Analysis*, 4th ed. Pearson Education, Boston, USA, 2010.