Compact Textbooks in Mathematics

Gunther Leobacher
Friedrich Pillichshammer

# Introduction to Quasi-Monte Carlo Integration and Applications

Birkhäuser

# Compact Textbooks in Mathematics

For further volumes:
http://www.springer.com/series/11225

Birkhäuser

## Compact Textbooks in Mathematics

This textbook series presents concise introductions to current topics in mathematics and mainly addresses advanced undergraduates and master students. The concept is to offer small books covering subject matter equivalent to 2- or 3-hour lectures or seminars which are also suitable for self-study. The books provide students and teachers with new perspectives and novel approaches. They feature examples and exercises to illustrate key concepts and applications of the theoretical contents. The series also includes textbooks specifically speaking to the needs of students from other disciplines such as physics, computer science, engineering, life sciences, finance.

Gunther Leobacher
Friedrich Pillichshammer

# Introduction to Quasi-Monte Carlo Integration and Applications

Birkhäuser

Gunther Leobacher
Institute of Financial Mathematics
University of Linz
Linz, Austria

Friedrich Pillichshammer
University of Linz
Linz, Austria

ISBN 978-3-319-03424-9 ISBN 978-3-319-03425-6 (eBook)
DOI 10.1007/978-3-319-03425-6
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014945104

2010 Mathematics Subject Classification : 11K06, 11K38, 11K45, 65C05, 65D30

*Cover design*: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*To Charlotte and Gisi*

# Preface

While Fubini's theorem states that the integral of a function on the $s$-dimensional unit cube can be computed simply by computing iterated integrals, the attempt of doing so for a function for which those integrals cannot be given by closed-form formulas is in most cases doomed to fail, if $s$, say, is greater than 10. The reason for this is that by iteration of a one-dimensional integration rule the number of function evaluation needed for the corresponding product rule grows exponentially in $s$.

This constraint on the practical computation of integrals led to the development of probabilistic methods. Here, the integral is interpreted as the expected value of the integrand evaluated at a random variable that is uniformly distributed on the $s$-dimensional unit cube. These methods were first applied by E. Fermi, S. Ulam and J. von Neumann, the latter also being the originator of the name "Monte Carlo simulation".

In contrast to Monte Carlo integration rules, which sample the integrand at random points, so-called *quasi*-Monte Carlo rules use deterministic sample points. The relationship between Monte Carlo and quasi-Monte Carlo corresponds to the relationship between two notions of "uniform distribution" in mathematics. The first is the probabilistic notion of a random variable for which the probability of taking values in a given subset of the unit cube is precisely the volume of that set. The second notion is that of a sequence of points, for which the proportion of points of the sequence lying in a given $s$-dimensional sub-interval of the unit cube equals the volume of the sub-interval.

While the notion of uniformly distributed sequences and examples thereof had been coined earlier, the birth of the theory of "Uniform Distribution Modulo One" is marked by H. Weyl's seminal paper "Über die Gleichverteilung mod. Eins", first published in the year 1916. It was already known by then that, at least in principle, uniformly distributed sequences could be used to integrate Riemann-integrable functions. However, under the weak assumptions the convergence of the sample mean to the integral can be arbitrarily slow, making the "method" impractical.

As the starting point for the analysis of quasi-Monte Carlo methods for numerical integration one can consider the establishment of the Koksma-Hlawka inequality, which was shown by J.F. Koksma in 1942 for the one-dimensional case and by E. Hlawka in 1961 for arbitrary dimensions. Since then the Koksma-Hlawka inequality is the prototypical error estimate for quasi-Monte Carlo integration. Its main feature is that it bounds the integration error by the product of two terms, the variation of the function and the star discrepancy of the underlying sample nodes. The second notion is related to that of uniform distribution of a sequence, but while the latter is an asymptotic quality, the star discrepancy allows to assess the quality of uniformity of a finite number of points. Knowing how well the points can be chosen with respect to that measure means – thanks to the Koksma-Hlawka inequality – knowing the possible convergence of the integration error. This is where concepts from Discrepancy Theory enter the game.

From the early 1960s on several people, among these N.M. Korobov, E. Hlawka, I.M. Sobol', J. Halton, H. Faure, H. Niederreiter and C.P. Xing provided constructions of point sets and sequences with excellent distribution properties, i.e., with low star discrepancy or related/alternative quality measures. The point sets and sequences constructed in this way are therefore suitable sample points for quasi-Monte Carlo rules. However, a certain disadvantageous dependence of the discrepancy bounds on the dimension led to the belief that quasi-Monte Carlo rules can only be applied in very moderate dimensions. Contrary to these opinions, quasi-Monte Carlo rules are nowadays used for numerical integration of functions in hundreds or even thousands of dimensions, and since recently there is also a stream of research which studies infinite-dimensional integration. The motivation for this paradigm change lies in results of numerical experiments published in 1995 by S.H. Paskov and J.F. Traub, who studied quasi-Monte Carlo rules for functions in 360 dimensions coming from Mathematical Finance. But, despite their apparent effectivity even for those very high-dimensional problems, the question of exactly *why* quasi-Monte Carlo rules should give these good results is still not completely resolved. In 2010, at the MCQMC meeting in Warsaw, I.H. Sloan spoke in this context about "The unreasonable effectiveness of quasi-Monte Carlo". Although in the meantime some partial answers come from the study of weighted function spaces and from tractability theory, the quest for an explanation of this unreasonable effectiveness of quasi-Monte Carlo is still a very active part of research.

As suggested by its title, this book is an introductory text to quasi-Monte Carlo methods and some of their applications, and it aims at giving a comprehensible treatment of the subject with detailed explanations of the basic concepts. Originating from a 2-h one semester undergraduate course, it should be accessible to students in mathematics or computer science with basic knowledge of algebra, calculus, linear algebra, and probability theory. Although the main focus is on the theory behind the concepts of quasi-Monte Carlo, several practical applications with an emphasis on financial problems are discussed.

The topics of the book roughly retrace the history of quasi-Monte Carlo methods as sketched above, but do so using up-to-date concepts and notations. Thus we start with the classical multi-dimensional integration problem and its first high-dimensional alternative, Monte Carlo integration. Chapter 2 is devoted to uniform distribution of sequences and several concepts of discrepancy. We give a discrepancy estimate for one of the oldest specimens of low discrepancy sequences, the Halton sequence. In Chap. 3 we introduce the modern framework of reproducing kernel Hilbert spaces for obtaining bounds on the integration error for functions in those spaces. The Koksma-Hlawka inequality, though not in its most general form, appears as a special case of that theory. The next two chapters are mostly devoted to constructions of low-discrepancy point sets and sequences, namely lattice point sets, $(t, m, s)$-nets and $(t, s)$-sequences. The chapter on lattice rules includes a section on integration in weighted Korobov spaces. The concept of weighted spaces has some bearing on the issue of effectiveness of quasi-Monte Carlo methods for very high-dimensional problems. Chapter 6 concludes the theoretical part by providing more information about the curse of dimensionality and tractability of discrepancy.

The last two chapters constitute the application part of the book. Chapter 7 gives a very condensed introduction to concepts from Mathematical Finance, in particular derivative pricing. We introduce some models and derivatives that can serve as specimens for trying out the simulation methods provided in Chap. 8. This last chapter covers some of the basics of simulation, like generation of non-uniform random variables and generation of Brownian paths. The emphasis is on (fast-)orthogonal transforms for speeding up convergence. Several examples serve to illustrate the methods.

The compilation of a textbook demands a great deal not only from the authors, but also from their families, colleagues, and students, some of the common time of which has to be diverted to the project. We want to thank all of them for their support and understanding.

We appreciate valuable comments, suggestions and improvements from several colleagues which we would like to mention here: Josef Dick, Aicke Hinrichs, Peter Kritzer, Gerhard Larcher, Harald Niederreiter, Klaus Ritter and Wolfgang Ch. Schmid.

We hope that the book will turn out to be useful for teaching, self-study, and as a reference, and that it will encourage many people to study quasi-Monte Carlo methods and/or apply them to problems from Mathematical Finance or other areas.

Linz, Austria
October 2013

Gunther Leobacher
Friedrich Pillichshammer

# Contents

# Introduction 1

## 1.1 The Univariate and Multivariate Integration Problem

In this book we consider the problem of numerical integration over the $s$-dimensional unit cube $[0,1]^s$,

$$\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_0^1 \cdots \int_0^1 f(x_1,\ldots,x_s)\,\mathrm{d}x_1\ldots\mathrm{d}x_s. \tag{1.1}$$

Here the dimension $s$ may be large in practical applications. The restriction to integration problems over the unit cube $[0,1]^s$ is mostly for simplicity and in many cases does not impose a big limitation, since most integrals over bounded or unbounded regions can be transformed into integrals over the unit cube (although one has to be careful in choosing suitable transformations which, of course, have influence on the behavior of the transformed integrand).

For most functions arising in practice the integral (1.1) cannot be solved analytically since, for instance, the given integrand does not have an antiderivative or its antiderivative, if existent, is not expressible as a finite combination of elementary functions. Hence we must solve such integrals numerically, that is, we want to find an algorithm which enables us to approximate the true value of the integral to any prescribed level of accuracy.

We will always assume that for a function $f$ we can compute its function values at finitely many points in $[0,1]^s$. However, we remark here already that the knowledge of finitely many function values is not enough to solve the integration problem, since in this case the integral still could be any number. So we will need further "global" information, as for example the smoothness of $f$, which restricts the class of integrands to certain function classes.

In the one-dimensional case there are many classical quadrature rules available, such as, for instance, the rectangle rule (midpoint rule), the trapezoidal rule, Simpson's rule, or the Gauss rule, which all have the general form

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_1,

$$T_m(f) = \sum_{n=0}^{m} q_n f(x_n) \tag{1.2}$$

with quadrature points $x_0, \ldots, x_m$ from $[0, 1]$ and with weights $q_0, \ldots, q_m \in \mathbb{R}$. As an example, we mention the trapezoidal rule for which $q_0 = q_m = 1/(2m)$, $q_n = 1/m$ for $n = 1, \ldots, m-1$ and $x_n = n/m$ for $n = 0, \ldots, m$. If $f \in C^2([0, 1])$, the error of the trapezoidal rule is of order $O(m^{-2})$ (cf. Exercise 1.2).

In the multi-dimensional case the classical methods use Cartesian products of one-dimensional quadrature rules which are then often called *product rules*. This means that one applies a one-dimensional quadrature rule of the form (1.2) to each one-dimensional integral in (1.1), which results in the product rule of the form

$$\sum_{n_1=0}^{m} \cdots \sum_{n_s=0}^{m} q_{n_1} \cdots q_{n_s} f(x_{n_1}, \ldots, x_{n_s}).$$

When recasting this product rule in the form

$$\sum_{n=0}^{M} w_n f(\boldsymbol{x}_n),$$

then the set of quadrature points $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_M\}$ is just the $s$-fold product of the one-dimensional quadrature points $\{x_0, \ldots, x_m\}$. Hence the total number of nodes of a product rule is $N = M + 1 = (m + 1)^s$, which grows dramatically with the dimension $s$. For example, if $s = 30$, a product rule based on a one-dimensional rule with only two points involves already $N = 2^{30} \approx 10^9$ nodes. But if $s = 500$, then a product rule based on a 2-element one-dimensional rule already requires $2^{500}$ nodes and the time until the end of our universe will not suffice to compute that many function evaluations.

The error analysis of product rules follows immediately from that for the underlying one-dimensional quadrature rule. For instance, the error of approximating an $s$-dimensional integral by a product rule based on the trapezoidal rule is of order $O(m^{-2})$, where $m + 1$ is the number of nodes used in the underlying one-dimensional rule, provided that all partial derivatives of order two in each variable are continuous on $[0, 1]^s$. This looks quite promising on first sight, but in terms of the actual number $N = M + 1$ of integration nodes, this error is of order $O(N^{-2/s})$. For large dimensions, which might be in the hundreds for practical problems, such an error convergence is less than satisfying. This phenomenon is often called the *curse of dimensionality*. The expression "curse of dimensionality" was coined already in 1957 by R. Bellman and means in our context that the minimal number of function values needed to compute an $\varepsilon$-approximation to multivariate integrals is an exponential function in $s$.

### Example 1.1

Let $\mathcal{F}_{\mathrm{Lip}} = \{f : [0,1]^s \to \mathbb{R} : \mathrm{Lip}(f) \leq L\}$ denote the class of Lipschitz continuous functions with Lipschitz constant of at most $L$, i.e., for $f \in \mathcal{F}_{\mathrm{Lip}}$ we have $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_\infty$, where $\|\boldsymbol{x}\|_\infty = \max_{j=1,\dots,s} |x_j|$ for $\boldsymbol{x} = (x_1, \dots, x_s)$. Assume we want to use the Cartesian product of the one-dimensional midpoint rule for the integration of functions $f \in \mathcal{F}_{\mathrm{Lip}}$. Then the set of integration nodes is

$$\Gamma_{m,s} = \left\{ \boldsymbol{x}_{\boldsymbol{k}} = \left( \frac{2k_1+1}{2m}, \dots, \frac{2k_s+1}{2m} \right) : \boldsymbol{k} = (k_1, \dots, k_s) \in \{0, 1, \dots, m-1\}^s \right\}.$$

Denote the integration error of the product midpoint rule for $f \in \mathcal{F}_{\mathrm{Lip}}$ by

$$e(f, \Gamma_{m,s}) = \int_{[0,1]^s} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \frac{1}{m^s} \sum_{k_1,\dots,k_s=0}^{m-1} f(\boldsymbol{x}_{\boldsymbol{k}})$$

and put $Q_{\boldsymbol{k}} = \prod_{j=1}^s \left[\frac{k_j}{m}, \frac{k_j+1}{m}\right)$ for $\boldsymbol{k} = (k_1, \dots, k_s) \in \{0, 1, \dots, m-1\}^s$. Then

$$|e(f, \Gamma_{m,s})| = \left| \sum_{k_1,\dots,k_s=0}^{m-1} \int_{Q_{\boldsymbol{k}}} (f(\boldsymbol{x}) - f(\boldsymbol{x}_{\boldsymbol{k}})) \,\mathrm{d}\boldsymbol{x} \right| \leq L \sum_{\substack{\boldsymbol{k} \in \mathbb{N}_0^s \\ \|\boldsymbol{k}\|_\infty < m}} \int_{Q_{\boldsymbol{k}}} \|\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{k}}\|_\infty \,\mathrm{d}\boldsymbol{x}.$$

For $\boldsymbol{x} \in Q_{\boldsymbol{k}}$ we have

$$\|\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{k}}\|_\infty = \max_{j=1,\dots,s} \left| x_j - \frac{2k_j+1}{2m} \right| \leq \frac{1}{2m},$$

and hence we find that

$$|e(f, \Gamma_{m,s})| \leq \frac{L}{2m} = \frac{L}{2} \frac{1}{N^{1/s}},$$

where $N = \#\Gamma_{m,s} = m^s$ is the number of employed integration nodes.

This result cannot be significantly improved. For example, consider the function

$$g(x_1, x_2, \dots, x_s) = \frac{L}{2\pi m}(1 + \cos(2\pi m x_1)).$$

It is easy to see that $g \in \mathcal{F}_{\mathrm{Lip}}$ and that

$$\int_{[0,1]^s} g(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \frac{L}{2\pi m} = \frac{L}{2\pi} \frac{1}{N^{1/s}} \quad \text{and} \quad \frac{1}{m^s} \sum_{k_1,\dots,k_s=0}^{m-1} g(\boldsymbol{x}_{\boldsymbol{k}}) = 0.$$

Hence $e(g, \Gamma_{m,s}) = \frac{L}{2\pi} \frac{1}{N^{1/s}}$. This means that

$$\frac{L}{2\pi} \frac{1}{N^{1/s}} \leq \sup_{f \in \mathcal{F}_{\mathrm{Lip}}} |e(f, \Gamma_{m,s})| \leq \frac{L}{2} \frac{1}{N^{1/s}}.$$

The question arises whether there are algorithms for multivariate integration for which the error convergence does not show such a poor dependence on the dimension $s$. This question can be answered in the affirmative, which can be seen by the following considerations.

## 1.2 Monte Carlo Integration

We aim at approximating the integral of a function $f : [0,1]^s \to \mathbb{R}$ by an equal weight quadrature rule of the form

$$\frac{1}{N}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n),$$

where the quadrature points $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ are from $[0,1)^s$. We are interested in the integration error

$$e(f,\mathcal{P}) := \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \frac{1}{N}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n).$$

But how should we choose the quadrature points? One idea is to choose realisations of $N$ independent and uniformly distributed random variables $X_0, \ldots, X_{N-1}$ in $[0,1]^s$ and to check what we can expect for the resulting error. This means that we use

$$Q_{N,s}(f) := \frac{1}{N}\sum_{n=0}^{N-1} f(X_n) \tag{1.3}$$

as an estimator for the integral. Note that a measurable function $f : [0,1]^s \to \mathbb{R}$ can be considered as a random variable on the probability space $([0,1]^s, \mathcal{B}, \lambda_s)$ where $\mathcal{B}$ is the Borel $\sigma$-algebra on $[0,1]^s$ and $\lambda_s$ the Lebesgue measure. Then the expectation of this random variable equals the integral we want to compute, i.e., $\mathbb{E}[f] = \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Using the linearity of the expected value we have

$$\mathbb{E}[Q_{N,s}(f)] = \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}[f] = \mathbb{E}[f] = \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

and hence $Q_{N,s}(f)$ is an unbiased estimator for the integral $\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The strong law of large numbers guarantees that

$$\mathbb{P}\left[\lim_{N\to\infty} Q_{N,s}(f) = \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right] = 1,$$

where $\mathbb{P}[\cdot]$ is the probability on an arbitrary probability space supporting an independent sequence $(X_n)_{n\in\mathbb{N}_0}$ of random variables uniformly distributed on $[0,1]^s$.

The *variance* of $f$ is given by $\mathrm{Var}[f] := \int_{[0,1]^s}(f(\boldsymbol{x}) - \int_{[0,1]^s} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y})^2\,\mathrm{d}\boldsymbol{x}$. Since $X_0, \ldots, X_{N-1}$ are independent, we obtain from the Bienaymé formula the following result for the variance of the estimator $Q_{N,s}(f)$.

**Theorem 1.2.** *Let $f \in L_2([0,1]^s)$. Then for any $N \in \mathbb{N}$ we have*

$$\mathrm{Var}[Q_{N,s}(f)] = \frac{\mathrm{Var}[f]}{N}.$$

Note that

$$\mathrm{Var}[Q_{N,s}(f)] = \mathbb{E}[(Q_{N,s}(f) - \mathbb{E}[f])^2] = \mathbb{E}[e^2(f,\cdot)],$$

where $e(f,\cdot)$ is the error estimator

$$e(f,\cdot) := \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \frac{1}{N}\sum_{n=0}^{N-1} f(X_n).$$

Hence it follows from Theorem 1.2 that

$$\mathbb{E}[|e(f,\cdot)|] \le \sqrt{\mathbb{E}[e^2(f,\cdot)]} = \frac{\sigma[f]}{\sqrt{N}},$$

where $\sigma[f] = (\mathrm{Var}[f])^{1/2}$ denotes the *standard deviation* of $f$. This means that the absolute value of the integration error is, on average, bounded by $\sigma[f]/\sqrt{N}$. It is remarkable that the convergence rate of the expected integration error does not depend on the dimension $s$. We have $N^{-1/2} < N^{-1/s}$ for $s > 2$. Hence, roughly speaking, for $s > 2$ it is on average better to use random points for the approximation of the integral of $f$ than classical product rules. Here $f$ does not even have to be continuous if one chooses random samples.

The method of using random sample points is called (*plain*) *Monte Carlo* (MC) *method* and the integration rule (1.3) is called a *Monte Carlo rule* (*or algorithm*). The phrase “Monte Carlo” goes back to J. von Neumann, who chose it as a code name for this method when he worked on a secret project together with S. Ulam in 1946 at Los Alamos Scientific Laboratory.

To summarize, the advantages of the MC method are:

- It suffices that the integrands are quadratic integrable.
- The convergence rate of order $O(N^{-1/2})$ is independent of the dimension $s$. This is a surprising fact, although it does not mean that MC breaks the curse of dimensionality, because the standard deviation $\sigma[f]$ is in general not independent of $s$.

On the other hand, the MC method also has some disadvantages, which are:

- The error bound is only “probabilistic”, that is, in any one instance one cannot be sure of the integration error. However, further probabilistic information is obtained from the central limit theorem, which states that, if $0 < \sigma[f] < \infty$, then

$$\lim_{N\to\infty} \mathbb{P}\left[|e(f,\{X_0,\ldots,X_{N-1}\})| \le \frac{c\,\sigma[f]}{\sqrt{N}}\right] = 2\Phi(c) - 1,$$

for any $c > 0$, where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)\,\mathrm{d}t$ is the cumulative distribution function of the standard normal distribution.

If we would like to have

$$\mathbb{P}\left[|e(f, \{X_0, \ldots, X_{N-1}\})| > \varepsilon\right] < \alpha$$

for small $\varepsilon, \alpha > 0$, then we need to require

$$\begin{aligned} 1-\alpha &< \mathbb{P}\left[|e(f, \{X_0, \ldots, X_{N-1}\})| \leq \varepsilon\right] \\ &= \mathbb{P}\left[-\frac{\varepsilon\sqrt{N}}{\sigma[f]} \leq \sqrt{N}\frac{e(f, \{X_0, \ldots, X_{N-1}\})}{\sigma[f]} \leq \frac{\varepsilon\sqrt{N}}{\sigma[f]}\right] \\ &\approx 2\Phi\left(\frac{\varepsilon\sqrt{N}}{\sigma[f]}\right) - 1. \end{aligned}$$

This means that $1 - \frac{\alpha}{2} < \Phi\left(\frac{\varepsilon\sqrt{N}}{\sigma[f]}\right)$, or

$$N > \frac{\mathrm{Var}[f]}{\varepsilon^2}\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)^2.$$

Note that $\mathrm{Var}[f]$ is in general not explicitly known. However, it can be estimated per MC by using the same function evaluations as for the MC integration of $f$. This will be explained in more detail shortly.

- A second problem is that the generation of random samples is difficult. This problem is a topic on its own which cannot be further discussed in this book. References for reading in this direction are provided at the end of this chapter.
- For some applications the convergence rate of $O(N^{-1/2})$ is too slow.
- The convergence rate of $O(N^{-1/2})$ does not reflect some possible regularity of the integrand.

## Example 1.3

A popular first example for the MC method is the approximation of the number $\pi$. The idea is to distribute a certain number of points independently and uniformly in the unit-square and to count the number of them which lie distance at most 1 from the origin; see Fig. 1.1. This number, normalized by the total number of distributed points, is an approximation for the number $\pi/4$, which is the area of a quadrant with radius 1.

Let $f : [0,1]^2 \rightarrow \mathbb{R}$, $f(x_1, x_2) = 1$ if $x_1^2 + x_2^2 \leq 1$ and 0 otherwise. Obviously, $\int_0^1 \int_0^1 f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \frac{\pi}{4}$. Then the method described above for the approximation of $\pi/4$ amounts to calculating the MC mean $\frac{1}{N}\sum_{n=0}^{N-1} f(X_n)$ with $N$ independent and uniformly distributed random variables $X_0, \ldots, X_{N-1}$ in $[0,1]^2$. Performing for fixed $N$ 100 random experiments and plotting the mean MC error for this random experiments leads to a picture as shown in Fig. 1.2.

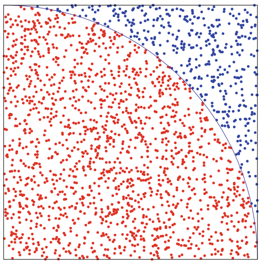

Fig. 1.1 2000 random points in $[0, 1]^2$

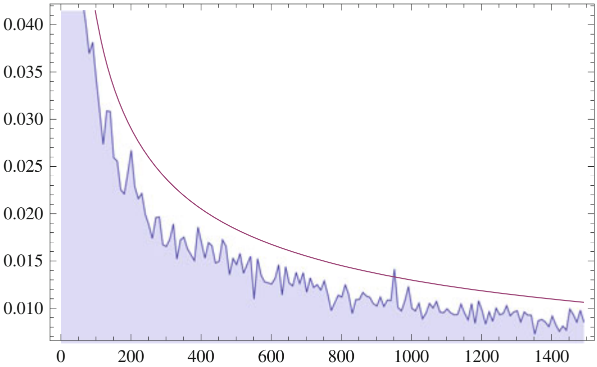

Fig. 1.2 Average MC error (*blue*) for $N = 1, \ldots, 1500$ and 100 random experiments per given $N$ compared to $\sigma[f]/\sqrt{N}$ (*red*)

A practical error estimate for the MC method can be obtained from the estimator

$$V_N = \frac{1}{N-1} \sum_{n=0}^{N-1} (f(X_n) - Q_{N,s}(f))^2$$

for the variance $\mathrm{Var}[f]$ whenever $f \in L_2([0,1]^s)$. This estimator is unbiased, since

$$V_N = \frac{1}{N-1} \sum_{n=0}^{N-1} (f(X_n) - \mathbb{E}[f])^2 - \frac{N}{N-1} (\mathbb{E}[f] - Q_{N,s}(f))^2,$$

and hence, by using Theorem 1.2

$$\begin{aligned}\mathbb{E}[V_N] &= \frac{1}{N-1}\sum_{n=0}^{N-1}\mathbb{E}[(f(X_n)-\mathbb{E}[f])^2] - \frac{N}{N-1}\mathbb{E}[(\mathbb{E}[f]-Q_{N,s}(f))^2]\\ &= \frac{1}{N-1}\sum_{n=0}^{N-1}\mathrm{Var}[f] - \frac{N}{N-1}\frac{\mathrm{Var}[f]}{N}\\ &= \mathrm{Var}[f].\end{aligned}$$

Hence,

$$\frac{V_N}{N} = \frac{1}{N(N-1)}\left(\sum_{n=0}^{N-1} f^2(X_n) - N(Q_{N,s}(f))^2\right)$$

provides an unbiased estimator for $\mathrm{Var}[Q_{N,s}(f)] = \mathbb{E}[e^2(f,\cdot)]$. Note that this estimator requires only $O(N)$ function evaluations and hence it is very efficient for practical implementation.

As already mentioned, one of the disadvantages of MC is the slow convergence rate of order $O(N^{-1/2})$. Although in practice one can use variance reduction techniques such as, e.g., importance sampling, stratified sampling, or correlated sampling, these techniques usually do not improve the rate of error convergence. One very simple variance reduction technique for dimension $s = 1$ is discussed in Exercise 1.10. This is the main motivation for the following strategy:

The aim is to find *deterministic* constructions of quadrature points which are at least as good as the average. This method is then called *quasi-Monte Carlo* (QMC) *method*, as opposed to MC, where one uses randomly chosen quadrature points, and

$$Q_{N,s}(f) = \frac{1}{N}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n)$$

with *deterministic* quadrature points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$ in $[0,1)^s$ is called a *quasi-Monte Carlo* (QMC) *rule (or algorithm)*. Hence in the deterministic case we need quadrature points which are in some sense “well” distributed in $[0,1)^s$. To this end we need to clarify what we mean by “well” distributed. This question is subject to the theory of Uniform Distribution Modulo One, which is the topic of our next chapter.

## 1.3 Further Reading and Exercises

### Further Reading

A coherent theory underlying numerical quadrature is provided in the book of Brass and Petras [10]. An excellent introduction to the Monte Carlo method is provided by the book of Müller-Gronbach, Novak and Ritter [67]. We also recommend the books of Niederreiter [69], Lemieux [61], and Glasserman [31], where the latter deals with the application of the Monte Carlo method to financial problems. Information concerning the generation of (pseudo) random samples can be found in the books of Knuth [54, Chapter 3], Niederreiter [69, Chapter 7–10], Tezuka [91], and in the survey articles of L'Ecuyer and Hellekalek [59].

## Exercises

1.1 Let $f \in C^1([0,1])$. For $m \in \mathbb{N}$ the trapezoidal rule is given by

$$T_m f := \sum_{n=0}^{m-1} \frac{f(n/m) + f((n+1)/m)}{2m} = \sum_{n=0}^{m} q_n f\left(\frac{n}{m}\right),$$

where $q_0 = q_m = 1/(2m)$ and $q_n = 1/m$ for $n = 1, \ldots, m-1$. Show that

$$\int_0^1 f(x)\,\mathrm{d}x - T_m f = \int_0^1 K(t) f'(t)\,\mathrm{d}t, \tag{1.4}$$

where $K(t) = \frac{n}{m} + \frac{1}{2m} - t$ for $\frac{n}{m} < t \leq \frac{n+1}{m}$ and $n = 0, \ldots, m-1$. Deduce from (1.4) that the integration error is of order $O(m^{-1})$. *Hint:* Consider first the term

$$\int_{n/m}^{(n+1)/m} f(x)\,\mathrm{d}x - \frac{f(n/m) + f((n+1)/m)}{2m}$$

and use that $f(x) = f(n/m) + \int_{n/m}^{x} f'(t)\,\mathrm{d}t$ for $n/m \leq x \leq (n+1)/m$.

1.2 Assume that $f \in C^2([0,1])$. Show that then the integration error of the trapezoidal rule is of order $O(m^{-2})$. *Hint:* Use (1.4) and integration by parts on $[\frac{n}{m}, \frac{n+1}{m}]$.

1.3 Let $f : [0,1]^s \to \mathbb{R}$ have continuous second order partial derivatives $\partial^2 f/\partial x_i^2$ for all $i = 1, \ldots, s$. Consider the product trapezoidal rule of the form

$$T_m^{(s)} f = \sum_{n_1=0}^{m} \cdots \sum_{n_s=0}^{m} w_{n_1} \cdots w_{n_s} f\left(\frac{n_1}{m}, \ldots, \frac{n_s}{m}\right).$$

Show that if $f$ depends only on one variable and is constant with respect to the remaining $s-1$ variables, then the integration error is of order $O(m^{-2})$. Is this result satisfactory if we consider the number of required integration nodes?

1.4 Let $f : [0,1]^2 \to \mathbb{R}$, $f(x_1, x_2) = 1$ if $x_1^2 + x_2^2 \leq 1$ and $f(x_1, x_2) = 0$ otherwise, be the function from Example 1.3. Write a computer program for MC integration and run several experiments. Compare the MC error with $1/\sqrt{N}$, where $N$ denotes the number of nodes involved.

1.5 Let $f : [0,1]^3 \to \mathbb{R}$, $f(x, y, z) = x^2 y - \exp(2y + z)$. Compute

$$\int_{[0,1]^3} f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

as well as $\mathrm{Var}[f]$ and $\mathbb{E}[e^2(f, \mathcal{P})]$.

1.6 Compute the integral from Exercise 1.5 with the MC method. Run for fixed $N$ 100 experiments and compare the average MC error with $\mathbb{E}[e^2(f, \mathcal{P})]$.

1.7 Let $f, g : [0, 1] \to \mathbb{R}$, $f(x) = x^{-1/4}$ and $g(x) = x^{-3/4}$. Use the MC method to estimate the integrals $\int_0^1 f(x)\,\mathrm{d}x$ and $\int_0^1 g(x)\,\mathrm{d}x$. Repeat the random experiment for fixed $N$ 100 times and plot the average error for $N \in \{1, \dots, 5000\}$ against $N$ (for example in steps of 50). Why is there such a big difference between the results for $f$ and $g$? *Remark:* This example is taken from [67], where one can find a very detailed discussion of it.

1.8 Let $f \in L_2([0, 1]^s)$ and let $X_1, \dots X_N$ be independent and uniformly distributed random variables in $[0, 1]^s$.

(a) Show that

$$\mathbb{P}\left[\left|Q_{N,s}(f) - \mathbb{E}[f]\right| \geq \varepsilon\right] \leq \frac{\operatorname{Var}[f]}{N\varepsilon^2}.$$

(b) Let $\delta > 0$ and $N \in \mathbb{N}$. Determine confidence intervals of the form $I_{N,\delta} = \left[Q_{N,s}(f) - L_N, Q_{N,s}(f) + L_N\right]$ such that

$$\mathbb{P}\left[\mathbb{E}[f] \in I_{N,\delta}\right] \geq 1 - \delta.$$

1.9 Show that the estimator $\frac{1}{N}\sum_{n=1}^{N}\left(f(X_n) - Q_{N,s}\right)^2$ for $\operatorname{Var}[f]$ is *not* unbiased.

1.10 Let $f : [0, 1] \to \mathbb{R}$ be continuous and monotone. Put $g(x) = (f(x) + f(1 - x))/2$. Show that $\int_0^1 f(x)\,\mathrm{d}x = \int_0^1 g(x)\,\mathrm{d}x$ and

$$\operatorname{Var}[g] \leq \frac{1}{2}\operatorname{Var}[f].$$

*Hint:* A proof of this simple variance reduction technique can be found in [69, Proposition 1.3].

# Uniform Distribution Modulo One 2

The theory of *Uniform Distribution Modulo One* is a branch of Number Theory which goes back to the seminal work of H. Weyl from 1916. For us the main motivation to study this topic lies in its application for numerical integration based on QMC rules.

## 2.1 Definition and Basic Properties

On first sight, it is not immediately clear what it should mean that a finite point set $\mathcal{P}$ is uniformly distributed in the unit cube. Intuitively one may suppose that every region contains a proper portion of elements from $\mathcal{P}$, that there are no accumulations of points, and that there are no big gaps between the points from $\mathcal{P}$. In other words, every region should contain a number of elements from $\mathcal{P}$ that is proportional to the size of the region. To put this intuition into exact mathematical terms we consider infinite sequences instead of finite point sets. Since a sequence is a discrete object, it is clear that we need some restriction on the demand "*every* region". In Uniform Distribution Theory these regions are usually restricted to intervals of the form $[\boldsymbol{a}, \boldsymbol{b})$, where $\boldsymbol{a} = (a_1, \ldots, a_s)$, $\boldsymbol{b} = (b_1, \ldots, b_s)$ are elements of $[0,1]^s$ and $[\boldsymbol{a}, \boldsymbol{b}) = [a_1, b_1) \times \cdots \times [a_s, b_s)$. See Fig. 2.1 for a memorable picture.

For a sequence $\mathcal{S} = (\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$ in $[0,1)^s$ and an interval $[\boldsymbol{a}, \boldsymbol{b}) \subseteq [0,1)^s$, the number of indices $n \in \{0, \ldots, N-1\}$ for which $\boldsymbol{x}_n \in [\boldsymbol{a}, \boldsymbol{b})$ is denoted by $A([\boldsymbol{a}, \boldsymbol{b}), \mathcal{S}, N)$ (or in short by $A([\boldsymbol{a}, \boldsymbol{b}), N)$, if there is no risk of confusion). Precisely,

$$A([\boldsymbol{a}, \boldsymbol{b}), \mathcal{S}, N) = \#\{n \in \mathbb{N}_0 \,:\, 0 \le n \le N-1 \text{ and } \boldsymbol{x}_n \in [\boldsymbol{a}, \boldsymbol{b})\}.$$

We use this notion also for finite point sets $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$, which we always interpret in the sense of the combinatorial notion of "multiset", i.e., a set where the multiplicity of elements matters. Precisely,

$$A([\boldsymbol{a}, \boldsymbol{b}), \mathcal{P}, N) = \#\{n \in \mathbb{N}_0 \,:\, 0 \le n \le N-1 \text{ and } \boldsymbol{x}_n \in [\boldsymbol{a}, \boldsymbol{b})\}.$$

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_2,

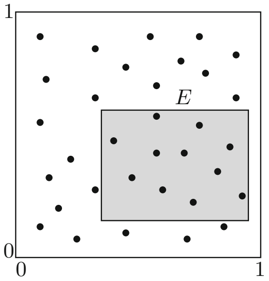

**Fig. 2.1** Every interval should contain the proper portion of points

Let $\lambda_s$ denote the $s$-dimensional Lebesgue measure. For intervals of the form $[\boldsymbol{a}, \boldsymbol{b})$ this reduces to $\lambda_s([\boldsymbol{a}, \boldsymbol{b})) = \prod_{j=1}^{s}(b_j - a_j)$.

**Definition 2.1**

An infinite sequence $\mathcal{S}$ in $[0,1)^s$ is said to be *uniformly distributed modulo one* (or equidistributed), if for every interval of the form $[\boldsymbol{a}, \boldsymbol{b}) \subseteq [0,1)^s$ we have

$$\lim_{N\to\infty} \frac{A([\boldsymbol{a}, \boldsymbol{b}), \mathcal{S}, N)}{N} = \lambda_s([\boldsymbol{a}, \boldsymbol{b})). \tag{2.1}$$

We will encounter sequences which enjoy this property later on.

Let $\chi_J$ be the characteristic function of a set $J \subseteq \mathbb{R}^s$, i.e., $\chi_J(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in J$, and 0 otherwise. Then we can write

$$A([\boldsymbol{a}, \boldsymbol{b}), \mathcal{S}, N) = \sum_{n=0}^{N-1} \chi_{[\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{x}_n),$$

and hence (2.1) is equivalent to

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \chi_{[\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{x}_n) = \int_{[0,1]^s} \chi_{[\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{2.2}$$

Relation (2.2) leads to the following characterization of uniform distribution modulo one:

**Theorem 2.2.** *A sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ in $[0,1)^s$ is uniformly distributed modulo one if and only if for every Riemann integrable function $f : [0,1]^s \to \mathbb{R}$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n) = \int_{[0,1]^s} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{2.3}$$

Observe that Theorem 2.2 gives the link to QMC integration. It is now clear that in order to obtain a QMC rule which converges to the actual value of the integral, the underlying nodes must come from a uniformly distributed sequence.

► **Remark 2.3** There is no sequence which satisfies (2.3) for all Lebesgue integrable functions. Moreover, it is known that for every Lebesgue integrable function which is not integrable in the Riemann sense there exists a uniformly distributed sequence for which (2.3) does not hold true.

*Proof of Theorem 2.2.* Assume that (2.3) holds for any Riemann integrable function $f : [0,1]^s \to \mathbb{R}$. But then (2.2) holds true as well, since for any sub-interval $J$ of $[0,1]^s$ the characteristic function $\chi_J$ is Riemann integrable. Thus the sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ is uniformly distributed modulo one.

In the opposite direction, assume that the sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ is uniformly distributed modulo one. We will show in a first step that then (2.3) holds for every step function on $[0,1)^s$ (i.e., every finite linear combination of characteristic functions of intervals of the form $[\boldsymbol{a},\boldsymbol{b})$). Let the intervals $E_1,\dots,E_m$ form a partition of $[0,1]^s$ and put

$$g(\boldsymbol{x}) = \sum_{i=1}^{m} d_i \chi_{E_i}(\boldsymbol{x}),$$

where $d_i \in \mathbb{R}$ for $i = 1,2,\dots,m$. Then we have

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} g(\boldsymbol{x}_n) &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=1}^{m} d_i \chi_{E_i}(\boldsymbol{x}_n)\\ &= \sum_{i=1}^{m} d_i \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\chi_{E_i}(\boldsymbol{x}_n) = \sum_{i=1}^{m} d_i\lambda(E_i) = \int_{[0,1]^s} g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Hence (2.3) holds for all step functions on $[0,1]^s$.

Now we consider a Riemann integrable function $f : [0,1]^s \to \mathbb{R}$. According to the definition of the Riemann integral, for every $\varepsilon > 0$ there exist step functions $g_1$ and $g_2$ defined on $[0,1]^s$ such that $g_1 \le f \le g_2$ and $\int_{[0,1]^s}(g_2(\boldsymbol{x}) - g_1(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} < \varepsilon$. Hence we have

$$\begin{aligned}\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \varepsilon \le \int_{[0,1]^s} g_1(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} g_1(\boldsymbol{x}_n)\\ &\le \liminf_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n) \le \limsup_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n)\end{aligned}$$

$$\leq \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} g_2(\boldsymbol{x}_n) = \int_{[0,1]^s} g_2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \leq \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \varepsilon.$$

Since $\varepsilon$ can be chosen arbitrarily close to zero, the result follows. □

In the same manner one can prove the following result:

**Theorem 2.4.** *A sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ in $[0,1)^s$ is uniformly distributed modulo one if and only if (2.3) holds for every continuous one-periodic and complex-valued function $f : [0,1]^s \to \mathbb{C}$.*

*Proof.* The necessity follows from applying Theorem 2.2 to the real and imaginary parts of $f$. The sufficiency of (2.3) for every continuous one-periodic and complex-valued function $f : [0,1]^s \to \mathbb{C}$ follows from the elementary fact that the characteristic function of every interval $[\boldsymbol{a},\boldsymbol{b}) \subseteq [0,1)^s$ can be approximated from above and below by continuous one-periodic functions, i.e., for all $\varepsilon > 0$ there exist continuous one-periodic $f_1, f_2$ on $[0,1]^s$ such that $f_1 \leq \chi_{[\boldsymbol{a},\boldsymbol{b})} \leq f_2$ and $0 \leq \int_{[0,1]^s}(f_2(\boldsymbol{x}) - f_1(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} < \varepsilon$. We leave the details as an exercise. □

In the already mentioned paper entitled *Über die Gleichverteilung von Zahlen mod. Eins* H. Weyl proved the following criterion for uniform distribution modulo one, which can be considered as a simplification of Theorem 2.4 in the sense that it is sufficient to check condition (2.3) for trigonometric functions of the form $\exp(2\pi\mathtt{i}\boldsymbol{h}\cdot\boldsymbol{x})$ for integer vectors $\boldsymbol{h} \neq \boldsymbol{0}$, rather than for all continuous one-periodic and complex-valued functions.

**Theorem 2.5 (Weyl's criterion).** *A sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ in $[0,1)^s$ is uniformly distributed modulo one if and only if*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \exp(2\pi\mathtt{i}\boldsymbol{h}\cdot\boldsymbol{x}_n) = 0 \tag{2.4}$$

*for all vectors $\boldsymbol{h} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\}$. Here $\boldsymbol{x}\cdot\boldsymbol{y}$ denotes the usual inner product of two elements $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^s$.*

*Proof.* If the sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ is uniformly distributed modulo one, then Theorem 2.4 with the special choice $f(\boldsymbol{x}) = \exp(2\pi\mathtt{i}\boldsymbol{h}\cdot\boldsymbol{x})$ implies (2.4).

Now let $f : [0,1]^s \to \mathbb{C}$ be a continuous one-periodic and complex-valued function. According to the trigonometric version of Weierstrass' approximation theorem, $f$ can be uniformly approximated as closely as desired by a finite linear combination of functions $\exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{x})$, $\boldsymbol{h} \in \mathbb{Z}^s$, with complex coefficients. In other words, for any $\varepsilon > 0$ there exists a trigonometric polynomial of the form

$$P(\boldsymbol{x}) = \sum_{\substack{\boldsymbol{h} \in \mathbb{Z}^s \\ \|\boldsymbol{h}\|_1 < R}} a_{\boldsymbol{h}} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{x})$$

with complex coefficients $a_{\boldsymbol{h}}$, such that

$$\|f - P\|_{L_\infty} < \varepsilon/3,$$

where $R = R(\varepsilon)$. Here for $\boldsymbol{h} = (h_1, \ldots, h_s)$ we put $\|\boldsymbol{h}\|_1 = |h_1| + \cdots + |h_s|$, and for a function $g : [0,1]^s \to \mathbb{C}$ we put $\|g\|_{L_\infty} := \sup_{\boldsymbol{x} \in [0,1]^s} |g(\boldsymbol{x})|$. Then we have

$$\begin{aligned} \left| \int_{[0,1]^s} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n) \right| \le & \left| \int_{[0,1]^s} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \int_{[0,1]^s} P(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \\ & + \left| \int_{[0,1]^s} P(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{n=0}^{N-1} P(\boldsymbol{x}_n) \right| \\ & + \left| \frac{1}{N} \sum_{n=0}^{N-1} P(\boldsymbol{x}_n) - \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n) \right|. \end{aligned}$$

We have

$$\left| \int_{[0,1]^s} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \int_{[0,1]^s} P(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \le \|f - P\|_{L_\infty} < \varepsilon/3$$

and

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} P(\boldsymbol{x}_n) - \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n) \right| \le \|f - P\|_{L_\infty} < \varepsilon/3.$$

Furthermore, it follows from (2.4), that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} P(\boldsymbol{x}_n) = \sum_{\substack{\boldsymbol{h} \in \mathbb{Z}^s \\ \|\boldsymbol{h}\|_1 < R}} a_{\boldsymbol{h}} \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{x}_n) = a_{\boldsymbol{0}}$$

and

$$\int_{[0,1]^s} P(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \sum_{\substack{\boldsymbol{h} \in \mathbb{Z}^s \\ \|\boldsymbol{h}\|_1 < R}} a_{\boldsymbol{h}} \int_{[0,1]^s} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = a_{\boldsymbol{0}}.$$

Hence, for large enough $N$,

$$\left| \int_{[0,1]^s} P(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{n=0}^{N-1} P(\boldsymbol{x}_n) \right| < \varepsilon/3.$$

Altogether, for large enough $N$, we obtain

$$\left| \int_{[0,1]^s} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n) \right| < \varepsilon$$

and the result follows. □

Based on Weyl's criterion we can give a first example of a uniformly distributed sequence.

**Proposition 2.6.** *Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s$. The sequence $(\{n\boldsymbol{\alpha}\})_{n \in \mathbb{N}_0}$, where the fractional part $\{\cdot\}$ is applied component-wise, is uniformly distributed modulo one if and only if the numbers $1, \alpha_1, \ldots, \alpha_s$ are linearly independent over the rationals. In particular, the one-dimensional sequence $(\{n\alpha\})_{n \in \mathbb{N}_0}$ is uniformly distributed modulo one if and only if $\alpha \in \mathbb{R} \setminus \mathbb{Q}$.*

*Proof.* Let $\boldsymbol{h} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\}$. Since the function $x \mapsto \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{x})$ is one-periodic, we have

$$\frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \{n\boldsymbol{\alpha}\}) = \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{\alpha})^n.$$

If the numbers $1, \alpha_1, \ldots, \alpha_s$ are linearly independent over the rationals, then $\boldsymbol{h} \cdot \boldsymbol{\alpha} \notin \mathbb{Q}$ and therefore $\exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{\alpha}) \neq 1$. Using the formula for geometric sums we obtain

$$\begin{aligned} \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{\alpha})^n \right| &= \frac{1}{N} \left| \frac{\exp(2\pi \mathtt{i} N \boldsymbol{h} \cdot \boldsymbol{\alpha}) - 1}{\exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{\alpha}) - 1} \right| \\ &\leq \frac{1}{N} \frac{2}{|\exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{\alpha}) - 1|}. \end{aligned}$$

An application of Weyl's criterion shows that the sequence is uniformly distributed modulo one. The other direction is left as an exercise. □

▶ **Remark 2.7** The sequences considered in Proposition 2.6 are sometimes called *Kronecker sequences* or *$n\alpha$-sequences*.

We provide another example of a uniformly distributed sequence. To this end we need the following definition:

**Definition 2.8**

Let $b \in \mathbb{N}, b \geq 2$.

- The *$b$-adic radical inverse function* is defined as $\phi_b : \mathbb{N}_0 \to [0, 1)$,

$$\phi_b(n) = \frac{n_0}{b} + \frac{n_1}{b^2} + \frac{n_2}{b^3} \cdots ,$$

  for $n \in \mathbb{N}_0$ with $b$-adic digit expansion $n = n_0 + n_1 b + n_2 b^2 + \cdots$, where $n_i \in \{0, 1, \ldots, b-1\}$. In other words, $\phi_b(n)$ is the reflection of the $b$-adic digit expansion of $n$ at the comma.
- The *van der Corput sequence in base $b$* is defined as $(x_n)_{n\in\mathbb{N}_0}$ with $x_n = \phi_b(n)$.

## Example 2.9

For example, for $b = 2$ we have

| $n$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $(n)_2$ | 0. | 1. | 10. | 11. | 100. | 101. | 110. | 111. | 1000. |
| $(\phi_2(n))_2$ | .0 | .1 | .01 | .11 | .001 | .101 | .011 | .111 | .0001 |
| $\phi_2(n)$ | 0 | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{3}{4}$ | $\frac{1}{8}$ | $\frac{5}{8}$ | $\frac{3}{8}$ | $\frac{7}{8}$ | $\frac{1}{16}$ |

**Proposition 2.10.** *The van der Corput sequence in base $b$ is uniformly distributed modulo one.*

*Proof.* Fix $m \in \mathbb{N}$. For every $a \in \{0, 1, \ldots, b^m - 1\}$ with $b$-adic digit expansion $a = a_0 b^{m-1} + a_1 b^{m-2} + \cdots + a_{m-2} b + a_{m-1}$ we consider the so-called elementary interval in base $b$ of the form $J_a = \left[\frac{a}{b^m}, \frac{a+1}{b^m}\right)$. For $n \in \mathbb{N}_0$ with $b$-adic digit expansion $n = n_0 + n_1 b + n_2 b^2 + \cdots$, the element $x_n = \phi_b(n)$ belongs to $J_a$ if and only if

$$\frac{a}{b^m} \leq \frac{n_0}{b} + \frac{n_1}{b^2} + \frac{n_2}{b^3} + \cdots < \frac{a+1}{b^m}.$$

Multiplying by $b^m$ gives

$$a \leq \underbrace{n_0 b^{m-1} + n_1 b^{m-2} + \cdots + n_{m-1}}_{\in\mathbb{N}_0} + \underbrace{\frac{n_m}{b} + \cdots}_{\in[0,1)} < a + 1.$$

This is equivalent to $a = n_0 b^{m-1} + \cdots + n_{m-1}$ or, in terms of $b$-adic digits of $a$, to

$$n_0 = a_0, n_1 = a_1, \ldots, n_{m-1} = a_{m-1}.$$

This, in turn, is equivalent to

$$n \equiv a' \pmod{b^m}, \quad \text{where} \quad a' = a_0 + a_1 b + \cdots + a_{m-1} b^{m-1}.$$

Since the congruence $x \equiv a' \pmod{b^m}$ has a unique solution modulo $b^m$, it follows that exactly one of $b^m$ consecutive elements of the van der Corput sequence belongs to $J_a$. Hence, for $N \in \mathbb{N}$, it holds that

$$A(J_a, N) = \left\lfloor \frac{N}{b^m} \right\rfloor + \theta,$$

with $\theta \in \{0, 1\}$. Therefore, and since $\lim_{N\to\infty} \frac{1}{N} \left\lfloor \frac{N}{b^m} \right\rfloor = \frac{1}{b^m}$, we obtain

$$\lim_{N\to\infty} \frac{A(J_a, N)}{N} = \frac{1}{b^m} = \lambda_1(J_a),$$

where $\lambda_1$ denotes the one-dimensional Lebesgue measure.

Now let $J = [\alpha, \beta) \subseteq [0, 1)$ and let

$$\underline{J} = \bigcup_{a: J_a \subseteq J} J_a \quad \text{and} \quad \overline{J} = \bigcup_{a: J_a \cap J \neq \emptyset} J_a.$$

Then we have $\underline{J} \subseteq J \subseteq \overline{J}$ and $\lambda_1(\overline{J}) - \lambda_1(\underline{J}) \leq 2/b^m$, and consequently

$$\frac{A(\underline{J}, N)}{N} - \lambda_1(\overline{J}) \leq \frac{A(J, N)}{N} - \lambda_1(J) \leq \frac{A(\overline{J}, N)}{N} - \lambda_1(\underline{J}). \tag{2.5}$$

Note that for arbitrary $x, y, z \in \mathbb{R}$ and $\delta > 0$ we have

$$y - \delta \leq x \leq z + \delta \quad \Rightarrow \quad |x| \leq \delta + \max(|y|, |z|). \tag{2.6}$$

Therefore, (2.5) yields

$$\begin{aligned} &\left| \frac{A(J, N)}{N} - \lambda_1(J) \right| \\ &\leq \lambda_1(\overline{J}) - \lambda_1(\underline{J}) + \max\left\{ \left| \frac{A(\underline{J}, N)}{N} - \lambda_1(\underline{J}) \right|, \left| \frac{A(\overline{J}, N)}{N} - \lambda_1(\overline{J}) \right| \right\} \\ &\leq \frac{2}{b^m} + \max\left\{ \left| \frac{A(\underline{J}, N)}{N} - \lambda_1(\underline{J}) \right|, \left| \frac{A(\overline{J}, N)}{N} - \lambda_1(\overline{J}) \right| \right\}. \end{aligned}$$

Since

$$\lim_{N\to\infty}\left|\frac{A(\underline{J},N)}{N}-\lambda_1(\underline{J})\right|\le\sum_{a\colon J_a\subseteq J}\lim_{N\to\infty}\left|\frac{A(J_a,N)}{N}-\lambda_1(J_a)\right|=0$$

and, in the same manner, $\lim_{N\to\infty}\left|\frac{A(\overline{J},N)}{N}-\lambda_1(\overline{J})\right|=0$, it follows that

$$\lim_{N\to\infty}\left|\frac{A(J,N)}{N}-\lambda_1(J)\right|\le\frac{2}{b^m}.$$

Since $m\in\mathbb{N}$ was arbitrary we obtain

$$\lim_{N\to\infty}\left|\frac{A(J,N)}{N}-\lambda_1(J)\right|=0.$$

This means that the van der Corput sequence in base $b$ is uniformly distributed modulo one. □

▶ **Remark 2.11** It is not difficult to show that every uniformly distributed sequence in $[0,1)^s$ is dense in $[0,1]^s$. The converse is not true. A proof of these facts is left as an exercise. In particular, a Kronecker sequence $(\{n\boldsymbol{\alpha}\})_{n\in\mathbb{N}_0}$ is dense in $[0,1]^s$ if the numbers $1,\alpha_1,\dots,\alpha_s$ are linearly independent over the rationals. This is a well-known result from Number Theory known as *Kronecker's approximation theorem*.

## 2.2 Discrepancy

In this section we introduce quantitative measures for the assessment of the quality of uniform distribution. For $\boldsymbol{a},\boldsymbol{b}\in[0,1]^s$ with $\boldsymbol{a}=(a_1,\dots,a_s)$ and $\boldsymbol{b}=(b_1,\dots,b_s)$, the inequality $\boldsymbol{a}\le\boldsymbol{b}$ means that $a_j\le b_j$ for all $j=1,\dots,s$.

**Definition 2.12**

Let $\mathcal{P}$ be an $N$-element point set in $[0,1)^s$. The *extreme discrepancy $D_N$* of this point set is defined as

$$D_N(\mathcal{P})=\sup_{\substack{\boldsymbol{a},\boldsymbol{b}\in[0,1]^s\\ \boldsymbol{a}\le\boldsymbol{b}}}\left|\frac{A([\boldsymbol{a},\boldsymbol{b}),\mathcal{P},N)}{N}-\lambda_s([\boldsymbol{a},\boldsymbol{b}))\right|.$$

For an infinite sequence $\mathcal{S}$ the discrepancy $D_N(\mathcal{S})$ is the discrepancy of the first $N$ elements of $\mathcal{S}$.

Often it is enough to consider a slightly simplified notion of discrepancy:

**Definition 2.13**
Let $\mathcal{P}$ be an $N$-element point set in $[0,1)^s$. The *star discrepancy* $D_N^*$ of this point set is defined as

$$D_N^*(\mathcal{P}) = \sup_{\boldsymbol{a}\in[0,1]^s} \left| \frac{A([\boldsymbol{0},\boldsymbol{a}),\mathcal{P},N)}{N} - \lambda_s([\boldsymbol{0},\boldsymbol{a}))\right|.$$

For an infinite sequence $\mathcal{S}$ the star discrepancy $D_N^*(\mathcal{S})$ is the star discrepancy of the first $N$ elements of $\mathcal{S}$.

We have the following relation between extreme and star discrepancy:

**Proposition 2.14.** *For every $N$-element point set $\mathcal{P}$ in $[0,1)^s$,*

$$D_N^*(\mathcal{P}) \le D_N(\mathcal{P}) \le 2^s\, D_N^*(\mathcal{P}).$$

*Proof.* The left inequality is clear from the definitions of extreme and star discrepancy. We prove the right-hand inequality for dimension $s = 1$ and $s = 2$ only. The case $s \ge 3$ can be shown in the same manner. Let first $s = 1$ and $[\alpha,\beta) \subseteq [0,1)$. Then $[\alpha,\beta) = [0,\beta) \setminus [0,\alpha)$,

$$A([\alpha,\beta),N) = A([0,\beta),N) - A([0,\alpha),N),$$

and

$$\lambda_1([\alpha,\beta)) = \lambda_1([0,\beta)) - \lambda_1([0,\alpha)).$$

Consequently,

$$\begin{aligned}\frac{A([\alpha,\beta),N)}{N} - \lambda_1([\alpha,\beta)) =& \frac{A([0,\beta),N)}{N} - \lambda_1([0,\beta)) - \\ &- \left(\frac{A([0,\alpha),N)}{N} - \lambda_1([0,\alpha))\right).\end{aligned}$$

Taking the absolute value and applying the triangle inequality leads to

$$\left|\frac{A([\alpha,\beta),N)}{N} - \lambda_1([\alpha,\beta))\right| \le 2\, D_N^*(\mathcal{P})$$

and therefore we get

$$D_N(\mathcal{P}) \le 2\, D_N^*(\mathcal{P}),$$

as claimed.

In the case $s = 2$ for $J = [\alpha_1,\beta_1) \times [\alpha_2,\beta_2)$ we have the following inclusion-exclusion principle (see also Fig. 2.2):

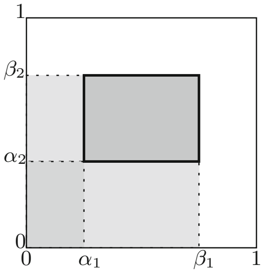

**Fig. 2.2** Inclusion-exclusion principle

$$\begin{aligned} J &= ([0,\beta_1)\setminus[0,\alpha_1))\times([0,\beta_2)\setminus[0,\alpha_2)) \\ &= \Big\{([0,\beta_1)\times[0,\beta_2))\setminus\{([0,\alpha_1)\times[0,\beta_2))\cup([0,\beta_1)\times[0,\alpha_2))\}\Big\} \\ &\cup([0,\alpha_1)\times[0,\alpha_2)) \end{aligned}$$

and hence

$$\begin{aligned} A(J,N) =& A([0,\beta_1)\times[0,\beta_2),N) \\ & - A([0,\alpha_1)\times[0,\beta_2),N) - A([0,\beta_1)\times[0,\alpha_2),N) \\ & + A([0,\alpha_1)\times[0,\alpha_2),N). \end{aligned}$$

A similar equality is true for $\lambda_2$. Now the result follows as for the case $s = 1$. □

Based on the notions of discrepancy we can give the following criterion for uniform distribution modulo one:

**Theorem 2.15.** *Let $\mathcal{S}$ be a sequence in $[0,1)^s$. The following assertions are equivalent:*
*(a)* *$\mathcal{S}$ is uniformly distributed modulo one;*
*(b)* $\lim_{N\to\infty} D_N^*(\mathcal{S}) = 0$;
*(c)* $\lim_{N\to\infty} D_N(\mathcal{S}) = 0$.

For the proof of this result we need the following lemma:

**Lemma 2.16.** *For $j = 1,\ldots,s$ let $u_j, v_j \in [0,1]$ with $|u_j - v_j| \le \delta$. Then*

$$\left|\prod_{j=1}^{s} u_j - \prod_{j=1}^{s} v_j\right| \le 1-(1-\delta)^s \le s\,\delta.$$

*Proof.* We use induction on $s$. Trivially, the assertion holds true for $s = 1$. Assume that it holds true for $s \in \mathbb{N}$. Without loss of generality we may assume that $u_{s+1} \geq v_{s+1}$. Then we have

$$\begin{aligned}\left|\prod_{j=1}^{s+1} u_j - \prod_{j=1}^{s+1} v_j\right| &= \left|(u_{s+1} - v_{s+1}) \prod_{j=1}^{s} u_j + v_{s+1}\left(\prod_{j=1}^{s} u_j - \prod_{j=1}^{s} v_j\right)\right| \\ &\leq |u_{s+1} - v_{s+1}| \cdot 1 + v_{s+1}(1 - (1-\delta)^s) \\ &= u_{s+1} - v_{s+1}(1-\delta)^s \\ &= u_{s+1}(1 - (1-\delta)^s) + (u_{s+1} - v_{s+1})(1-\delta)^s \\ &\leq 1 - (1-\delta)^s + \delta(1-\delta)^s \\ &= 1 - (1-\delta)^{s+1}.\end{aligned}$$

Hence the left-hand inequality is shown.

According to the Mean Value Theorem, we have for all $y, z \in \mathbb{R}$ with $z \leq y$ that $y^s - z^s = s\xi^{s-1}(y - z)$ for some $\xi \in (z, y)$. Choosing $y = 1$ and $z = 1 - \delta$ yields the right-hand inequality of the lemma. □

*Proof of Theorem 2.15.* The equivalence of ($b$) and ($c$) follows from Proposition 2.14.

($c$) ⇒ ($a$): Assume that $\lim_{N\to\infty} D_N(\mathcal{S}) = 0$, i.e.,

$$\lim_{N\to\infty} \sup_{\substack{\boldsymbol{a},\boldsymbol{b}\in[0,1]^s \\ \boldsymbol{a}\leq\boldsymbol{b}}} \left|\frac{A([\boldsymbol{a},\boldsymbol{b}),\mathcal{S},N)}{N} - \lambda_s([\boldsymbol{a},\boldsymbol{b}))\right| = 0.$$

Then obviously

$$\lim_{N\to\infty} \frac{A([\boldsymbol{a},\boldsymbol{b}),\mathcal{S},N)}{N} = \lambda_s([\boldsymbol{a},\boldsymbol{b}))$$

for all intervals of the form $[\boldsymbol{a},\boldsymbol{b}) \subseteq [0,1)^s$, and so the sequence $\mathcal{S}$ is uniformly distributed modulo one.

($a$) ⇒ ($b$): Assume that the sequence $\mathcal{S}$ is uniformly distributed modulo one. Let $J$ be a sub-interval of $[0,1)^s$ of the form $[\boldsymbol{0},\boldsymbol{a})$ and choose $m \in \mathbb{N}$, $m \geq 2$. Now we proceed in a similar way to what we did in the proof of Proposition 2.10. For $\boldsymbol{k} = (k_1,\ldots,k_s) \in \{0,1,\ldots,m-1\}^s$, put $Q_{\boldsymbol{k}} := \prod_{j=1}^{s}\left[\frac{k_j}{m}, \frac{k_j+1}{m}\right)$ and define

$$\underline{J} = \bigcup_{\boldsymbol{k}:Q_{\boldsymbol{k}}\subseteq J} Q_{\boldsymbol{k}} \quad \text{and} \quad \overline{J} = \bigcup_{\boldsymbol{k}:Q_{\boldsymbol{k}}\cap J\neq\emptyset} Q_{\boldsymbol{k}}.$$

Obviously, $\underline{J}$ and $\overline{J}$ are intervals anchored at the origin ($\underline{J}$ may be empty) which satisfy $\underline{J} \subseteq J \subseteq \overline{J}$ and corresponding edge-lengths of $\underline{J}$ and $\overline{J}$ differ by at most $1/m$. Hence we obtain from Lemma 2.16 that

$$\lambda_s(\overline{J}) - \lambda_s(\underline{J}) \le \frac{s}{m}.$$

Now we have

$$\frac{A(\underline{J}, \mathcal{S}, N)}{N} - \lambda_s(\overline{J}) \le \frac{A(J, \mathcal{S}, N)}{N} - \lambda_s(J) \le \frac{A(\overline{J}, \mathcal{S}, N)}{N} - \lambda_s(\underline{J}),$$

whence

$$\frac{A(J, \mathcal{S}, N)}{N} - \lambda_s(J) \begin{cases} \ge \frac{A(\underline{J}, \mathcal{S}, N)}{N} - \lambda_s(\underline{J}) - (\lambda_s(\overline{J}) - \lambda_s(\underline{J})), \\ \le \frac{A(\overline{J}, \mathcal{S}, N)}{N} - \lambda_s(\overline{J}) + (\lambda_s(\overline{J}) - \lambda_s(\underline{J})). \end{cases}$$

Using (2.6) we obtain

$$\begin{aligned} \left| \frac{A(J, \mathcal{S}, N)}{N} - \lambda_s(J) \right| \le & \lambda_s(\overline{J}) - \lambda_s(\underline{J}) \\ & + \max\left( \left| \frac{A(\underline{J}, \mathcal{S}, N)}{N} - \lambda_s(\underline{J}) \right|, \left| \frac{A(\overline{J}, \mathcal{S}, N)}{N} - \lambda_s(\overline{J}) \right| \right) \\ \le & \frac{s}{m} + \max_{\boldsymbol{k} \in \{0,1,\dots,m\}^s} \left| \frac{A([\boldsymbol{0}, \frac{1}{m}\boldsymbol{k}), \mathcal{S}, N)}{N} - \lambda_s\left( \left[ \boldsymbol{0}, \frac{1}{m}\boldsymbol{k} \right) \right) \right|. \end{aligned}$$

Note that the final term does not depend on $J$.

Since the sequence $\mathcal{S}$ is uniformly distributed modulo one we have, for every $\boldsymbol{k} \in \{0, 1, \dots, m\}^s$,

$$\lim_{N\to\infty} \left| \frac{A([\boldsymbol{0}, \frac{1}{m}\boldsymbol{k}), \mathcal{S}, N)}{N} - \lambda_s\left( \left[ \boldsymbol{0}, \frac{1}{m}\boldsymbol{k} \right) \right) \right| = 0,$$

and thus also

$$\lim_{N\to\infty} \max_{\boldsymbol{k} \in \{0,1,\dots,m\}^s} \left| \frac{A([\boldsymbol{0}, \frac{1}{m}\boldsymbol{k}), \mathcal{S}, N)}{N} - \lambda_s\left( \left[ \boldsymbol{0}, \frac{1}{m}\boldsymbol{k} \right) \right) \right| = 0.$$

Consequently,

$$0 \le \limsup_{N\to\infty} D_N^*(\mathcal{S}) \le \frac{s}{m},$$

and since $m$ can be chosen arbitrarily large, the result follows. □

▶ **Remark 2.17** We learn from the proof of Theorem 2.15 that for the star discrepancy of every $N$-element point set in $[0,1)^s$ and for every $m \in \mathbb{N}$, $m \geq 2$, we have

$$D_N^*(\mathcal{P}) \leq \frac{s}{m} + \max_{\boldsymbol{k}\in\{0,1,\ldots,m\}^s} \left| \frac{A([\mathbf{0}, \frac{1}{m}\boldsymbol{k}), \mathcal{P}, N)}{N} - \lambda_s\left(\left[\mathbf{0}, \frac{1}{m}\boldsymbol{k}\right)\right) \right|.$$

This means that the supremum in the definition of star discrepancy can be replaced by a maximum over a finite set consisting of $(m+1)^s$ elements with an error of at most $s/m$.

Theorem 2.15 shows that a sequence is uniformly distributed modulo one if and only if its (star) discrepancy converges to zero as $N$ approaches infinity. We now show that the order of convergence to zero cannot be faster than $1/N$.

**Proposition 2.18.** *For every $N$-element point set $\mathcal{P}$ in $[0,1)^s$ we have*

$$D_N(\mathcal{P}) \geq \frac{1}{N} \quad \textit{and} \quad D_N^*(\mathcal{P}) \geq \frac{1}{2^s N}.$$

*Proof.* Choose $0 < \varepsilon \leq 1/N$ and let $\boldsymbol{x} = (x_1, \ldots, x_s)$ be an element of the point set $\mathcal{P}$. Put

$$J = \left([x_1, x_1 + \varepsilon^{1/s}) \times \cdots \times [x_s, x_s + \varepsilon^{1/s})\right) \cap [0,1]^s.$$

Since $\boldsymbol{x} \in J$, it follows that

$$D_N(\mathcal{P}) \geq \frac{A(J,N)}{N} - \lambda_s(J) \geq \frac{1}{N} - \varepsilon.$$

Since $\varepsilon$ can be chosen arbitrarily close to zero, it follows that $D_N(\mathcal{P}) \geq \frac{1}{N}$. The result for the star discrepancy can be obtained from Proposition 2.14. □

If one thinks of a point set whose elements are evenly distributed in the unit cube, on first sight one might have a regular lattice in mind. By a *regular lattice* in $[0,1)^s$ consisting of $m^s$ elements, where $m \in \mathbb{N}$, $m \geq 2$, we mean the point set

$$\Gamma_{m,s} = \left\{ \left( \frac{2n_1+1}{2m}, \ldots, \frac{2n_s+1}{2m} \right) : n_1, \ldots, n_s \in \{0, \ldots, m-1\} \right\}. \tag{2.7}$$

Sometimes one also speaks of a *centered* regular lattice.

However, for dimension $s \geq 2$ it turns out that this is not a good choice.

**Theorem 2.19.** *Let $s, m \in \mathbb{N}$, $m \geq 2$. For the star discrepancy of the regular lattice $\Gamma_{m,s}$ with $N = m^s$ elements in $[0,1)^s$ it holds that*

$$D_N^*(\Gamma_{m,s}) = 1 - \left(1 - \frac{1}{2m}\right)^s.$$

*Proof.* Since $\Gamma_{m,s} \subseteq [0, 1 - 1/(2m)]^s =: B_{m,s}$, we find that

$$D_N^*(\Gamma_{m,s}) \geq \left| \frac{A(B_{m,s}, \Gamma_{m,s}, N)}{N} - \lambda_s(B_{m,s}) \right| = 1 - \left(1 - \frac{1}{2m}\right)^s.$$

To prove an upper bound, let $J = [0, \alpha_1) \times \cdots \times [0, \alpha_s) \subseteq [0,1)^s$. If $\min_{j=1,\ldots,s} \alpha_j < \frac{1}{2m}$, then

$$A(J, \Gamma_{m,s}, N) = 0 \quad \text{and} \quad \lambda_s(J) \leq \frac{1}{2m} \leq 1 - \left(1 - \frac{1}{2m}\right)^s.$$

Otherwise, let $a_j \in \{0, \ldots, m-1\}$ for $j = 1, \ldots, s$ be such that $\frac{2a_j+1}{2m} < \alpha_j \leq \frac{2a_j+3}{2m}$. Then we have $A(J, \Gamma_{m,s}, N) = \prod_{j=1}^s (a_j + 1)$ and

$$0 \leq \frac{A(J, \Gamma_{m,s}, N)}{N} - \lambda_s(J) \leq \prod_{j=1}^s \frac{a_j + 1}{m} - \prod_{j=1}^s \frac{2a_j + 1}{2m}.$$

Therefore and by Lemma 2.16 we obtain

$$\left| \frac{A(J, \Gamma_{m,s}, N)}{N} - \lambda_s(J) \right| \leq \left| \prod_{j=1}^s \frac{a_j + 1}{m} - \prod_{j=1}^s \frac{2a_j + 1}{2m} \right| \leq 1 - \left(1 - \frac{1}{2m}\right)^s.$$

Since this upper bound is independent of the specific choice of the interval $J$, we obtain that

$$D_N^*(\Gamma_{m,s}) \leq 1 - \left(1 - \frac{1}{2m}\right)^s.$$

Hence the result follows. □

► **Remark 2.20** For $s = 1$ we obtain $D_N^*(\Gamma_{m,1}) = 1/(2N)$, which is best possible, as we can see from Proposition 2.18. Hence in dimension $s = 1$ we have found, for any $N \in \mathbb{N}$, the $N$-element point set with lowest star discrepancy. Unfortunately, this is not true for dimensions $s \geq 2$. We have

$$\frac{1}{2N^{1/s}} = \frac{1}{2m} \leq 1 - \left(1 - \frac{1}{2m}\right)^s \leq \frac{s}{2m} = \frac{s}{2N^{1/s}},$$

and so

$$D_N^*(\Gamma_{m,s}) \asymp_s \frac{1}{N^{1/s}}.$$

For large dimensions $s$ this convergence rate is rather weak, as we shall see in Sect. 2.4, where we will encounter point sets with much lower discrepancy.

The star discrepancy of a point set $\mathcal{P}$ may also be defined as the supremum norm of the function $\Delta_{\mathcal{P},N} : [0,1]^s \to \mathbb{R}$,

$$\Delta_{\mathcal{P},N}(\boldsymbol{y}) = \frac{A([\boldsymbol{0},\boldsymbol{y}),\mathcal{P},N)}{N} - \lambda_s([\boldsymbol{0},\boldsymbol{y})),$$

i.e.,

$$D_N^*(\mathcal{P}) = \|\Delta_{\mathcal{P},N}\|_{L_\infty}.$$

The function $\Delta_{\mathcal{P},N}$ is sometimes called *local discrepancy* or *discrepancy function* of the point set $\mathcal{P}$. From this point of view it is near at hand to also consider other norms of the local discrepancy, e.g., the $L_p$-norm for $p \geq 1$.

**Definition 2.21**

Let $\mathcal{P}$ be an $N$-element point set in $[0,1)^s$ and let $p \in [1,\infty)$. The $L_p$ *discrepancy* (sometimes also referred to as $L_p$ star discrepancy) of $\mathcal{P}$ is defined as

$$L_{p,N}(\mathcal{P}) = \|\Delta_{\mathcal{P},N}\|_{L_p} = \left(\int_{[0,1]^s} \left|\frac{A([\boldsymbol{0},\boldsymbol{y}),\mathcal{P},N)}{N} - \lambda_s([\boldsymbol{0},\boldsymbol{y}))\right|^p \mathrm{d}\boldsymbol{y}\right)^{1/p}.$$

For an infinite sequence $\mathcal{S}$ the $L_p$ discrepancy $L_{p,N}(\mathcal{S})$ is the $L_p$ discrepancy of the first $N$ elements of $\mathcal{S}$.

From the monotonicity of the $L_p$-norm it follows immediately that $L_{p_1,N}(\mathcal{P}) \leq L_{p_2,N}(\mathcal{P})$ whenever $p_1 \leq p_2$. Also there is a relation between the star- and the $L_p$ discrepancy which we state here without proof.

**Proposition 2.22.** *For every $N$-element point set $\mathcal{P}$ in $[0,1)^s$ we have*

$$L_{p,N}(\mathcal{P}) \leq D_N^*(\mathcal{P}) \leq c(s,p) L_{p,N}^{\frac{p}{p+s}}(\mathcal{P}),$$

*where the positive quantity $c(s,p)$ only depends on $s$ and $p$, but not on $N$.*

As a consequence we obtain a further criterion for uniform distribution modulo one of infinite sequences.

**Corollary 2.23.** *Let $p \in [1, \infty)$. A sequence $\mathcal{S}$ in $[0,1)^s$ is uniformly distributed modulo one if and only if $\lim_{N\to\infty} L_{p,N}(\mathcal{S}) = 0$.*

## 2.3 Bounds on the Discrepancy

The lower bound on the star discrepancy from Proposition 2.18 is rather weak, at least for large dimensions. Only for dimension one it is best possible. Now we present a much stronger result, which was first shown by K. F. Roth in 1954.

**Theorem 2.24 (Roth).** *For every dimension $s \in \mathbb{N}$ there exists a quantity $c_s > 0$ with the following property: for every $N$-element point set $\mathcal{P}$ in $[0,1)^s$,*

$$D_N(\mathcal{P}) \geq D_N^*(\mathcal{P}) \geq L_{2,N}(\mathcal{P}) \geq c_s \frac{(\log N)^{\frac{s-1}{2}}}{N}.$$

For the proof of this result we use the method of Roth, but we follow the exposition in [65]. For the sake of simplicity, we restrict ourselves to the case of dimension $s = 2$. The general case can be shown in the same manner, but with more technical effort. The proof is technical and can be skipped by beginners.

*Proof of Theorem 2.24.* For $\boldsymbol{x} \in [0,1]^2$ let

$$D(\boldsymbol{x}) = N\lambda_2([\boldsymbol{0}, \boldsymbol{x})) - A([\boldsymbol{0}, \boldsymbol{x}), \mathcal{P}, N).$$

By the Cauchy-Schwarz-inequality, for any function $F : [0,1]^2 \to \mathbb{R}$

$$\int_{[0,1]^2} F(\boldsymbol{x}) D(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \leq \left(\int_{[0,1]^2} F^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\right)^{1/2} \left(\int_{[0,1]^2} D^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\right)^{1/2},$$

and hence, provided that $\int_{[0,1]^2} F^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} > 0$,

$$N L_{2,N}(\mathcal{P}) = \left(\int_{[0,1]^2} D^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\right)^{1/2} \geq \frac{\int_{[0,1]^2} F(\boldsymbol{x}) D(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}}{\left(\int_{[0,1]^2} F^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\right)^{1/2}}.$$

Now it is our aim to chose the function $F$ such that $\int_{[0,1]^2} F^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = O(\log N)$, but $\int_{[0,1]^2} F(\boldsymbol{x}) D(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$ is at least of the order of magnitude $\log N$.

The function $F$ will depend on the given point set $\mathcal{P}$. For $x \in \mathbb{R}$ define

$$\psi(x) = \begin{cases} 1 & \text{if } x \in [k, k+1/2) \text{ for some } k \in \mathbb{Z}, \\ -1 & \text{if } x \in [k+1/2, k+1) \text{ for some } k \in \mathbb{Z}. \end{cases}$$

Note that for $i \in \mathbb{N}_0$ and $a \in \{0, \ldots, 2^i - 1\}$ we have

$$\int_{a/2^i}^{(a+1)/2^i} \psi(2^i y)\,\mathrm{d}y = \frac{1}{2^i} \int_a^{a+1} \psi(y)\,\mathrm{d}y = 0.$$

Choose $m \in \mathbb{N}$ such that $2N \le 2^m < 4N$. For $j \in \{0, 1, \ldots, m\}$ define functions $f_j : [0,1]^2 \to \{-1, 0, 1\}$ in the following way: let

$$R_{a,b}^{(j)} = \left[\frac{a}{2^{m-j}}, \frac{a+1}{2^{m-j}}\right) \times \left[\frac{b}{2^j}, \frac{b+1}{2^j}\right),$$

where $a \in \{0, 1, \ldots, 2^{m-j} - 1\}$ and $b \in \{0, 1, \ldots, 2^j - 1\}$. Then, for $\boldsymbol{y} = (y_1, y_2) \in R_{a,b}^{(j)}$, we set

$$f_j(\boldsymbol{y}) = \begin{cases} 0 & \text{if } R_{a,b}^{(j)} \cap \mathcal{P} \neq \emptyset, \\ \psi(2^{m-j} y_1)\psi(2^j y_2) & \text{if } R_{a,b}^{(j)} \cap \mathcal{P} = \emptyset. \end{cases} \tag{2.8}$$

Now we show that the functions $f_j$ are mutually orthogonal. For $i < j$ there exists some $b \in \{0, \ldots, 2^i - 1\}$ such that either

$$\left[\frac{a}{2^j}, \frac{a+1}{2^j}\right) \subseteq \left[\frac{b}{2^i}, \frac{b}{2^i} + \frac{1}{2^{i+1}}\right), \quad \text{or} \quad \left[\frac{a}{2^j}, \frac{a+1}{2^j}\right) \subseteq \left[\frac{b}{2^i} + \frac{1}{2^{i+1}}, \frac{b+1}{2^i}\right).$$

Hence $\psi(2^i y) = c \in \{-1, 0, 1\}$ is constant on the intervals $\left[a/2^j, (a+1)/2^j\right)$ and therefore

$$\int_{a/2^j}^{(a+1)/2^j} \psi(2^j y)\psi(2^i y)\,\mathrm{d}y = c \int_{a/2^j}^{(a+1)/2^j} \psi(2^j y)\,\mathrm{d}y = 0.$$

For $a \in \{0, \ldots, 2^{m-i} - 1\}$ and $b \in \{0, \ldots, 2^j - 1\}$ let

$$R_{a,b}^{(i,j)} = \left[\frac{a}{2^{m-i}}, \frac{a+1}{2^{m-i}}\right) \times \left[\frac{b}{2^j}, \frac{b+1}{2^j}\right).$$

Note that $i < j$ implies that $R_{a,b}^{(i,j)} \subseteq R_{a,\lfloor b2^{i-j} \rfloor}^{(i)}$. Hence if $R_{a,b}^{(i,j)} \cap \mathcal{P} \neq \emptyset$, then also $R_{a,\lfloor b2^{i-j} \rfloor}^{(i)} \cap \mathcal{P} \neq \emptyset$. This implies that $f_i(\boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in R_{a,\lfloor b2^{i-j} \rfloor}^{(i)}$ which in turn yields $f_i(\boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in R_{a,b}^{(i,j)}$. Consequently,

$$
\begin{aligned}
&\int_{[0,1]^2} f_i(\boldsymbol{y}) f_j(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\
&\quad = \sum_{a=0}^{2^{m-i}-1}\sum_{b=0}^{2^j-1} \int_{R_{a,b}^{(i,j)}} f_i(\boldsymbol{y}) f_j(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\
&\quad = \underbrace{\sum_{a=0}^{2^{m-i}-1}\sum_{b=0}^{2^j-1}}_{R_{a,b}^{(i,j)}\cap\mathcal{P}=\emptyset} \int_{R_{a,b}^{(i,j)}} \psi(2^{m-i}y_1)\psi(2^i y_2)\psi(2^{m-j}y_1)\psi(2^j y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2\\
&\quad = \underbrace{\sum_{a=0}^{2^{m-i}-1}\sum_{b=0}^{2^j-1}}_{R_{a,b}^{(i,j)}\cap\mathcal{P}=\emptyset} \int_{a/2^{m-i}}^{(a+1)/2^{m-i}} \psi(2^{m-i}y_1)\psi(2^{m-j}y_1)\,\mathrm{d}y_1 \int_{b/2^j}^{(b+1)/2^j} \psi(2^i y_2)\psi(2^j y_2)\,\mathrm{d}y_2\\
&\quad = 0,
\end{aligned}
$$

i.e., the $f_j$ are mutually orthogonal.

Now we put

$$
F(\boldsymbol{x}) = f_0(\boldsymbol{x}) + f_1(\boldsymbol{x}) + \cdots + f_m(\boldsymbol{x}). \tag{2.9}
$$

Using the orthogonality of the $f_j$ and the fact that $2^m < 4N$ it follows that

$$
\begin{aligned}
\int_{[0,1]^2} F^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \sum_{i,j=0}^{m}\int_{[0,1]^2} f_i(\boldsymbol{x}) f_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{i=0}^{m}\int_{[0,1]^2} f_i^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\
&\leq \sum_{i=0}^{m} 1 = m+1 < \log_2(4N)+1 = \log_2 N + 3.
\end{aligned}
$$

Next we show that for all $j \in \{0,1,\ldots,m\}$ we have

$$
\int_{[0,1]^2} f_j(\boldsymbol{x}) D(\boldsymbol{x})\,\mathrm{d}x \geq \frac{1}{2^8}. \tag{2.10}
$$

To this end we consider $f_j D$ on the rectangles $R_{a,b}^{(j)}$ from the definition of the function $f_j$. We only have to consider empty $R_{a,b}^{(j)}$, since otherwise $f_j(\boldsymbol{x}) = 0$ for every $\boldsymbol{x} \in R_{a,b}^{(j)}$. Here and in the following we speak of an *empty* rectangle $R_{a,b}^{(j)}$ whenever $R_{a,b}^{(j)} \cap \mathcal{P} = \emptyset$.

In total, there are $2^m \geq 2N$ rectangles $R_{a,b}^{(j)}$, of which at most $N$ contain a point from $\mathcal{P}$. Hence, there are at least $N$ empty rectangles $R_{a,b}^{(j)}$. Thus it suffices to show that for every empty rectangle $R_{a,b}^{(j)}$ we have

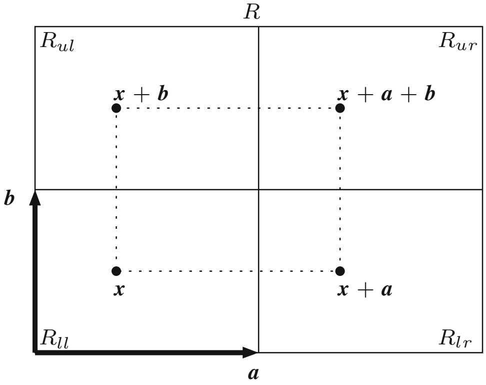
$R$
$R_{ul}$
$R_{ur}$
$\boldsymbol{x}+\boldsymbol{b}$
$\boldsymbol{x}+\boldsymbol{a}+\boldsymbol{b}$
$\boldsymbol{b}$
$\boldsymbol{x}$
$\boldsymbol{x}+\boldsymbol{a}$
$R_{ll}$
$R_{lr}$
$\boldsymbol{a}$

**Fig. 2.3** An empty rectangle $R_{a,b}^{(j)}$ with quadrants $R_{ll}$, $R_{lr}$, $R_{ul}$, $R_{ur}$

$$\int_{R_{a,b}^{(j)}} f_j(\boldsymbol{x}) D(\boldsymbol{x})\,\mathrm{d}x \geq \frac{1}{2^8 N}. \tag{2.11}$$

Let $R_{a,b}^{(j)}$ be an empty rectangle and let $R_{ll}$ the lower left quadrant of this rectangle and accordingly $R_{lr}$, $R_{ul}$, $R_{ur}$ (see Fig. 2.3).

Let $\boldsymbol{a}, \boldsymbol{b}$ be given as in Fig. 2.3. Then

$$\begin{aligned}
&\int_{R_{a,b}^{(j)}} f_j(\boldsymbol{x}) D(\boldsymbol{x})\,\mathrm{d}x \\
&= \int_{R_{ll}} D(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{R_{lr}} D(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{R_{ul}} D(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_{R_{ur}} D(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= \int_{R_{ll}} [D(\boldsymbol{x}) - D(\boldsymbol{x}+\boldsymbol{a}) - D(\boldsymbol{x}+\boldsymbol{b}) + D(\boldsymbol{x}+\boldsymbol{a}+\boldsymbol{b})]\,\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Since

$$\begin{aligned}
\lambda_2(R_{ll}) &= \lambda_2([\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{a}+\boldsymbol{b})) \\
&= \lambda_2([\boldsymbol{0}, \boldsymbol{x})) - \lambda_2([\boldsymbol{0}, \boldsymbol{x}+\boldsymbol{a})) - \lambda_2([\boldsymbol{0}, \boldsymbol{x}+\boldsymbol{b})) + \lambda_2([\boldsymbol{0}, \boldsymbol{x}+\boldsymbol{a}+\boldsymbol{b})),
\end{aligned}$$

and since

$$\begin{aligned}
0 =& A([\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{a}+\boldsymbol{b}), \mathcal{P}, N) \\
=& A([\boldsymbol{0}, \boldsymbol{x}), \mathcal{P}, N) - A([\boldsymbol{0}, \boldsymbol{x}+\boldsymbol{a}), \mathcal{P}, N) \\
&- A([\boldsymbol{0}, \boldsymbol{x}+\boldsymbol{b}), \mathcal{P}, N) + A([\boldsymbol{0}, \boldsymbol{x}+\boldsymbol{a}+\boldsymbol{b}), \mathcal{P}, N),
\end{aligned}$$

it follows that

$$D(\boldsymbol{x}) - D(\boldsymbol{x}+\boldsymbol{a}) - D(\boldsymbol{x}+\boldsymbol{b}) + D(\boldsymbol{x}+\boldsymbol{a}+\boldsymbol{b}) = N\lambda_2(R_{ll}).$$

Hence

$$\int_{R_{a,b}^{(j)}} f_j(\boldsymbol{x})D(\boldsymbol{x})\,\mathrm{d}x = \int_{R_{ll}} N\lambda_2(R_{ll})\,\mathrm{d}\boldsymbol{x} = N\lambda_2(R_{ll})^2 = \frac{N}{2^{2m+4}} > \frac{1}{2^{m+6}} > \frac{1}{2^8 N}.$$

Hence (2.11) as well as (2.10) are shown.

From (2.10) we obtain

$$\int_{[0,1]^2} F(\boldsymbol{x})D(\boldsymbol{x})\,\mathrm{d}x \geq \frac{m+1}{2^8} \geq \frac{\log_2 N}{2^8}.$$

Altogether it follows that

$$NL_{2,N}(\mathcal{P}) \geq \frac{\log_2 N}{2^8\sqrt{\log_2 N + 3}} \geq \frac{\sqrt{\log_2 N}}{2^9},$$

as claimed. □

It is known that Roth's lower bound is best possible for the $L_2$ discrepancy in the sense that for every $s, N \in \mathbb{N}$, $N \geq 2$, there exists an $N$-element point set $\mathcal{P}$ in $[0,1)^s$ with

$$L_{2,N}(\mathcal{P}) \leq \tilde{c}_s \frac{(\log N)^{\frac{s-1}{2}}}{N},$$

where the positive $\tilde{c}_s$ depends on the dimension $s$, but not on $N$. A first explicit construction of such point sets in arbitrary dimension $s$ was provided by W.W.L. Chen and M.M. Skriganov in 2002. Recently, J. Dick and F. Pillichshammer presented an alternative construction.

On the other hand, Roth's lower bound is not best possible for the star discrepancy. In 2008 D. Bilyk, M.T. Lacey and A. Vagharshakyan proved in a joint work that, for every dimension $s \in \mathbb{N}$, $s \geq 2$, there exist quantities $c_s > 0$ and $\eta_s \in (0, \frac{1}{2})$ with the property that for every $N$-element point set $\mathcal{P}$ in $[0,1)^s$ we have

$$D_N^*(\mathcal{P}) \geq c_s \frac{(\log N)^{(s-1)/2+\eta_s}}{N}.$$

This result is currently the best lower estimate for the star discrepancy of finite point sets in dimension $s \geq 3$. For dimension $s = 2$ W.M. Schmidt proved the following result in 1972:

**Theorem 2.25 (Schmidt).** *There exists a constant $c > 0$ such that for the star discrepancy of any $N$-element point set $\mathcal{P}$ in $[0,1)^2$ we have*

$$D_N^*(\mathcal{P}) \geq c\frac{\log N}{N}.$$

We shall prove this result in Exercise 2.15. Schmidt's lower bound in dimension two is best possible in the order of magnitude in $N$. For $s \geq 3$ the exact asymptotic order of star discrepancy is still unknown. Although it is conjectured that the lower bound on the star discrepancy is significantly larger than Roth's lower bound, there is presently no consensus among the experts about its sharp form. There are many people who conjecture that the sharp exponent of the logarithm for $s \geq 3$ should be $s-1$. But there are also other opinions such as, for example, $s/2$. Both conjectures are consistent with Schmidt's lower bound in dimension $s = 2$. From above we know constructions of $N$-element point sets in dimension $s$ with star discrepancy of order $(\log N)^{s-1}/N$. We will encounter such constructions later on. However, the exact determination of the sharp lower bound on star discrepancy seems to be a very difficult problem.

For infinite sequences the situation is slightly different. The following lower bound can be deduced from Roth's result for finite point sets:

**Theorem 2.26.** *For every $s \in \mathbb{N}$ there exists a quantity $c_s' > 0$ such that for every infinite sequence $\mathcal{S}$ in $[0,1)^s$ we have*

$$D_N^*(\mathcal{S}) \geq c_s' \frac{(\log N)^{s/2}}{N} \quad \textit{for infinitely many } N \in \mathbb{N}.$$

*Proof.* Let $\mathcal{S} = (\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$, where $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,s})$. For fixed $N \in \mathbb{N}$ we consider the $N$-element point set $\mathcal{P} = \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{N-1}\}$ in $[0,1)^{s+1}$ given by

$$\boldsymbol{y}_n := (x_{n,1}, \ldots, x_{n,s}, n/N), \quad \text{for } n = 0, \ldots, N-1.$$

According to Theorem 2.24, there exist $a_1, \ldots, a_{s+1} \in [0,1]$ such that

$$\left| \frac{A(\prod_{j=1}^{s+1}[0,a_j), \mathcal{P}, N)}{N} - \prod_{j=1}^{s+1} a_j \right| \geq c_{s+1} \frac{(\log N)^{s/2}}{N}.$$

Choose $m \in \mathbb{N}$ such that $\frac{m-1}{N} < a_{s+1} \leq \frac{m}{N}$ or, equivalently, $-1 < Na_{s+1} - m \leq 0$.

By definition $A(\prod_{j=1}^{s+1}[0,a_j), \mathcal{P}, N)$ is the number of all $n \in \{0, \ldots, N-1\}$ for which we have $x_{n,j} \in [0, a_j)$ for all $j = 1, \ldots, s$ and $\frac{n}{N} \in [0, a_{s+1})$. Here the latter condition is equivalent to $n \in \{0, \ldots, m-1\}$. Therefore,

$$A\left(\prod_{j=1}^{s+1}[0,a_j), \mathcal{P}, N\right) = A\left(\prod_{j=1}^{s}[0,a_j), \mathcal{S}, m\right).$$

From this and the triangle inequality we obtain

$$\begin{aligned}
&\left| A\left(\prod_{j=1}^{s}[0,a_j),\mathcal{S},m\right) - m\,a_1\cdots a_s\right| \\
&\geq \left| A\left(\prod_{j=1}^{s+1}[0,a_j),\mathcal{P},N\right) - N\,a_1\cdots a_{s+1}\right| - |Na_1\cdots a_{s+1} - m\,a_1\cdots a_s| \\
&\geq c_{s+1}\,(\log N)^{s/2} - a_1\cdots a_s\,|Na_{s+1}-m| \geq c_{s+1}\,(\log N)^{s/2} - 1.
\end{aligned}$$

It follows that for $c_s' < c_{s+1}$ and $N \in \mathbb{N}$ large enough there exists some $m \in \{1,\ldots,N\}$ such that

$$mD_m^*(\mathcal{S}) \geq \left| A\left(\prod_{j=1}^{s}[0,a_j),\mathcal{S},m\right) - m\,a_1\cdots a_s\right| > c_s'\,(\log N)^{s/2} \geq c_s'(\log m)^{s/2}.$$

Now we show that the inequality

$$mD_m^*(\mathcal{S}) > c_s'(\log m)^{s/2} \tag{2.12}$$

holds true for infinitely many $m \in \mathbb{N}$. Assume, on the contrary, that (2.12) only holds for finitely many $m \in \mathbb{N}$ and let $m^*$ be the maximal integer with this property. Then choose $N \in \mathbb{N}$ such that

$$c_s'(\log N)^{s/2} > \max_{1\leq k\leq m^*} kD_k^*(\mathcal{S}).$$

For this $N$ it is possible, as shown above, to find an $m \in \{1,\ldots,N\}$ with the property that

$$mD_m^*(\mathcal{S}) > c_s'(\log N)^{s/2} \geq c_s'(\log m)^{s/2}.$$

If $m \leq m^*$, then we would have a contradiction, since

$$mD_m^*(\mathcal{S}) > c_s'(\log N)^{s/2} > \max_{1\leq k\leq m^*} kD_k^*(\mathcal{S}).$$

Hence, $m > m^*$. But since $mD_m^*(\mathcal{S}) \geq c_s'(\log m)^{s/2}$, this contradicts the maximality of $m^*$. Therefore, (2.12) is true for infinitely many $m \in \mathbb{N}$. □

For dimension $s = 1$ there is an improvement due to W.M. Schmidt from 1972 whose proof we leave as an exercise (cf. Exercise 2.16).

**Theorem 2.27 (Schmidt).** *There exists a constant $c > 0$ such that for every infinite sequence $\mathcal{S}$ in $[0,1)$ we have*

$$D_N^*(\mathcal{S}) \geq c\frac{\log N}{N} \quad \textit{for infinitely many } N \in \mathbb{N}.$$

Schmidt's lower bound in dimension $s = 1$ is best possible in the order of magnitude in $N$. For dimensions $s \geq 2$ it is conjectured that the sharp lower bound for the star discrepancy is significantly larger than the result from Theorem 2.26. Corresponding to the conjectures for finite point sets, there are some opinions that $s$ is the correct exponent of the $\log N$ term, and others that $\frac{s+1}{2}$ is the correct exponent. Again, both conjectures are consistent with Schmidt's result for dimension $s = 1$.

One can extend the method of proof of Theorem 2.26 to show that for every $s \in \mathbb{N}$ there exists a quantity $c_s'' > 0$ such that for every infinite sequence $\mathcal{S}$ in $[0,1)^s$ we have

$$L_{2,N}(\mathcal{S}) \geq c_s'' \frac{(\log N)^{s/2}}{N} \quad \text{for infinitely many } N \in \mathbb{N}.$$

This was first shown by P.D. Proinov in 1985. Recently J. Dick and F. Pillichshammer presented the first explicit construction of an infinite sequence whose $L_2$ discrepancy is of order of magnitude $(\log N)^{s/2}/N$, which proves that Proinov's lower bound is asymptotically best possible in $N$.

Now we turn our attention to inequalities which are often useful for proving upper bounds on the star discrepancy of point sets and sequences. The first result that we mention, though without proof, is the so-called Erdős-Turán-Koksma inequality. This fundamental inequality connects the behavior of discrepancy with that of exponential sums of the form as they appear in the Weyl Criterion. This way, the Erdős-Turán-Koksma inequality can be viewed as quantitative version of the Weyl Criterion.

**Theorem 2.28 (Erdős-Turán-Koksma Inequality).** *For every $s \in \mathbb{N}$ there exists a $c_s > 0$, such that for every $N$-element point set $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ in $[0,1)^s$ we have*

$$D_N(\mathcal{P}) \leq c_s \left( \frac{1}{m+1} + \sum_{0 < \|\boldsymbol{h}\|_\infty \leq m} \frac{1}{r_1(\boldsymbol{h})} \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{x}_n) \right| \right),$$

*where $m \in \mathbb{N}$ and where $r_1(\boldsymbol{h}) = \prod_{j=1}^s \max(1, |h_j|)$ and $\|\boldsymbol{h}\|_\infty = \max_{j=1,\ldots,s} |h_j|$ for $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{Z}^s$.*

References for the proof of this result can be found at the end of this section.

### Example 2.29

For $N \in \mathbb{N}$ let $\mathcal{P} = \{0, 1/N, 2/N, \ldots, (N-1)/N\}$. For all $h = 1, \ldots, N-1$ we have $\sum_{n=0}^{N-1} \exp(2\pi \mathrm{i} hn/N) = 0$. Hence, with the choice $m = N-1$, from Theorem 2.28 we obtain $D_N(\mathcal{P}) \leq c_1/N$.

Sometimes one has to estimate the star discrepancy of point sets whose components are rational numbers. In 1977 H. Niederreiter proved a discrepancy estimate for such point sets which is related to the Erdős-Turán-Koksma inequality. To state Niederreiter's result we have to introduce some notation.

For $M \in \mathbb{N}$, $M \geq 2$, put $C(M) = (-M/2, M/2] \cap \mathbb{Z}$ and $C_s(M) = C(M)^s$, the $s$-fold Cartesian product of $C(M)$. Further, write $C_s^*(M) = C_s(M) \setminus \{\mathbf{0}\}$. For $h \in C(M)$ put

$$r(h, M) = \begin{cases} M \sin(\pi |h|/M) & \text{if } h \neq 0, \\ 1 & \text{if } h = 0, \end{cases}$$

and for $\boldsymbol{h} = (h_1, \ldots, h_s) \in C_s(M)$ put $r(\boldsymbol{h}, M) = \prod_{j=1}^{s} r(h_j, M)$.

**Theorem 2.30 (Niederreiter).** *For an integer $M \geq 2$ and $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{N-1} \in \mathbb{Z}^s$, let $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ be the point set consisting of the fractional parts $\boldsymbol{x}_n = \{\boldsymbol{y}_n/M\}$ for $n = 0, \ldots, N-1$. Then*

$$D_N(\mathcal{P}) \leq 1 - \left(1 - \frac{1}{M}\right)^s + \sum_{\boldsymbol{h} \in C_s^*(M)} \frac{1}{r(\boldsymbol{h}, M)} \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{y}_n / M) \right|.$$

*Proof.* For $\boldsymbol{k} = (k_1, \ldots, k_s) \in \mathbb{Z}^s$, let

$$A(\boldsymbol{k}) := \#\{n \in \mathbb{N}_0 \,:\, 0 \leq n < N \text{ and } \boldsymbol{y}_n \equiv \boldsymbol{k} \pmod{M}\},$$

where a congruence between vectors is understood component wise. We have

$$\frac{1}{M} \sum_{h \in C(M)} \exp(2\pi \mathrm{i} ha/M) = \begin{cases} 1 & \text{if } a \equiv 0 \pmod{M}, \\ 0 & \text{if } a \not\equiv 0 \pmod{M}, \end{cases}$$

and therefore

$$\sum_{n=0}^{N-1} \frac{1}{M^s} \sum_{\boldsymbol{h} \in C_s(M)} \exp(2\pi \mathrm{i} \boldsymbol{h} \cdot (\boldsymbol{y}_n - \boldsymbol{k})/M)$$

$$= \sum_{n=0}^{N-1} \prod_{j=1}^{s} \underbrace{\left( \frac{1}{M} \sum_{h_j \in C(M)} \exp(2\pi \mathrm{i} h_j (y_{n,j} - k_j)/M) \right)}_{= \begin{cases} 1 & \text{if } y_{n,j} \equiv k_j \pmod{M}, \\ 0 & \text{otherwise,} \end{cases}} = A(\boldsymbol{k}).$$

Consequently,

$$A(\boldsymbol{k}) - \frac{N}{M^s} = \frac{1}{M^s} \sum_{\boldsymbol{h} \in C_s^*(M)} \exp(-2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{k}/M) \sum_{n=0}^{N-1} \exp(2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{y}_n/M).$$

Now let $J = \prod_{j=1}^{s} [u_j, v_j) \subseteq [0, 1)^s$. For $j = 1, \ldots, s$ let $a_j \in \mathbb{Z}$ be minimal such that $u_j \le a_j/M$ and $b_j \in \mathbb{Z}$ maximal such that $b_j/M < v_j$. In particular, we have $[a_j/M, b_j/M] \subseteq [u_j, v_j)$.

If $[a_j/M, b_j/M] = \emptyset$ for some $j \in \{1, \ldots, s\}$, then $A(J, \mathcal{P}, N) = 0$ and $v_j - u_j < 1/M$. Therefore,

$$\left| \frac{A(J, \mathcal{P}, N)}{N} - \lambda_s(J) \right| = \lambda_s(J) < \frac{1}{M} \le 1 - \left(1 - \frac{1}{M}\right)^s.$$

Now assume that $[a_j/M, b_j/M] \neq \emptyset$ for all $j = 1, \ldots, s$. We have

$$A(J, \mathcal{P}, N) = \sum_{\boldsymbol{a} \le \boldsymbol{k} \le \boldsymbol{b}} A(\boldsymbol{k}) \quad \text{and} \quad \sum_{\boldsymbol{a} \le \boldsymbol{k} \le \boldsymbol{b}} \frac{1}{M^s} = \frac{1}{M^s} \prod_{j=1}^{s} (b_j - a_j + 1),$$

where $\boldsymbol{a} \le \boldsymbol{k} \le \boldsymbol{b}$ means summation over all $\boldsymbol{k} \in \mathbb{Z}^s$ for which $a_j \le k_j \le b_j$ for all $j = 1, 2, \ldots, s$, and hence

$$\frac{A(J, \mathcal{P}, N)}{N} - \lambda_s(J) = \sum_{\boldsymbol{a} \le \boldsymbol{k} \le \boldsymbol{b}} \left( \frac{A(\boldsymbol{k})}{N} - \frac{1}{M^s} \right) + \frac{1}{M^s} \prod_{j=1}^{s} (b_j - a_j + 1) - \lambda_s(J)$$

$$= \frac{1}{M^s} \sum_{\boldsymbol{h} \in C_s^*(M)} \left( \sum_{\boldsymbol{a} \le \boldsymbol{k} \le \boldsymbol{b}} \exp(-2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{k}/M) \right) \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{y}_n/M)$$

$$+ \prod_{j=1}^{s} \frac{b_j - a_j + 1}{M} - \prod_{j=1}^{s} (v_j - u_j).$$

For all $j \in \{1, \ldots, s\}$ we have

$$\left| \frac{b_j - a_j + 1}{M} - (v_j - u_j) \right| < \frac{1}{M},$$

and so it follows from Lemma 2.16 that

$$\left|\prod_{j=1}^{s}\frac{b_j-a_j+1}{M}-\prod_{j=1}^{s}(v_j-u_j)\right|\le 1-\left(1-\frac{1}{M}\right)^s.$$

Consequently,

$$\begin{aligned}\left|\frac{A(J,\mathcal{P},N)}{N}-\lambda_s(J)\right| &\le 1-\left(1-\frac{1}{M}\right)^s\\ &+\frac{1}{M^s}\sum_{\boldsymbol{h}\in C_s^*(M)}\underbrace{\left|\sum_{\boldsymbol{a}\le\boldsymbol{k}\le\boldsymbol{b}}\exp(-2\pi\mathrm{i}\boldsymbol{h}\cdot\boldsymbol{k}/M)\right|}_{=:r^*(\boldsymbol{h},M)}\cdot\left|\frac{1}{N}\sum_{n=0}^{N-1}\exp(2\pi\mathrm{i}\boldsymbol{h}\cdot\boldsymbol{y}_n/M)\right|.\end{aligned}$$

Since $|z| = |\overline{z}|$ for any complex number $z$ and $|\exp(2\pi\mathrm{i}t)| = 1$ for any real number $t$, we obtain

$$\begin{aligned}r^*(\boldsymbol{h},M) &= \prod_{j=1}^{s}\left|\sum_{k_j=a_j}^{b_j}\exp(2\pi\mathrm{i}h_jk_j/M)\right|\\ &= \prod_{j=1}^{s}\left|\sum_{k_j=0}^{b_j-a_j}\exp(2\pi\mathrm{i}h_jk_j/M)\exp(2\pi\mathrm{i}h_ja_j/M)\right|\\ &= \prod_{j=1}^{s}\left|\sum_{k_j=0}^{b_j-a_j}\exp(2\pi\mathrm{i}h_jk_j/M)\right|.\end{aligned}$$

If $h_j = 0$, then

$$\left|\sum_{k_j=0}^{b_j-a_j}\exp(2\pi\mathrm{i}h_jk_j/M)\right| = b_j-a_j+1\le M = \frac{M}{r(h_j,M)}.$$

If now $h_j\in C^*(M)$, then

$$\begin{aligned}\left|\sum_{k_j=0}^{b_j-a_j}\exp(2\pi\mathrm{i}h_jk_j/M)\right| &= \left|\frac{\exp(2\pi\mathrm{i}h_j(b_j-a_j+1)/M)-1}{\exp(2\pi\mathrm{i}h_j/M)-1}\right|\\ &= \left|\frac{\sin(\pi h_j(b_j-a_j+1)/M)}{\sin(\pi h_j/M)}\right|\\ &\le \frac{1}{\sin(\pi|h_j|/M)} = \frac{M}{r(h_j,M)}.\end{aligned}$$

In any case,

$$r^*(\boldsymbol{h}, M) \le \prod_{j=1}^{s} \frac{M}{r(h_j, M)} = \frac{M^s}{r(\boldsymbol{h}, M)},$$

and therefore

$$\left|\frac{A(J,\mathcal{P},N)}{N} - \lambda_s(J)\right| \le 1 - \left(1 - \frac{1}{M}\right)^s$$
$$+ \sum_{\boldsymbol{h}\in C_s^*(M)} \frac{1}{r(\boldsymbol{h}, M)} \left|\frac{1}{N}\sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{y}_n / M)\right|.$$

The right-hand side of this inequality is independent of the specific choice of the interval $J$ and hence the result follows. □

We close this section with a formula for the star discrepancy of finite one-dimensional point sets. This formula was first proved in 1972 by H. Niederreiter.

**Proposition 2.31 (Niederreiter).** *Let* $\mathcal{P} = \{x_1, \ldots, x_N\}$ *be a point set in the unit interval* $[0, 1)$ *satisfying* $x_1 \le x_2 \le \cdots \le x_N$. *Then*

$$D_N^*(\mathcal{P}) = \frac{1}{2N} + \max_{n=1,2,\ldots,N} \left|x_n - \frac{2n-1}{2N}\right|.$$

*(We remark that the requirement* $x_1 \le x_2 \le \cdots \le x_N$ *does not impose any restriction on the point set.)*

▶ **Remark 2.32** Applying Proposition 2.31 to the regular lattice $\Gamma_{m,1} = \{\frac{2n-1}{2N} : n = 1, \ldots, N\}$ we find again that $D_N^*(\Gamma_{m,1}) = 1/(2N)$, and this is best possible for any $N$-element point set in dimension $s = 1$ (cf. Remark 2.20).

*Proof.* Put $x_0 := 0$ and $x_{N+1} := 1$. We have

$$D_N^*(\mathcal{P}) = \sup_{0<\alpha\le 1} \left|\frac{A([0,\alpha), N)}{N} - \alpha\right| = \max_{\substack{n=0,\ldots,N \\ x_n < x_{n+1}}} \sup_{x_n < \alpha \le x_{n+1}} \left|\frac{A([0,\alpha), N)}{N} - \alpha\right|$$
$$= \max_{\substack{n=0,\ldots,N \\ x_n < x_{n+1}}} \sup_{x_n < \alpha \le x_{n+1}} \left|\frac{n}{N} - \alpha\right|. \tag{2.13}$$

Assume that $x_n < x_{n+1}$. The function $g_n(x) = \left|\frac{n}{N} - x\right|$ is convex and hence it achieves its maximum on the compact interval $[x_n, x_{n+1}]$ in one of the end-points of this interval. Therefore,

$$D_N^*(\mathcal{P}) = \max_{\substack{n=0,\ldots,N \\ x_n < x_{n+1}}} \max\left(\left|\frac{n}{N} - x_n\right|, \left|\frac{n}{N} - x_{n+1}\right|\right). \tag{2.14}$$

Now we show that we may omit the condition $x_n < x_{n+1}$ in the first of the above maxima. Assume that

$$x_n < x_{n+1} = x_{n+2} = \ldots = x_{n+r} < x_{n+r+1},$$

with some $r \geq 2$. Hence the indices $n + j$ for $j = 1, \ldots, r - 1$ do not appear in the above maximum. We show that

$$\left|\frac{n+j}{N} - x_{n+j}\right| \quad \text{and} \quad \left|\frac{n+j}{N} - x_{n+j+1}\right|$$

for $1 \leq j \leq r - 1$ are always less or equal to some number which already appears in the first maximum in (2.14). Indeed, let $1 \leq j \leq r - 1$. Then

$$\begin{aligned}\left|\frac{n+j}{N} - x_{n+j}\right| = \left|\frac{n+j}{N} - x_{n+1}\right| &\leq \max\left(\left|\frac{n}{N} - x_{n+1}\right|, \left|\frac{n+r}{N} - x_{n+1}\right|\right) \\ &= \max\left(\left|\frac{n}{N} - x_{n+1}\right|, \left|\frac{n+r}{N} - x_{n+r}\right|\right).\end{aligned}$$

However, both terms in the latter maximum already appear in the first maximum in (2.14). The same can be shown for the term $\left|\frac{n+j}{N} - x_{n+j+1}\right|$. Hence,

$$\begin{aligned} D_N^*(\mathcal{P}) &= \max_{n=0,\ldots,N} \max\left(\left|\frac{n}{N} - x_n\right|, \left|\frac{n}{N} - x_{n+1}\right|\right) \\ &= \max_{n=1,\ldots,N} \max\left(\left|\frac{n}{N} - x_n\right|, \left|\frac{n-1}{N} - x_n\right|\right) \\ &= \max_{n=1,\ldots,N} \max\left(\left|\frac{2n-1}{2N} - x_n + \frac{1}{2N}\right|, \left|\frac{2n-1}{2N} - x_n - \frac{1}{2N}\right|\right) \\ &= \frac{1}{2N} + \max_{n=1,\ldots,N} \left|\frac{2n-1}{2N} - x_n\right|, \end{aligned}$$

since $\max(|a - b|, |a + b|) = |a| + |b|$. □

## 2.4 A Classical Construction: The Halton Sequence

In Sect. 2.1 we introduced the one-dimensional van der Corput sequence in base $b$ and showed its uniform distribution modulo one. This sequence is in fact the prototype of many uniformly distributed sequences and point sets with low star discrepancy, even in higher dimensions. In this section we introduce two

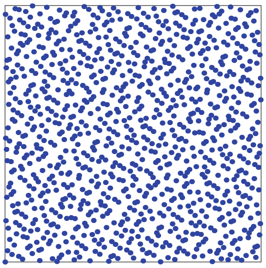

**Fig. 2.4** The first 1000 elements of the Halton sequence $\mathcal{S}_{2,3}$

generalizations of the van der Corput sequence which are already classical. The first is the infinite Halton sequence and the second is its finite version, the so-called Hammersley point set.

## The Halton Sequence

We construct an infinite sequence by component-wise concatenation of corresponding elements of van der Corput sequences in different bases. Recall that $\phi_b(n)$ denotes the $b$-adic radical inverse function as defined in Definition 2.8.

**Definition 2.33**

Let $s \in \mathbb{N}$ and let $b_1, \ldots, b_s \geq 2$ be integers. The *Halton sequence* in bases $b_1, \ldots, b_s$ is the sequence $\mathcal{S}_{b_1,\ldots,b_s} = (\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$ whose $n$th element is given by

$$\boldsymbol{x}_n := (\phi_{b_1}(n), \phi_{b_2}(n), \ldots, \phi_{b_s}(n)) \quad \text{for} \quad n \in \mathbb{N}_0.$$

For $s = 1$ we recover the van der Corput sequence $(\phi_b(n))_{n \in \mathbb{N}_0}$ in base $b = b_1$.

## Example 2.34

For example, in dimension $s = 2$ with bases $b_1 = 2$ and $b_2 = 3$, we have $\boldsymbol{x}_0 = (0, 0)$, $\boldsymbol{x}_1 = (1/2, 1/3)$, $\boldsymbol{x}_2 = (1/4, 2/3)$, $\boldsymbol{x}_3 = (3/4, 1/9)$, $\boldsymbol{x}_4 = (1/8, 4/9)$, .... Figure 2.4 shows the first 1000 elements of this sequence.

**Theorem 2.35.** *Let $s \in \mathbb{N}$, and let $b_1, \ldots, b_s \geq 2$ be pairwise coprime integers. For the star discrepancy of the Halton sequence $\mathcal{S}_{b_1,\ldots,b_s}$ in bases $b_1, \ldots, b_s$ we have*

$$D_N^*(\mathcal{S}_{b_1,\ldots,b_s}) \leq \left( \prod_{j=1}^{s} \frac{b_j \log(b_j N)}{\log b_j} \right) \frac{1}{N} \quad \textit{for} \quad N \in \mathbb{N}.$$

*Hence, asymptotically for $N \to \infty$ we have $D_N^*(\mathcal{S}_{b_1,\ldots,b_s}) = O((\log N)^s / N)$.*

For the proof of Theorem 2.35 we require the following lemmas:

**Lemma 2.36.** *Let $a, n, m \in \mathbb{Z}$. The congruence $x \equiv a \pmod{m}$ has exactly $\lfloor \frac{n}{m} \rfloor + \theta$ solutions $x \in \{0, \dots, n-1\}$, where $\theta$ is either 0 or 1.*

*Proof.* The result follows from the elementary fact that among any $m$ consecutive integers the congruence has exactly one solution. □

We also need the Chinese Remainder Theorem, which we recall here without proof.

**Lemma 2.37 (Chinese Remainder Theorem).** *Let $m_1, \dots, m_s$ be pairwise coprime integers. Then the system of congruences*

$$\begin{aligned} x &\equiv a_1 \pmod{m_1}, \\ x &\equiv a_2 \pmod{m_2}, \\ &\vdots \\ x &\equiv a_s \pmod{m_s}, \end{aligned}$$

*has exactly one solution modulo $m_1 m_2 \cdots m_s$.*

Now we prove Theorem 2.35 for dimensions $s = 1$ and $s = 2$. The general case follows by the same arguments. The proof presented here follows [46].

*Proof of Theorem 2.35.* For $N < \max_{j=1,2,\dots,s} b_j$ the bound on the right-hand side is always larger than 1 and therefore the assertion trivially holds true. Hence we may assume in the following that $N \geq \max_{j=1,2,\dots,s} b_j$.

**The case $s = 1$:** Let $b = b_1$. Let $\alpha \in (0, 1]$ with infinite $b$-adic expansion

$$\alpha = \frac{\alpha_0}{b} + \frac{\alpha_1}{b^2} + \frac{\alpha_2}{b^3} + \cdots .$$

If the $b$-adic expansion of $\alpha$ is finite, say $\alpha = (0, \alpha_0 \alpha_1 \dots \alpha_M)_b$ with $\alpha_M \neq 0$, then we switch to the infinite expansion by writing

$$\alpha = (0, \alpha_0 \dots \alpha_{M-1} \alpha'_M \alpha'_{M+1} \dots)_b$$

with $\alpha'_M = \alpha_M - 1$ and $\alpha'_i = b - 1$ for $i \geq M + 1$.
Let $k \in \{0, \dots, N-1\}$ with $b$-adic expansion

$$k = k_0 + k_1 b + k_2 b^2 + \cdots + k_m b^m,$$

where $k_m \neq 0$. Then $b^m \leq k < b^{m+1}$ and hence $m = \lfloor \frac{\log k}{\log b} \rfloor$. This means that $k$ has exactly $m + 1 = \lfloor \frac{\log k}{\log b} \rfloor + 1 \leq M + 1$ digits in its $b$-adic expansion, where $M = \lfloor \frac{\log N}{\log b} \rfloor$. Hence for $k \in \{0, \ldots, N-1\}$ we have

$$k = k_0 + k_1 b + \cdots + k_M b^M$$

with $b$-adic digits $k_0, k_1, \ldots, k_M \in \{0, \ldots, b-1\}$, and the corresponding value of the $b$-adic radical inverse function equals

$$\phi_b(k) = \frac{k_0}{b} + \frac{k_1}{b^2} + \cdots + \frac{k_M}{b^{M+1}}.$$

Let us determine the number

$$A([0,\alpha), \mathcal{S}_b, N) = \#\{k \in \{0, \ldots, N-1\} \colon \phi_b(k) \in [0,\alpha)\}$$

of indices $k \in \{0, \ldots, N-1\}$ for which $x_k = \phi_b(k)$ belongs to the interval $[0,\alpha)$. According to the definition of the radical inverse function $\phi_b(k)$, we have $\phi_b(k) < \alpha$ if and only if one of the following conditions is satisfied:

$$\begin{array}{rl}
(1) & k_0 < \alpha_0 \\
(2) & k_0 = \alpha_0, k_1 < \alpha_1 \\
(3) & k_0 = \alpha_0, k_1 = \alpha_1, k_2 < \alpha_2 \\
 & \vdots \\
(M{+}1) & k_0 = \alpha_0, \ldots, k_{M-1} = \alpha_{M-1}, k_M < \alpha_M \\
(M{+}2) & k_0 = \alpha_0, \ldots, k_{M-1} = \alpha_{M-1}, k_M = \alpha_M
\end{array}$$

Note that conditions (1), ..., $(M{+}2)$ are mutually exclusive. Now recast these $M + 2$ conditions as congruences in the following way:

$$\begin{array}{rl}
(1) & k \equiv k_0 \pmod{b},\ 0 \leq k_0 < \alpha_0 \\
(2) & k \equiv \alpha_0 + k_1 b \pmod{b^2},\ 0 \leq k_1 < \alpha_1 \\
(3) & k \equiv \alpha_0 + \alpha_1 b + k_2 b^2 \pmod{b^3},\ 0 \leq k_2 < \alpha_2 \\
 & \vdots \\
(M{+}1) & k \equiv \alpha_0 + \cdots + \alpha_{M-1} b^{M-1} + k_M b^M \pmod{b^{M+1}},\ 0 \leq k_M < \alpha_M \\
(M{+}2) & k \equiv \alpha_0 + \cdots + \alpha_M b^M \pmod{b^{M+2}}
\end{array}$$

By Lemma 2.36, the number of solutions of the respective congruences in $\{0, \ldots, N-1\}$ equals

$$\begin{array}{rl}
(1) & \alpha_0\left(\lfloor N/b\rfloor+\theta_0\right)\\
(2) & \alpha_1\left(\lfloor N/b^2\rfloor+\theta_1\right)\\
 & \vdots\\
(M{+}1) & \alpha_M\left(\lfloor N/b^{M+1}\rfloor+\theta_M\right)\\
(M{+}2) & \lfloor N/b^{M+2}\rfloor+\theta_{M+1}=\theta_{M+1}
\end{array}$$

with appropriate $\theta_j\in\{0,1\}$ for $j=0,\dots,M+1$. For the last line we mention that $b^M\le N<b^{M+1}$ implies $\lfloor N/b^{M+2}\rfloor=0$. Now we have

$$A([0,\alpha),\mathcal{S}_b,N)=\sum_{i=0}^{M}\alpha_i\left(\left\lfloor\frac{N}{b^{i+1}}\right\rfloor+\theta_i\right)+\theta_{M+1}$$

and

$$\begin{aligned}
\left|A([0,\alpha),\mathcal{S}_b,N)-N\sum_{i=0}^{M}\frac{\alpha_i}{b^{i+1}}\right| &= \left|\sum_{i=0}^{M}\alpha_i\left(\left\lfloor\frac{N}{b^{i+1}}\right\rfloor+\theta_i-\frac{N}{b^{i+1}}\right)+\theta_{M+1}\right|\\
&\le\sum_{i=0}^{M}\alpha_i\left|\frac{N}{b^{i+1}}-\left\lfloor\frac{N}{b^{i+1}}\right\rfloor-\theta_i\right|+1\\
&\le\sum_{i=0}^{M}\alpha_i+1\\
&\le(M+1)(b-1)+1,
\end{aligned}$$

since $\alpha_i\le b-1$. Therefore,

$$|A([0,\alpha),\mathcal{S}_b,N)-\alpha N|\le(M+1)(b-1)+1+N\sum_{i=M+1}^{\infty}\frac{\alpha_i}{b^{i+1}}.$$

Using again the fact that $\alpha_i\le b-1$, we obtain

$$N\sum_{i=M+1}^{\infty}\frac{\alpha_i}{b^{i+1}}\le N\frac{b-1}{b^{M+2}}\sum_{i=0}^{\infty}\frac{1}{b^i}=\frac{N}{b^{M+1}}<1, \tag{2.15}$$

and hence

$$|A([0,\alpha),\mathcal{S}_b,N)-\alpha N|\le(M+1)(b-1)+2.$$

For the final term we have

$$\begin{aligned}(M+1)(b-1)+2 &\leq \left(\frac{\log N}{\log b}+1\right)(b-1)+2\\ &= b\frac{\log(Nb)}{\log b}+2-\frac{\log(Nb)}{\log b}\\ &\leq b\frac{\log(Nb)}{\log b},\end{aligned} \tag{2.16}$$

since

$$2-\frac{\log(Nb)}{\log b}=\frac{\log b^2-\log(Nb)}{\log b}<0$$

for $N \geq b$. Thus we have shown that for all $\alpha \in (0,1]$ and all $N \geq b$

$$\left|\frac{A([0,\alpha),\mathcal{S}_b,N)}{N}-\alpha\right| \leq \frac{b}{N}\frac{\log(Nb)}{\log b}.$$

Since this bound is independent of the specific choice of $\alpha$ the result follows for $s=1$.

**The case** $s=2$: We write $b=b_1$ and $c=b_2$. Let $\alpha,\beta \in (0,1]$ with infinite digit expansions

$$\alpha=(0,\alpha_0\alpha_1\ldots)_b \quad \text{and} \quad \beta=(0,\beta_0\beta_1\ldots)_c,$$

in bases $b$ and $c$, respectively. Put

$$M:=\left\lfloor\frac{\log N}{\log b}\right\rfloor \quad \text{and} \quad L:=\left\lfloor\frac{\log N}{\log c}\right\rfloor.$$

The number $A([0,\alpha)\times[0,\beta),\mathcal{S}_{b,c},N)$ is the number of indices $k \in \{0,\ldots,N-1\}$ with

$$\phi_b(k)<\alpha \text{ and } \phi_c(k)<\beta.$$

Let

$$k=l_0+l_1c+\cdots+l_Lc^L,$$

with $0 \leq l_i \leq c-1$, be the $c$-adic expansion of $k$. Then we have $\phi_b(k)<\alpha$ and $\phi_c(k)<\beta$ if and only if, in addition to the conditions (1), (2), ..., ($M$+2) for the first component from the case $s=1$ also one of the following conditions for the second component is satisfied:

$$\begin{array}{rl}
(\overline{1}) & k \equiv l_0 \pmod{c},\ 0 \le l_0 < \beta_0 \\
(\overline{2}) & k \equiv \beta_0 + l_1 c \pmod{c^2},\ 0 \le l_1 < \beta_1 \\
(\overline{3}) & k \equiv \beta_0 + \beta_1 c + l_2 c^2 \pmod{c^3},\ 0 \le l_2 < \beta_2 \\
 & \vdots \\
(\overline{L+1}) & k \equiv \beta_0 + \cdots + \beta_{L-1} c^{L-1} + l_L c^L \pmod{c^{L+1}},\ 0 \le l_L < \beta_L \\
(\overline{L+2}) & k \equiv \beta_0 + \cdots + \beta_L c^L \pmod{c^{L+2}}
\end{array}$$

Recall that by our global assumption we have $\gcd(b, c) = 1$ and thus also $\gcd(b^m, c^l) = 1$ for $m, l \in \mathbb{N}$. According to Lemmas 2.36 and 2.37, the number of $k \in \{0, \ldots, N-1\}$ which satisfy the congruences $(m)$ and $(\overline{l})$ for

- $m \in \{1, \ldots, M+1\}$ and $l \in \{1, \ldots, L+1\}$ equals
$$\alpha_{m-1}\beta_{l-1}\left(\left\lfloor \frac{N}{b^m c^l} \right\rfloor + \theta_{m-1,l-1}\right),$$
where $\theta_{m-1,l-1} \in \{0, 1\}$,
- For $m \in \{1, \ldots, M+1\}$ and $l = L+2$ equals
$$\alpha_{m-1}\theta_{m-1},$$
where $\theta_{m-1} \in \{0, 1\}$,
- For $m = M+2$ and $l \in \{1, \ldots, L+1\}$ equals
$$\beta_{l-1}\theta_{l-1},$$
where $\theta_{l-1} \in \{0, 1\}$, and
- For $m = M+2$ and $l = L+2$ equals $\theta_{M+1,L+1} \in \{0, 1\}$.

Therefore,

$$\begin{aligned}
A([0, \alpha) \times [0, \beta), \mathcal{S}_{b,c}, N) = \sum_{m=1}^{M+1} \sum_{l=1}^{L+1} \alpha_{m-1}\beta_{l-1}\left(\left\lfloor \frac{N}{b^m c^l} \right\rfloor + \theta_{m-1,l-1}\right) + \\
+ \sum_{m=1}^{M+1} \alpha_{m-1}\theta_{m-1} + \sum_{l=1}^{L+1} \beta_{l-1}\theta_{l-1} + \theta_{M+1,L+1}.
\end{aligned}$$

We obtain

$$\begin{aligned}
&\left| A([0, \alpha) \times [0, \beta), \mathcal{S}_{b,c}, N) - \sum_{m=1}^{M+1} \sum_{l=1}^{L+1} \alpha_{m-1}\beta_{l-1} \frac{N}{b^m c^l} \right| \\
&\le \sum_{m=1}^{M+1} \sum_{l=1}^{L+1} \alpha_{m-1}\beta_{l-1} + \sum_{m=1}^{M+1} \alpha_{m-1} + \sum_{l=1}^{L+1} \beta_{l-1} + 1
\end{aligned}$$

$$= \left(\sum_{m=1}^{M+1} \alpha_{m-1} + 1\right)\left(\sum_{l=1}^{L+1} \beta_{l-1} + 1\right),$$

and further, using (2.15) and (2.16),

$$\begin{aligned}
&|A([0,\alpha) \times [0,\beta), \mathcal{S}_{b,c}, N) - \alpha\beta N| \\
&\leq \left(\sum_{m=1}^{M+1} \alpha_{m-1} + 1\right)\left(\sum_{l=1}^{L+1} \beta_{l-1} + 1\right) \\
&\quad +N \sum_{m=M+2}^{\infty} \frac{\alpha_{m-1}}{b^m} + N \sum_{l=L+2}^{\infty} \frac{\beta_{l-1}}{c^l} + N \sum_{m=M+2}^{\infty} \sum_{l=L+2}^{\infty} \frac{\alpha_{m-1}}{b^m}\frac{\beta_{l-1}}{c^l} \\
&\leq ((b-1)(M+1)+1)((c-1)(L+1)+1)+3 \\
&\leq ((b-1)(M+1)+2)((c-1)(L+1)+2) \\
&\leq \frac{b\log(bN)}{\log b}\frac{c\log(cN)}{\log c},
\end{aligned}$$

since $N \geq b$ and $N \geq c$. This proves also the case $s = 2$ and indicates how the proof works for arbitrary dimensions $s \geq 3$. □

Theorem 2.35 shows that the star discrepancy of the Halton sequence is asymptotically of the order of magnitude $(\log N)^s/N$ for $N \to \infty$ and this is the best we can show. In the course of time it has become customary to speak of *low-discrepancy sequences* in case where their star discrepancy is of order of magnitude $(\log N)^s/N$. Later, we will become acquainted with other examples of low-discrepancy sequences. Recall from Sect. 2.3 that many people conjecture, though without proof, that $(\log N)^s/N$ is the best that one achieve at all for the star discrepancy. In dimension $s = 1$ the Halton result shows that Schmidt's lower bound from Theorem 2.27 is exact.

On the other hand, the bound from Theorem 2.35 has a slightly defective appearance. To explain this, we consider the quantity

$$d^*(\mathcal{S}) := \limsup_{N\to\infty} \frac{ND_N^*(\mathcal{S})}{(\log N)^s} \tag{2.17}$$

for Halton sequences $\mathcal{S}_{b_1,\ldots,b_s}$. From Theorem 2.35 it follows that

$$d^*(\mathcal{S}_{b_1,\ldots,b_s}) \leq \prod_{j=1}^{s} \frac{b_j}{\log b_j} =: c_s.$$

This bound $c_s$ is large and grows very fast to infinity as the dimension $s$ grows. For example, assume that $b_1,\ldots,b_s$ are the first $s$ prime numbers, which is the "smallest" choice in order to keep the bases pairwise coprime. The prime number theorem implies that $b_j \approx j \log j$ for large $j$ and hence

$$c_s = \prod_{j=1}^{s} \frac{b_j}{\log b_j} \approx \prod_{j=1}^{s} \frac{j \log j}{\log(j \log j)} \approx \prod_{j=1}^{s} \frac{j \log j}{\log j} = s!.$$

In 2004, Atanassov [5] was able to overcome this particular disadvantage of the bound in Theorem 2.35. We state his result without proof here:

**Theorem 2.38 (Atanassov).** *Let $s \in \mathbb{N}$, and let $b_1, \ldots, b_s \geq 2$ be pairwise coprime integers. Then for the star discrepancy of the Halton sequence $\mathcal{S}_{b_1,\ldots,b_s}$ in bases $b_1, \ldots, b_s$ for $N \geq 2$ we have*

$$D_N^*(\mathcal{S}_{b_1,\ldots,b_s})$$
$$\leq \left[\frac{1}{s!}\prod_{j=1}^{s}\left(\frac{\lfloor b_j/2\rfloor \log N}{\log b_j} + s\right) + \sum_{k=0}^{s-1}\frac{b_{k+1}}{k!}\prod_{j=1}^{k}\left(\frac{\lfloor b_j/2\rfloor \log N}{\log b_j} + k\right)\right]\frac{1}{N}.$$

This result implies that

$$d^*(\mathcal{S}_{b_1,\ldots,b_s}) \leq \frac{1}{s!}\prod_{j=1}^{s}\frac{\lfloor b_j/2\rfloor}{\log b_j}.$$

If again $b_1, \ldots, b_s$ are the first $s$ prime numbers, then it can be shown that $\frac{1}{s!}\prod_{j=1}^{s}\frac{\lfloor b_j/2\rfloor}{\log b_j} \leq \frac{7}{2^s s}$, and hence

$$\limsup_{s\to\infty}\frac{\log d^*(\mathcal{S}_{b_1,\ldots,b_s})}{s} \leq -\log 2. \tag{2.18}$$

This shows that $d^*(\mathcal{S}_{b_1,\ldots,b_s})$ tends to zero at an exponential rate as $s \to \infty$.

The discrepancy bounds for the Halton sequence suggest to choose the bases $b_1, \ldots, b_s$ as small as possible such that they are still pairwise coprime. Hence, the best we can do is to choose them to be the first $s$ prime numbers. However, in higher dimensions already this "minimal" choice leads to the problem that the initial segments of the Halton sequence, or, more exact, projections to lower dimensional faces thereof, have very poor distribution properties. For many practical applications such a behavior is not desirable. Just as an example, consider the projection of a Halton sequence in dimension $s \geq 36$ with prime bases $b_1, b_2, \ldots, b_s$ to the components with numbers 35 and 36, that is, with $b_{35} = 149$ and $b_{36} = 151$. Then most points of the initial segment are concentrated around the main diagonal and very large areas of the unit square contain no point; see Fig. 2.5 where we

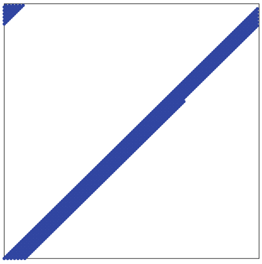

**Fig. 2.5** The projection of the first 1000 elements of the Halton sequence to the coordinates 35 and 36

plotted the first 1000 elements. Similar pictures appear when we consider other two-dimensional projections to "high" coordinates. There are two ways to overcome this problem, at least to a certain extent: the first one is the concept of *generalised Halton sequences*, which uses $s$ sequences of permutations $(\pi_{j,k})_{k\in\mathbb{N}_0}$ for $j = 1, 2, \ldots, s$, where $\pi_{j,k}$ is a permutation of $\{0, 1, \ldots, b_j - 1\}$ for each $k \in \mathbb{N}_0$. Then the $j$th component of the $n$th element of a generalised Halton sequence in bases $b_1, \ldots, b_s$ is given by

$$x_{n,j} = \sum_{k=0}^{\infty} \frac{\pi_{j,k}(n_{j,k})}{b_j^{k+1}} \quad \text{whenever} \quad n = n_{j,0} + n_{j,1}b_j + n_{j,2}b_j^2 + \cdots \tag{2.19}$$

for $j = 1, 2, \ldots, s$ and $n \in \mathbb{N}_0$. A second possibility is a generalisation of the Halton sequence to *digital* $(t, s)$*-sequences* (see Chap. 5) which use for all $s$ components the same fixed base $b$ in the digital construction. However, also in this case there might appear some defects in lower dimensional projections.

## The Hammersley Point Set

We use the infinite Halton sequence to construct for every $N \in \mathbb{N}$, $N \geq 2$, an $N$-element point set with low star discrepancy.

**Definition 2.39**

Let $s \in \mathbb{N}$, and let $b_1, \ldots, b_{s-1} \geq 2$ be pairwise coprime integers. The $N$-element *Hammersley point set* $\mathcal{H}_{N,b_1,\ldots,b_{s-1}}$ in bases $b_1, \ldots, b_{s-1}$ is the finite point set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ in $[0, 1)^s$ where

$$\boldsymbol{x}_n := \left(\frac{n}{N}, \phi_{b_1}(n), \ldots, \phi_{b_{s-1}}(n)\right) \quad \text{for} \quad n = 0, \ldots, N - 1.$$

## Example 2.40

For example, in dimension $s = 2$ with $N = 8$ and $b = 2$ the Hammersley point set $\mathcal{H}_{8,2}$ consists of the elements $\boldsymbol{x}_0 = (0, 0)$, $\boldsymbol{x}_1 = (1/8, 1/2)$, $\boldsymbol{x}_2 = (2/8, 1/4)$, $\boldsymbol{x}_3 = (3/8, 3/4)$, $\boldsymbol{x}_4 = (4/8, 1/8)$,

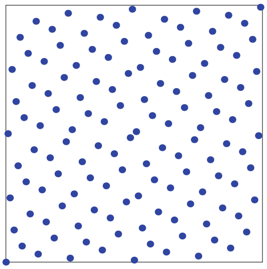

**Fig. 2.6** The 128-element Hammersley point set $\mathcal{H}_{128,2}$ in base 2

$\boldsymbol{x}_5 = (5/8, 5/8)$, $\boldsymbol{x}_6 = (6/8, 3/8)$ and $\boldsymbol{x}_7 = (7/8, 7/8)$. A further example is illustrated in Fig. 2.6.

**Theorem 2.41.** *For the star discrepancy of the $N$-element Hammersley point set $\mathcal{H}_{N,b_1,\ldots,b_{s-1}}$ in $[0,1)^s$ with pairwise coprime bases $b_1, \ldots, b_{s-1}$ we have*

$$D_N^*(\mathcal{H}_{N,b_1,\ldots,b_{s-1}}) \le \frac{1}{N}\left(\prod_{j=1}^{s-1} \frac{b_j \log(b_j N)}{\log b_j} + 1\right).$$

Hence, the Hammersley point set in pairwise coprime bases satisfies

$$D_N^*(\mathcal{H}_{N,b_1,\ldots,b_{s-1}}) = O\left(\frac{(\log N)^{s-1}}{N}\right).$$

Note that this is a substantial improvement compared to the discrepancy of the regular lattice, which is of the order of magnitude $N^{-1/s}$ only. It has become customary to speak of *low-discrepancy point sets* in the case where their star discrepancy is of order of magnitude $(\log N)^{s-1}/N$ in dimension $s$, with $N \ge 2$ being the cardinality of the considered point set.

For the proof of Theorem 2.41 we use the following general principle which is usually attributed to Roth.

**Lemma 2.42.** *Let $s \in \mathbb{N}$, $s \ge 2$, and let $\mathcal{S} = (\boldsymbol{y}_n)_{n\in\mathbb{N}_0}$ be an infinite sequence in $[0,1)^{s-1}$ with star discrepancy $D_M^*(\mathcal{S})$ for $M \in \mathbb{N}$. Let $N \in \mathbb{N}$ and let $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$, where $\boldsymbol{x}_n := \left(\frac{n}{N}, \boldsymbol{y}_n\right)$ for $n = 0, 1, \ldots, N-1$. Then*

$$D_N^*(\mathcal{P}) \le \frac{1}{N}\left(\max_{M=1,2,\ldots,N} M D_M^*(\mathcal{S}) + 1\right).$$

*Proof.* Consider an interval of the form $J := \prod_{j=1}^{s}[0, \alpha_j) \subseteq [0, 1)^s$. An element $\boldsymbol{x}_n$ from $\mathcal{P}$ belongs to $J$ if and only if

$$0 \leq \frac{n}{N} < \alpha_1 \quad \text{and} \quad \boldsymbol{y}_n \in \prod_{j=2}^{s}[0, \alpha_j) =: \tilde{J}.$$

Now we have $A(J, \mathcal{P}, N) = A(\tilde{J}, \mathcal{S}, M)$, where $M = \lceil N\alpha_1 \rceil$. Consequently,

$$\begin{aligned} &|A(J, \mathcal{P}, N) - N\lambda_s(J)| \\ &\leq |A(\tilde{J}, \mathcal{S}, M) - M\lambda_{s-1}(\tilde{J})| + |M\lambda_{s-1}(\tilde{J}) - N\lambda_s(J)|. \end{aligned}$$

Since $|M\lambda_{s-1}(\tilde{J}) - N\lambda_s(J)| = |(\lceil N\alpha_1 \rceil - N\alpha_1) \prod_{j=2}^{s} \alpha_j| \leq 1$, it follows that

$$\begin{aligned} |A(J, \mathcal{P}, N) - N\lambda_s(J)| &\leq |A(\tilde{J}, \mathcal{S}, M) - M\lambda_{s-1}(\tilde{J})| + 1 \\ &\leq MD_M^*(\mathcal{S}) + 1 \\ &\leq \max_{M=1,2,\ldots,N} MD_M^*(\mathcal{S}) + 1, \end{aligned}$$

and the result follows. □

*Proof of Theorem 2.41.* Let $\mathcal{S}_{b_1,\ldots,b_{s-1}}$ be an $(s-1)$-dimensional Halton sequence in pairwise coprime bases $b_1, \ldots, b_{s-1}$. Then we obtain from Theorem 2.35

$$MD_M^*(\mathcal{S}_{b_1,\ldots,b_{s-1}}) \leq \prod_{j=1}^{s-1} \frac{b_j \log(b_j M)}{\log b_j},$$

whence

$$\max_{M=1,2,\ldots,N} MD_M^*(\mathcal{S}_{b_1,\ldots,b_{s-1}}) \leq \prod_{j=1}^{s-1} \frac{b_j \log(b_j N)}{\log b_j}.$$

Now the result follows from Lemma 2.42. □

## 2.5 Further Reading and Exercises

### Further Reading

The theory of Uniform Distribution Modulo One was initiated by the seminal work of Weyl [95]. A standard reference is the book of Kuipers and Niederreiter [56]. Furthermore, the book of Matoušek [65] offers a readable introduction to discrepancy theory. Also the books of Beck and Chen [7], Dick and Pillichshammer [18], Drmota and Tichy [25], and Hlawka [43] can be warmly recommended. The survey article [8] of Bilyk and Lacey gives an overview on recent results concerning the sharp order of magnitude for the smallest possible value of star discrepancy and the survey [19] discusses constructions of point sets with best order of $L_2$ discrepancy. A proof of Proposition 2.22 can be found in [25] and of the inequality of Erdős-Turán-Koksma (Theorem 2.28) in [25] or in [56] (for $s = 1$). The current best value for $c_s$ follows from Cochrane [14]. The proof of Theorem 2.30 follows the exposition in [69]. More information on the Halton sequence can be found in the books of Hua and Wang [46], Niederreiter [69], and Dick and Pillichshammer [18]. The currently best asymptotic estimate for the star discrepancy of the Halton sequence was shown by Atanassov [5]. The presented version of this result (Theorem 2.38) is due to [18, Theorem 3.36]. The problem of defects in lower dimensional projections of Halton sequences and other sequences based on digit expansions is discussed, e.g., by Lemieux [61] and Schmid [84].

Known approaches and algorithms to compute discrepancy measures of point sets are surveyed in the book chapter [24] by Doerr, Gnewuch and Wahlström. Information concerning applications of discrepancy theory to a wide variety of topics in computer science can be found in the book by Chazelle [13].

## Exercises

2.1 Let $\mathcal{S}$ be a sequence in $[0,1)^s$. Provide a Lebesgue integrable function $f : [0,1] \to \mathbb{R}$ for which (2.3) does not hold.

2.2 Show that an infinite sequence $\mathcal{S}$ in $[0,1)^s$ is uniformly distributed modulo one if and only if for every closed interval of the form $[\boldsymbol{a}, \boldsymbol{b}] \subseteq [0,1)^s$ we have

$$\lim_{N\to\infty} \frac{A([\boldsymbol{a},\boldsymbol{b}],\mathcal{S},N)}{N} = \lambda_s([\boldsymbol{a},\boldsymbol{b}]),$$

where for $\boldsymbol{a} = (a_1,\dots,a_s)$, $\boldsymbol{b} = (b_1,\dots,b_s)$ we write $[\boldsymbol{a},\boldsymbol{b}] = [a_1,b_1] \times \cdots \times [a_s,b_s]$. An analogous assertion can be shown for open intervals $(\boldsymbol{a},\boldsymbol{b})$.

2.3 Give a detailed proof of Theorem 2.4.

2.4 Show that a necessary condition for the uniform distribution of $(\{n\boldsymbol{\alpha}\})_{n\in\mathbb{N}_0}$, where $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_s) \in \mathbb{R}^s$, is that $1, \alpha_1, \dots, \alpha_s$ are linearly independent over the rationals.

2.5 Let $b, c \geq 2$ be integers satisfying $\gcd(b,c) = 1$. Show directly, i.e., without the knowledge of the Halton result, that the two-dimensional sequence given by $\boldsymbol{x}_n = (\phi_b(n), \phi_c(n))$ for $n \in \mathbb{N}_0$ is uniformly distributed modulo one.

2.6 Determine both the $L_1$ discrepancy and the star discrepancy of the following point sets:
(a) $\mathcal{P} = \{n/N : n = 0,\dots,N-1\}$.
(b) $\mathcal{P} = \{n/(2N) : n = 0,\dots,N-1\}$.

2.7 Show the so-called *Warnock formula*

$$(L_{2,N}(\mathcal{P}))^2 = \frac{1}{3^s} - \frac{2}{N}\sum_{n=0}^{N-1}\prod_{j=1}^{s}\frac{1-x_{n,j}^2}{2} + \frac{1}{N^2}\sum_{n,m=0}^{N-1}\prod_{j=1}^{s}\min(1-x_{m,j}, 1-x_{n,j})$$

for the squared $L_2$ discrepancy of a point set $\mathcal{P} = \{\boldsymbol{x}_0,\dots,\boldsymbol{x}_{N-1}\}$ in $[0,1)^s$, where $x_{n,j}$ denotes the $j$th component of the point $\boldsymbol{x}_n$.

2.8 Determine the $L_2$ discrepancy of the point sets from Exercise 2.6.

2.9 Determine the $L_1$ discrepancy and the $L_2$ discrepancy of the one-dimensional centered regular lattice $\Gamma_{N,1} = \{(2n+1)/(2N) : n = 0,\dots,N-1\}$.

2.10 Compute $\int_{[0,1]^{sN}} [L_{2,N}(\{\boldsymbol{t}_1,\dots,\boldsymbol{t}_N\})]^2 \,\mathrm{d}\boldsymbol{t}_1 \dots \mathrm{d}\boldsymbol{t}_N$ and interpret the result.

2.11 Let $\mathcal{T}_{s,N}(\alpha)$ be the set of all $N$-tuples $(\boldsymbol{t}_1,\dots,\boldsymbol{t}_N) \in [0,1)^{sN}$ for which

$$L_{2,N}(\{\boldsymbol{t}_1,\ldots,\boldsymbol{t}_N\}) \le \frac{\alpha}{\sqrt{N}}\left(\frac{1}{2^s}-\frac{1}{3^s}\right)^{1/2}.$$

Show that for all $\alpha \ge 1$ we have $\lambda_{sN}(\mathcal{T}_{s,N}(\alpha)) > 1-\alpha^{-2}$.

2.12 For $s, m \in \mathbb{N}$, $m \ge 2$, let

$$\Gamma_{m,s}^* = \left\{\left(\frac{n_1}{m},\ldots,\frac{n_s}{m}\right) \,:\, n_1,\ldots,n_s \in \{0,\ldots,m-1\}\right\}.$$

(a) Show that $D_N^*(\Gamma_{m,s}^*) = 1-(1-1/m)^s$ and deduce $D_N^*(\Gamma_{m,s}^*) \asymp_s N^{-1/s}$.

(b) Determine the $L_2$ discrepancy $L_{2,N}(\Gamma_{m,s}^*)$ and deduce from the obtained formula that $L_{2,N}(\Gamma_{m,s}^*) \asymp_s N^{-1/s}$.

2.13 Let $m_1,\ldots,m_s \ge 2$ be integers and let

$$\Gamma_{m_1,\ldots,m_s} = \left\{\left(\frac{k_1}{m_1},\ldots,\frac{k_s}{m_s}\right) \,:\, k_j \in \{0,\ldots,m_j-1\} \text{ for } j=1,\ldots,s\right\}$$

be a regular lattice consisting of $N = m_1\cdots m_s$ elements in $[0,1)^s$. Show that

$$D_N^*(\Gamma_{m_1,\ldots,m_s}) = 1-\prod_{j=1}^{s}\left(1-\frac{1}{m_j}\right).$$

2.14 For $i = 1,\ldots,k$ let $\mathcal{P}_i$ be an $N_i$-element point set with discrepancy $D_{N_i}(\mathcal{P}_i)$. Let $\mathcal{P}$ be a superposition of the point sets $\mathcal{P}_i$, $i = 1,\ldots,k$ and put $N = N_1+\cdots+N_k$. Show the so-called *triangle inequality for the discrepancy*:

$$D_N(\mathcal{P}) \le \sum_{i=1}^{k}\frac{N_i}{N}D_{N_i}(\mathcal{P}_i),$$

and similarly for the star discrepancy.

2.15 We prove Theorem 2.25: The proof is based on that of Theorem 2.24. Choose $n \in \mathbb{N}$ such that $2^{n-1} < 2N \le 2^n$. For $i = 0,\ldots,n$ define the functions $f_i : [0,1)^s \to \{-1,0,1\}$ as in (2.8). Furthermore, for $\alpha \in (0,1/2)$ define the auxiliary function $H : [0,1)^2 \to \mathbb{R}$ by

$$H(\boldsymbol{x}) = \prod_{i=0}^{n}(1+\alpha f_i(\boldsymbol{x}))-1.$$

Put $D(\boldsymbol{x}) = Nx_1x_2 - A([\boldsymbol{0},\boldsymbol{x}),\mathcal{P},N)$, where $\boldsymbol{x} = (x_1,x_2)$. We split the proof into several steps:

(a) Show the following lemma:

**Lemma 2.43.** *For any set of integers* $\{i_1,\ldots,i_k\}$ *satisfying* $0 \le i_1 < \cdots < i_k \le n$ *we have* $\int_{[0,1]^2} f_{i_1}(\boldsymbol{x})\cdots f_{i_k}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$.

*Hint:* Partition the square $[0,1)^2$ into $2^{n-i_1+i_k}$ rectangles by dividing $[0,1)$ on the $x$-axes into $2^{i_k}$ and on the $y$-axes into $2^{n-i_1}$ intervals of the same length. Then argue as in the case $k=2$.

(b) Show that

$$\prod_{i=0}^{n}(1+\alpha f_i(\boldsymbol{x})) = 1+\alpha F(\boldsymbol{x}) + \sum_{k=2}^{n+1}\alpha^k F_k(\boldsymbol{x}),$$

where the function $F$ is defined as in (2.9) and, for $k = 2, \ldots, n+1$,

$$F_k(\boldsymbol{x}) = \sum_{0 \le i_1 < \ldots < i_k \le n} f_{i_1}(\boldsymbol{x}) \cdots f_{i_k}(\boldsymbol{x}).$$

(c) Prove that $\int_{[0,1]^2} |H(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} \le 2$. *Hint:* From $\alpha < 1/2$ we obtain $|H(\boldsymbol{x})| \le \prod_{i=0}^{n}(1+\alpha f_i(\boldsymbol{x}))+1$. Now use Lemma 2.43.

(d) Show the following lemma:

**Lemma 2.44.** *For any set of integers* $\{i_1, \ldots, i_k\}$ *satisfying* $0 \le i_1 < \cdots < i_k \le n$ *we have*

$$\left| \int_{[0,1]^2} f_{i_1}(\boldsymbol{x}) \cdots f_{i_k}(\boldsymbol{x}) D(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \le N 2^{-n+i_1-i_k-4}.$$

*Hint:* Let $R$ be one of the $2^{n-i_1+i_k}$ rectangles from the proof of Lemma 2.43. Show that $|\int f_{i_1} \cdots f_{i_k} D \,\mathrm{d}\boldsymbol{x}| = (\lambda_2(R)/4)^2$. Proceed as in the special case $k=1$ in the proof of Theorem 2.24.

(e) Show the following lemma:

**Lemma 2.45.** *For every* $k = 2, \ldots, n+1$ *we have*

$$\left| \int_{[0,1]^2} F_k(\boldsymbol{x}) D(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \le \sum_{i=0}^{n-k+1} \sum_{h=1}^{n-i} \frac{N}{2^{n+h+4}} \binom{h-1}{k-2}.$$

*Hint:* Use Lemma 2.44 and put $i_1 = i \in \{0, \ldots, n-k+1\}$ and $i_k = i+h$, where $h \in \{1, \ldots, n-i\}$.

(f) Show that $\left| \sum_{k=2}^{n+1} \alpha^k \int_{[0,1]^2} F_k(\boldsymbol{x}) D(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \le \alpha^2 N \frac{n}{2^{n+2}}$.

(g) Show that for sufficiently small $\alpha \in (0, 1/2)$ there exists a constant $C > 0$ such that $\left| \int_{[0,1]^2} H(\boldsymbol{x}) D(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \ge C \log N$. *Hint:* From the triangle inequality it follows that $\left|\int HD\right| \ge \alpha \left|\int FD\right| - \left|\sum_{k=2}^{n+1} \alpha^k \int F_k D\right|$.

(h) Deduce from Exercises 2.15c and 2.15g that

$$N D_N^*(\mathcal{P}) = \sup_{\boldsymbol{x} \in [0,1]^2} |D(\boldsymbol{x})| > c \log N$$

for some positive constant $c$.

2.16 Deduce Theorem 2.27 from Theorem 2.25. *Hint:* Use Lemma 2.42.

2.17 Let $\mathcal{S}_{b_1,\ldots,b_s}^{\Pi_1,\ldots,\Pi_s}$ be the generalised Halton sequence in pairwise coprime bases $b_1, \ldots, b_s$ and with $s$ sequences of permutations $\Pi_j = (\pi_{j,k})_{k \in \mathbb{N}_0}$ of $\{0, 1, \ldots, b_j - 1\}$ with $\pi_{j,k}(0) = 0$ for $j = 1, 2, \ldots, s$ and $k \in \mathbb{N}_0$, as defined in (2.19). Show that the star discrepancy of $\mathcal{S}_{b_1,\ldots,b_s}^{\Pi_1,\ldots,\Pi_s}$ satisfies the bound from Theorem 2.35,

$$D_N^*(\mathcal{S}_{b_1,\ldots,b_s}^{\Pi_1,\ldots,\Pi_s}) \le \left( \prod_{j=1}^{s} \frac{b_j \log(b_j N)}{\log b_j} \right) \frac{1}{N} \quad \text{for all} \quad N \in \mathbb{N}.$$

# 3 QMC Integration in Reproducing Kernel Hilbert Spaces

We return to the problem of numerical integration of multivariate functions. As already mentioned in Sect. 1.1, we normalize the integration domain to be the compact unit cube $[0,1]^s$, and hence the integrals considered are of the form (1.1). We aim at approximating such integrals by QMC rules of the form $Q_{N,s}(f) = \frac{1}{N}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n)$ with fixed integration nodes $\boldsymbol{x}_0,\ldots,\boldsymbol{x}_{N-1}$ taken from $[0,1)^s$, i.e.,

$$\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx Q_{N,s}(f).$$

On first sight this approach looks quite simple, but the crux of this method is the choice of underlying nodes. On the other hand, as already mentioned, the knowledge of the integration nodes is insufficient for solving the integration problem in full generality. It is easy to construct two functions $f, g : [0,1]^s \to \mathbb{R}$ for which we have $f(\boldsymbol{x}_n) = g(\boldsymbol{x}_n)$ for all $n = 0,1,\ldots,N-1$, but $\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{[0,1]^s} g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ can be any number. This means that we require some global information on the functions to be integrated. To tackle this problem we consider function classes with certain smoothness properties and study the worst-case integration error of QMC rules therein. Based on this error analysis, we find criteria for point sets to be used as underlying nodes of QMC rules for the specific function class.

## 3.1 Univariate QMC Integration

To begin with, we consider QMC integration of univariate real valued functions $f : [0,1] \to \mathbb{R}$ with continuous first derivative on $[0,1]$. For such functions we obtain from the fundamental theorem of calculus that for any $x \in [0,1]$

$$f(x) = f(1) - \int_x^1 f'(y)\,\mathrm{d}y.$$

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_3,

For the error

$$e(f,\mathcal{P}) = \int_0^1 f(x)\,\mathrm{d}x - \frac{1}{N}\sum_{n=0}^{N-1} f(x_n)$$

of a QMC rule based on sample nodes $\mathcal{P} = \{x_0,\ldots,x_{N-1}\}$ from $[0,1)$ we now have

$$\begin{aligned}
e(f,\mathcal{P}) &= \frac{1}{N}\sum_{n=0}^{N-1}\int_{x_n}^1 f'(y)\,\mathrm{d}y - \int_0^1\int_x^1 f'(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_0^1 \frac{1}{N}\sum_{n=0}^{N-1}\chi_{(x_n,1]}(y) f'(y)\,\mathrm{d}y - \int_0^1\int_0^y f'(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^1 f'(y)\left[\frac{1}{N}\sum_{n=0}^{N-1}\chi_{(x_n,1]}(y) - y\right]\mathrm{d}y.
\end{aligned}$$

Since

$$\sum_{n=0}^{N-1}\chi_{(x_n,1]}(y) = \sum_{n=0}^{N-1}\chi_{[0,y)}(x_n) = A([0,y),\mathcal{P},N)$$

is the number of indices $n \in \{0,\ldots,N-1\}$ for which $x_n$ belongs to $[0,y)$, we find that

$$\frac{1}{N}\sum_{n=0}^{N-1}\chi_{(x_n,1]}(y) - y = \Delta_{\mathcal{P},N}(y),$$

i.e., the local discrepancy of $\mathcal{P}$ in $y$. Hence,

$$e(f,\mathcal{P}) = \int_0^1 f'(y)\Delta_{\mathcal{P},N}(y)\,\mathrm{d}y. \tag{3.1}$$

Taking the absolute value and applying first the triangle inequality for integrals and then the Hölder inequality, we get

$$\begin{aligned}
|e(f,\mathcal{P})| &\le \int_0^1 |f'(y)||\Delta_{\mathcal{P},N}(y)|\,\mathrm{d}y \\
&\le \left(\int_0^1 |f'(y)|^q\,\mathrm{d}y\right)^{1/q}\left(\int_0^1 |\Delta_{\mathcal{P},N}(y)|^p\,\mathrm{d}y\right)^{1/p} \\
&= \|f'\|_{L_q}\|\Delta_{\mathcal{P},N}\|_{L_p},
\end{aligned} \tag{3.2}$$

where, as usual, $p,q \in [0,\infty]$ with $\frac{1}{p}+\frac{1}{q} = 1$.

The estimate in (3.2) separates the influence of the integrand $f$ and of the underlying point set $\mathcal{P}$ on the absolute integration error. Hence, for QMC integration of functions $f$ for which $\|f'\|_{L_q} < \infty$, one should choose sample nodes $\mathcal{P}$ with low $L_p$ discrepancy $L_{p,N}(\mathcal{P}) = \|\Delta_{\mathcal{P},N}\|_{L_p}$.

Note that $\|f'\|_{L_q}$ is just a semi-norm of the function $f$, while

$$\|f\|_{1,q} := \left(|f(1)|^q + \int_0^1 |f'(y)|^q \,\mathrm{d}y\right)^{1/q} \tag{3.3}$$

is a proper norm. For this reason, (3.2) is often stated in the form

$$|e(f,\mathcal{P})| \le \|f\|_{1,q}\, L_{p,N}(\mathcal{P}) \quad \text{for} \quad p,q \ge 1 \quad \text{and} \quad \frac{1}{p} + \frac{1}{q} = 1.$$

For $p = \infty$ and $q = 1$ this is a simplified version of the inequality of Koksma.

**Theorem 3.1 (Koksma inequality).** *For any $f \in C^1([0,1])$ and any $N$-element point set $\mathcal{P}$ in $[0,1)$ we have*

$$|e(f,\mathcal{P})| \le \|f\|_{1,1} D_N^*(\mathcal{P}).$$

Now we aim at developing a similar theory for multivariate functions. This can be done in a very elegant way by using the notion of reproducing kernel Hilbert space.

## 3.2 Reproducing Kernel Hilbert Spaces

For an integrable function $f$ defined on $[0,1]^s$ and a point set $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ in $[0,1)^s$ we use the notation

$$e(f,\mathcal{P}) = \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \frac{1}{N}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n),$$

in full analogy to the univariate case.

We want to introduce the notion of the worst-case error of a QMC rule in a Hilbert space of functions. We denote the inner product in a Hilbert space $\mathcal{H}$ by $\langle\cdot,\cdot\rangle$ and the corresponding norm by $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$.

**Definition 3.2**

The *worst-case error* of a QMC rule based on a point set $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ in $[0,1)^s$ in a Hilbert space $\mathcal{H}$ of integrable functions on $[0,1]^s$ is defined as

$$e(\mathcal{H},\mathcal{P}) = \sup_{f\in\mathcal{H}, \|f\|\le 1} |e(f,\mathcal{P})|.$$

One can develop a particularly nice error theory by using a reproducing kernel for $\mathcal{H}$.

**Definition 3.3**

A Hilbert space $\mathcal{H}$ of functions on $[0,1]^s$ is called a *reproducing kernel Hilbert space* on $[0,1]^s$ if there exists a function $K : [0,1]^s \times [0,1]^s \to \mathbb{C}$ such that

**K1:** $K(\cdot,\boldsymbol{y}) \in \mathcal{H}$ for all $\boldsymbol{y} \in [0,1]^s$ and
**K2:** $\langle f, K(\cdot,\boldsymbol{y})\rangle = f(\boldsymbol{y})$ for all $\boldsymbol{y} \in [0,1]^s$ and for all $f \in \mathcal{H}$.

The function $K$ is called the *reproducing kernel* of $\mathcal{H}$.

Note that in **K1** and **K2** we consider the reproducing kernel $K$ as a function of the first variable, denoted by $\cdot$, and in $\langle f, K(\cdot,y)\rangle$ we apply the inner product with respect to the first variable of $K$. The second property **K2** is the so-called *reproducing property*, i.e., the evaluation of the function $f$ can be expressed as the inner product of the function and the kernel function.

A function which satisfies the properties **K1** and **K2** is automatically symmetric, uniquely defined and positive semi-definite:

**K3 (symmetry):** $K(\boldsymbol{x},\boldsymbol{y}) = \overline{K(\boldsymbol{y},\boldsymbol{x})}$ for all $\boldsymbol{x},\boldsymbol{y} \in [0,1]^s$;
**K4 (uniqueness):** for any function $\tilde{K}$ satisfying **K1** and **K2** we have $\tilde{K} = K$;
**K5 (positive semi-definiteness):** for any choice of $a_0,\ldots,a_{N-1} \in \mathbb{C}$ and $\boldsymbol{x}_0,\ldots,\boldsymbol{x}_{N-1} \in [0,1]^s$ we have $\sum_{m,n=0}^{N-1} \overline{a}_m a_n K(\boldsymbol{x}_m,\boldsymbol{x}_n) \ge 0$.

*Proof.* **K3:** $K(\boldsymbol{x},\boldsymbol{y}) = \langle K(\cdot,\boldsymbol{y}), K(\cdot,\boldsymbol{x})\rangle = \overline{\langle K(\cdot,\boldsymbol{x}), K(\cdot,\boldsymbol{y})\rangle} = \overline{K(\boldsymbol{y},\boldsymbol{x})}$,
**K4:** $\tilde{K}(\boldsymbol{x},\boldsymbol{y}) = \langle \tilde{K}(\cdot,\boldsymbol{y}), K(\cdot,\boldsymbol{x})\rangle = \overline{\langle K(\cdot,\boldsymbol{x}), \tilde{K}(\cdot,\boldsymbol{y})\rangle} = \overline{K(\boldsymbol{y},\boldsymbol{x})} = K(\boldsymbol{x},\boldsymbol{y})$,
**K5:**

$$\begin{aligned}\sum_{m,n=0}^{N-1} \overline{a}_m a_n K(\boldsymbol{x}_m,\boldsymbol{x}_n) &= \sum_{m,n=0}^{N-1} \overline{a}_m a_n \langle K(\cdot,\boldsymbol{x}_n), K(\cdot,\boldsymbol{x}_m)\rangle \\ &= \left\langle \sum_{n=0}^{N-1} a_n K(\cdot,\boldsymbol{x}_n), \sum_{m=0}^{N-1} a_m K(\cdot,\boldsymbol{x}_m)\right\rangle \\ &= \left\| \sum_{m=0}^{N-1} a_m K(\cdot,\boldsymbol{x}_m)\right\|^2 \ge 0.\end{aligned}$$

□

Conversely, one can show that a function $K$ satisfying properties **K3** and **K5** uniquely determines a Hilbert space of functions for which **K1** and **K2** hold (and hence also **K4**). Therefore, it makes sense to speak of a reproducing kernel without explicitly specifying the corresponding Hilbert space of functions.

Now we revisit Sect. 3.1 and consider the one-dimensional integration problem in the light of reproducing kernel Hilbert spaces.

For $f, g \in C^1([0,1])$ define an inner product by

$$\langle f, g\rangle_1 := f(1)g(1) + \int_0^1 f'(x)g'(x)\,\mathrm{d}x. \tag{3.4}$$

The corresponding norm $\|f\|_{1,2} := \sqrt{\langle f, f\rangle_1}$ is exactly the norm considered in (3.3) for $q = 2$. We have $\|f\|_{1,2} < \infty$ whenever the first derivative $f'$ is in $L_2([0,1])$. Based on this norm we define a Hilbert space $\mathcal{H}_1$ by

$$\mathcal{H}_1 = \{f : [0,1] \to \mathbb{R} : f \text{ absolutely continuous and } \|f\|_{1,2} < \infty\}.$$

Recall that a function $f : [0,1] \to \mathbb{R}$ is called *absolutely continuous*, if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $n \in \mathbb{N}$ and any choice of pairwise disjoint intervals $(x_k, y_k)$, $k = 1, \ldots, n$, in $[0,1]$ with $\sum_{k=1}^n (y_k - x_k) < \delta$ we have $\sum_{k=1}^n |f(x_k) - f(y_k)| < \varepsilon$. For example, any Lipschitz continuous function is absolutely continuous. In particular, if $f : [0,1] \to \mathbb{R}$ is differentiable on $[0,1]$ with bounded first derivative $f'$, then $f$ is absolutely continuous.

The definition of an absolutely continuous function is a bit unwieldy, but there is a well-known equivalent condition that is more suited to our purpose: one can show that a function $f : [0,1] \to \mathbb{R}$ is absolutely continuous, if and only if there exists a Lebesgue integrable function $g : [0,1] \to \mathbb{R}$ such that $f$ can be written in the form

$$f(x) = f(0) + \int_0^x g(t)\,\mathrm{d}t = f(1) - \int_x^1 g(t)\,\mathrm{d}t \quad \text{for all} \quad x \in [0,1].$$

In this case $f$ is almost everywhere differentiable and $g = f'$ almost everywhere.

Now we show that the function $K_1$ defined by

$$K_1(x, y) = 1 + \min(1 - x, 1 - y) \tag{3.5}$$

is the reproducing kernel of $\mathcal{H}_1$. For fixed $y \in [0,1]$ we have

$$K_1(x, y) = \begin{cases} 2 - x & \text{if } x > y, \\ 2 - y & \text{if } x \le y, \end{cases}$$

and

$$\frac{\partial K_1(x, y)}{\partial x} = \begin{cases} -1 & \text{if } x > y, \\ 0 & \text{if } x \le y. \end{cases}$$

Thus with $g : [0,1] \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} -1 & \text{if } x > y, \\ 0 & \text{if } x \le y, \end{cases}$$

we obtain

$$K_1(1,y) - \int_x^1 g(t)\,\mathrm{d}t = 1 + \int_{\max(x,y)}^1 \mathrm{d}t = 1 + \min(1-x, 1-y) = K_1(x,y).$$

Therefore, the function $K_1(\cdot, y)$ is absolutely continuous. In particular, $K_1(\cdot, y)$ is also integrable. Furthermore,

$$\begin{aligned}\langle K_1(\cdot,y), K_1(\cdot,y)\rangle_1 &= K_1(1,y)^2 + \int_0^1 \left(\frac{\partial K_1(x,y)}{\partial x}\right)^2 \mathrm{d}x \\ &= 1 + \int_y^1 \mathrm{d}x = 2 - y < \infty,\end{aligned}$$

and hence $\|K_1(\cdot, y)\|_{1,2} < \infty$. This implies that $K_1(\cdot, y) \in \mathcal{H}_1$. It remains to check the reproducing property of $K_1$. We have

$$\begin{aligned}\langle f, K_1(\cdot,y)\rangle_1 &= f(1)K_1(1,y) + \int_0^1 f'(x)\frac{\partial K_1(x,y)}{\partial x}\,\mathrm{d}x \\ &= f(1) - \int_y^1 f'(x)\,\mathrm{d}x = f(y).\end{aligned}$$

Altogether, we have shown that $K_1$ is the reproducing kernel of the space $\mathcal{H}_1$.

Let us consider QMC integration for the Hilbert space $\mathcal{H}_1$. Using the reproducing kernel $K_1$ we can write the integration functional as

$$\begin{aligned}\int_0^1 f(y)\,\mathrm{d}y &= \int_0^1 \langle f, K_1(\cdot,y)\rangle_1\,\mathrm{d}y \\ &= \int_0^1 \left(f(1)K_1(1,y) + \int_0^1 f'(x)\frac{\partial K_1(x,y)}{\partial x}\,\mathrm{d}x\right)\mathrm{d}y \\ &= f(1)\int_0^1 K_1(1,y)\,\mathrm{d}y + \int_0^1 f'(x)\int_0^1 \frac{\partial K_1(x,y)}{\partial x}\,\mathrm{d}y\ \mathrm{d}x \\ &= f(1)\int_0^1 K_1(1,y)\,\mathrm{d}y + \int_0^1 f'(x)\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_0^1 K_1(x,y)\,\mathrm{d}y\right)\mathrm{d}x \\ &= \left\langle f, \int_0^1 K_1(\cdot,y)\,\mathrm{d}y\right\rangle_1,\end{aligned}$$

where we used Fubini's theorem to change the order of integration, and where the interchange of integration and differentiation can be justified by direct calculation of the terms $\int_0^1 \frac{\partial K_1(x,y)}{\partial x}\,\mathrm{d}y$ and $\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_0^1 K_1(x,y)\,\mathrm{d}y\right)$. In a similar manner we can write the QMC functional $Q_{N,1}(f)$ as

$$Q_{N,1}(f) = \frac{1}{N}\sum_{n=0}^{N-1} f(x_n) = \frac{1}{N}\sum_{n=0}^{N-1}\langle f, K_1(\cdot, x_n)\rangle_1 = \left\langle f, \frac{1}{N}\sum_{n=0}^{N-1} K_1(\cdot, x_n)\right\rangle_1 .$$

Put

$$h(x) = \int_0^1 K_1(x,y)\,\mathrm{d}y - \frac{1}{N}\sum_{n=0}^{N-1} K_1(x, x_n).$$

Note that $h \in \mathcal{H}_1$, since $K_1(\cdot, y) \in \mathcal{H}_1$ as well as $\int_0^1 K_1(\cdot, y)\,\mathrm{d}y \in \mathcal{H}_1$. Now we have

$$e(f, \mathcal{P}) = \int_0^1 f(y)\,\mathrm{d}y - Q_{N,1}(f) = \langle f, h\rangle_1$$

and hence, using the Cauchy-Schwarz inequality,

$$|e(f, \mathcal{P})| = |\langle f, h\rangle_1| \le \|f\|_{1,2}\,\|h\|_{1,2}\,. \tag{3.6}$$

In (3.6) we have equality if $f(x) = h(x)$.

Due to the linearity of the inner product we obtain for $f \in \mathcal{H}_1$ with $\|f\|_{1,2} \neq 0$ that

$$\frac{|e(f,\mathcal{P})|}{\|f\|_{1,2}} = |e(f/\|f\|_{1,2}, \mathcal{P})| \le \|h\|_{1,2} = \frac{|\langle h, h\rangle_1|}{\|h\|_{1,2}} = \frac{|e(h,\mathcal{P})|}{\|h\|_{1,2}},$$

with equality if $f = h$. Consequently,

$$e(\mathcal{H}_1, \mathcal{P}) = \frac{|e(h,\mathcal{P})|}{\|h\|_{1,2}} = \frac{|\langle h, h\rangle_1|}{\|h\|_{1,2}} = \|h\|_{1,2}.$$

Observe that for $x \notin \mathcal{P}$

$$\Delta_{\mathcal{P},N}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\int_0^1 K_1(x,y)\,\mathrm{d}y - \frac{1}{N}\sum_{n=0}^{N-1} K_1(x, x_n)\right) = h'(x)$$

(a proof of this fact is left as Exercise 3.8), and hence

$$L_{2,N}(\mathcal{P}) = \|h\|_{1,2}\,.$$

This implies that

$$e(\mathcal{H}_1, \mathcal{P}) = L_{2,N}(\mathcal{P}), \tag{3.7}$$

i.e., the worst-case error is exactly the $L_2$ discrepancy of the underlying point set. Furthermore, among all functions from the unit ball of $\mathcal{H}_1$ the function $h/\|h\|_{1,2}$ is the hardest to integrate (using the point set $\mathcal{P}$), since for $h$ we have equality in (3.6).

## 3.3 The Worst-Case Error in Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of integrable functions $f : [0,1]^s \to \mathbb{C}$, equipped with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$.

Recall from the general theory of normed spaces that a linear functional $T$ on $\mathcal{H}$ is called *bounded* if there exists $M < \infty$ such that $|T(f)| \le M$ for all $f$ satisfying $\|f\| \le 1$, and that for linear functionals boundedness is equivalent to continuity.

Recall further from Definition 3.2 that the worst-case error of a QMC rule based on a point set $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ in $[0,1)^s$ in $\mathcal{H}$ is defined as

$$e(\mathcal{H}, \mathcal{P}) = \sup_{f \in \mathcal{H}, \|f\| \le 1} |e(f, \mathcal{P})|,$$

where

$$e(f, \mathcal{P}) = \int_{[0,1]^s} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n)$$

for $f \in \mathcal{H}$. A priori it is not clear that $e(\mathcal{H}, \mathcal{P})$ is finite, or, in other words, that the linear functional $e(\cdot, \mathcal{P})$ is continuous. In the following we shall find conditions under which this is the case.

For $\boldsymbol{y} \in [0,1]^s$ consider the linear functional $T_{\boldsymbol{y}}$ which evaluates $f \in \mathcal{H}$ at $\boldsymbol{y}$, i.e.,

$$T_{\boldsymbol{y}}(f) = f(\boldsymbol{y}) \quad \text{for } f \in \mathcal{H}.$$

$T_{\boldsymbol{y}}$ is called the *evaluation functional* in $\boldsymbol{y}$. If $T_{\boldsymbol{y}}$ is continuous for all $\boldsymbol{y} \in [0,1]^s$, then so is every QMC rule.

It turns out that continuity of the evaluation functionals is equivalent to the existence of a reproducing kernel. For the proof of this remarkable fact we use a well-known result for continuous linear functionals from functional analysis, the Fréchet-Riesz representation theorem.

**Theorem 3.4 (Fréchet-Riesz Representation Theorem).** *Let $\mathcal{X}$ be a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$ and let $T : \mathcal{X} \to \mathbb{C}$ be a continuous linear functional. Then there exists exactly one element $z \in \mathcal{X}$ with the property that*

$$T(x) = \langle x, z \rangle \ \textit{for all} \ x \in \mathcal{X}.$$

**Theorem 3.5.** *Let $\mathcal{H}$ be a Hilbert space of functions on $[0,1]^s$. Then $\mathcal{H}$ is a reproducing kernel Hilbert space on $[0,1]^s$ if and only if the evaluation functionals*

$$T_{\boldsymbol{y}}(f) = f(\boldsymbol{y}) \text{ for } f \in \mathcal{H},\ \boldsymbol{y} \in [0,1]^s$$

*are continuous.*

*Proof.* If the evaluation functionals are continuous, then the Fréchet-Riesz representation theorem guarantees, for every $\boldsymbol{y}$, the existence of a uniquely determined function $k_{\boldsymbol{y}} \in \mathcal{H}$ with the property that

$$T_{\boldsymbol{y}}(f) = \langle f, k_{\boldsymbol{y}} \rangle \quad \text{for all} \quad f \in \mathcal{H}.$$

If we now define $K(\boldsymbol{x},\boldsymbol{y}) = k_{\boldsymbol{y}}(\boldsymbol{x})$ for $\boldsymbol{x},\boldsymbol{y} \in [0,1]^s$, then $K$ satisfies the properties **K1** and **K2** and thus $\mathcal{H}$ is a reproducing kernel Hilbert space with reproducing kernel $K$.

Conversely, assume that $K$ is a reproducing kernel for $\mathcal{H}$ and let $\boldsymbol{y} \in [0,1]^s$. Using the Cauchy-Schwarz inequality, we get for every $f \in \mathcal{H}$

$$|T_{\boldsymbol{y}}(f)| = |f(\boldsymbol{y})| = |\langle f, K(\cdot,\boldsymbol{y})\rangle| \le \|f\|\,\|K(\cdot,\boldsymbol{y})\|.$$

From the reproducing property we get $\|K(\cdot,\boldsymbol{y})\|^2 = \langle K(\cdot,\boldsymbol{y}), K(\cdot,\boldsymbol{y})\rangle = K(\boldsymbol{y},\boldsymbol{y})$, so that $|T_{\boldsymbol{y}}(f)| \le M$ for every $f$ with $\|f\| \le 1$, where $M = \sqrt{K(\boldsymbol{y},\boldsymbol{y})}$. That means that $T_{\boldsymbol{y}}$ is continuous. □

Next consider the integration functional $I(f) = \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. If $\mathcal{H}$ has a reproducing kernel $K$, then for any $f \in \mathcal{H}$ with $\|f\| \le 1$

$$\left|\int_{[0,1]^s} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\right| = \left|\int_{[0,1]^s} T_{\boldsymbol{y}}(f)\,\mathrm{d}\boldsymbol{y}\right| \le \int_{[0,1]^s} |T_{\boldsymbol{y}}(f)|\,\mathrm{d}\boldsymbol{y} \le \int_{[0,1]^s} \sqrt{K(\boldsymbol{y},\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y}\,.$$

Hence if the kernel $K$ satisfies the condition

**C:** $\int_{[0,1]^s} \sqrt{K(\boldsymbol{y},\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y} < \infty$,

then the integration functional $I$ is continuous.

We can summarize: if $\mathcal{H}$ has a reproducing kernel $K$ which satisfies condition **C**, then function evaluation and integration are continuous linear functionals, and so is $e(\cdot,\mathcal{P})$ for any point set $\mathcal{P}$. Moreover, under these conditions $e(\mathcal{H},\mathcal{P})$ is a well-defined finite number.

► **Remark 3.6** After studying a couple of examples of reproducing kernel Hilbert spaces, the property that the evaluation functionals are continuous becomes increasingly natural. So one may justly ask the question whether there exist any examples at all of Hilbert spaces of functions for which function evaluation is not continuous.

Such an example can indeed be constructed, however, this example makes use of the fact that any linear subspace of a vector space has an algebraic complement. Thus the example is rather non-constructive and we may state informally that all “natural” examples of Hilbert spaces of functions have continuous evaluation functionals.

The example cited above also serves as an example of a Hilbert space of functions where the integration functional is defined on the whole space, but is not continuous.

► **Remark 3.7** Condition **C** is not strictly necessary for continuity of the integration functional. In fact, under our global assumption that we have a Hilbert space of integrable functions, continuity of the evaluation functionals is sufficient for continuity of the integration functional. The proof for this uses the closed graph theorem and can be found in [76, Section 23.4]. In the notes on the same section one can find an example of a reproducing kernel Hilbert space for which $K(\cdot, \boldsymbol{y})$ is integrable for every $\boldsymbol{y}$, but which contains functions that are not integrable.

In the last section we used the important property that for the specific kernel $K_1(x, y) = 1 + \min(1 - x, 1 - y)$ one can interchange the order of integration and the inner product, i.e.,

$$\int_0^1 \langle f, K_1(\cdot, y)\rangle_1 \,\mathrm{d}y = \left\langle f, \int_0^1 K_1(\cdot, y)\,\mathrm{d}y \right\rangle_1 . \tag{3.8}$$

There, this interchange reduced to a change of the order of integration and of the order of integration and differentiation. Property (3.8) is essential for our error analysis, and in the following we show that it holds for any reproducing kernel Hilbert space for which the integration functional is continuous.

**Lemma 3.8.** *Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel $K$ and inner product $\langle \cdot, \cdot \rangle$. Assume that the mapping*

$$I(f) = \int_{[0,1]^s} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \quad \textit{for } f \in \mathcal{H}$$

*is a continuous linear functional on $\mathcal{H}$. Then*

$$\int_{[0,1]^s} \langle f, K(\cdot, \boldsymbol{y})\rangle \,\mathrm{d}\boldsymbol{y} = \left\langle f, \int_{[0,1]^s} K(\cdot, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \right\rangle.$$

*Proof.* Since $I$ is continuous, the Fréchet-Riesz representation theorem guarantees the existence of a unique function $R \in \mathcal{H}$, such that

$$\int_{[0,1]^s} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = I(f) = \langle f, R\rangle \quad \text{for all } f \in \mathcal{H}.$$

Since $R \in \mathcal{H}$, the evaluation of $R$ at some $\boldsymbol{x}$ can be expressed in terms of the inner product with the kernel function $K(\cdot, \boldsymbol{x})$, and hence

$$R(\boldsymbol{x}) = \langle R, K(\cdot, \boldsymbol{x}) \rangle = \overline{\langle K(\cdot, \boldsymbol{x}), R \rangle} = \overline{\int_{[0,1]^s} K(\boldsymbol{y}, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{y}}.$$

Thus we have

$$\begin{aligned}\int_{[0,1]^s} \langle f, K(\cdot, \boldsymbol{y}) \rangle \,\mathrm{d}\boldsymbol{y} &= \int_{[0,1]^s} f(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = \langle f, R \rangle \\ &= \left\langle f, \overline{\int_{[0,1]^s} K(\boldsymbol{y}, \cdot) \,\mathrm{d}\boldsymbol{y}} \right\rangle = \left\langle f, \int_{[0,1]^s} K(\cdot, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \right\rangle.\end{aligned}$$

□

From now on we will always tacitly assume that $K$ satisfies condition **C**.

With this assumption we can proceed in our worst-case error analysis as in the special case from the previous section. We have

$$I(f) = \int_{[0,1]^s} \langle f, K(\cdot, \boldsymbol{y}) \rangle \,\mathrm{d}\boldsymbol{y} = \left\langle f, \int_{[0,1]^s} K(\cdot, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \right\rangle$$

and

$$Q_{N,s}(f) = \frac{1}{N} \sum_{n=0}^{N-1} \langle f, K(\cdot, \boldsymbol{x}_n) \rangle = \left\langle f, \frac{1}{N} \sum_{n=0}^{N-1} K(\cdot, \boldsymbol{x}_n) \right\rangle.$$

Therefore, the integration error of the QMC rule $Q_{N,s}$ in $\mathcal{H}$ can again be expressed as an inner product

$$e(f, \mathcal{P}) = \langle f, h \rangle, \tag{3.9}$$

where

$$h(\boldsymbol{x}) = \int_{[0,1]^s} K(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} - \frac{1}{N} \sum_{n=0}^{N-1} K(\boldsymbol{x}, \boldsymbol{x}_n).$$

This function is often called the *representer* of the integration error. Taking the absolute value and applying the Cauchy-Schwarz inequality leads to

$$|e(f, \mathcal{P})| \leq \|f\| \, \|h\|.$$

Moreover, it follows from (3.9) that among all functions in the unit ball of $\mathcal{H}$ the normalized representer $h/\|h\|$ is the hardest to integrate. Therefore, the worst-case error can be written as

$$e(\mathcal{H}, \mathcal{P}) = \|h\|.$$

From this result we obtain for the squared worst-case error the formula $e^2(\mathcal{H}, \mathcal{P}) = \langle h, h \rangle$, which is very convenient for the calculation of the worst-case error of QMC rules in specific reproducing kernel Hilbert spaces.

**Theorem 3.9.** *Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel $K$ that satisfies condition* **C** *and let $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ be an $N$-element point set in $[0,1)^s$. Then*

$$e^2(\mathcal{H}, \mathcal{P}) = \int_{[0,1]^s} \int_{[0,1]^s} K(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} - \frac{2}{N} \sum_{n=0}^{N-1} \int_{[0,1]^s} K(\boldsymbol{x}_n, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y}$$
$$+ \frac{1}{N^2} \sum_{n,m=0}^{N-1} K(\boldsymbol{x}_n, \boldsymbol{x}_m).$$

In the next section we apply the theory developed here to show some classical results for the QMC integration error.

## 3.4 The Koksma-Hlawka Inequality

The Koksma-Hlawka inequality is the fundamental error estimate for QMC rules which separates the influence of the integrand $f$ and of the underlying integration nodes on the integration error. It was proved by J.F. Koksma in 1942 for dimension $s = 1$ and later, in 1961, generalized by E. Hlawka to arbitrary dimensions $s \in \mathbb{N}$. The Koksma-Hlawka inequality is a very general error estimate valid for all functions of finite variation in the sense of Hardy and Krause, which reduces to the total variation in dimension $s = 1$.

We consider the reproducing kernel $K_s : [0,1]^s \times [0,1]^s \to \mathbb{R}$ given by

$$K_s(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{s} K_1(x_j, y_j) = \prod_{j=1}^{s} (1 + \min(1 - x_j, 1 - y_j)), \tag{3.10}$$

where $\boldsymbol{x} = (x_1, \ldots, x_s) \in [0,1]^s$, $\boldsymbol{y} = (y_1, \ldots, y_s) \in [0,1]^s$, and $K_1$ is defined as in (3.5). The corresponding inner product is given by

$$\langle f, g \rangle_s = \sum_{\mathfrak{u} \subseteq [s]} \int_{[0,1]^{|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} f}{\partial \boldsymbol{x}_{\mathfrak{u}}}(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{1}) \frac{\partial^{|\mathfrak{u}|} g}{\partial \boldsymbol{x}_{\mathfrak{u}}}(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{1}) \, \mathrm{d}\boldsymbol{x}_{\mathfrak{u}}, \tag{3.11}$$

where $[s] = \{1, \ldots, s\}$. For $\mathfrak{u} \subseteq [s]$ and $\boldsymbol{x} = (x_1, \ldots, x_s) \in [0,1]^s$ we put $(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{1}) = (z_1, \ldots, z_s)$, where

$$z_j = \begin{cases} x_j & \text{if } j \in \mathfrak{u}, \\ 1 & \text{if } j \notin \mathfrak{u}. \end{cases}$$

Furthermore $\partial^{|\mathfrak{u}|}/\partial \boldsymbol{x}_{\mathfrak{u}}$ denotes the mixed first partial derivative with respect to the components of $\boldsymbol{x}$ whose index belongs to $\mathfrak{u}$.

Now let $\mathcal{H}_s$ be the reproducing kernel Hilbert space with reproducing kernel $K_s$ and norm $\|\cdot\|_{s,2} = \langle \cdot, \cdot \rangle_s^{1/2}$. We call $\mathcal{H}_s$ an *anchored Sobolev space* with anchor in 1.

This function space contains all functions $f$ on $[0,1]^s$ whose mixed partial derivatives $\partial^{|\mathfrak{u}|} f/\partial \boldsymbol{x}_{\mathfrak{u}}$ up to order one in each variable belong to $L_2([0,1]^s)$ and that are expressible in the form

$$\begin{aligned} f(\boldsymbol{x}) &= \langle f, K_s(\cdot, \boldsymbol{x}) \rangle_s \\ &= \sum_{\mathfrak{u} \subseteq [s]} (-1)^{|\mathfrak{u}|} \int_{[0,1]^{|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} f}{\partial \boldsymbol{y}_{\mathfrak{u}}}(\boldsymbol{y}_{\mathfrak{u}}, \mathbf{1}) \chi_{[\mathbf{0}, (\boldsymbol{y}_{\mathfrak{u}}, \mathbf{1}))}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \quad \text{for} \quad \boldsymbol{x} \in [0,1]^s. \end{aligned}$$

In Sect. 3.2 we already considered the one-dimensional case $\mathcal{H}_1$ and learned that it consists of all absolutely continuous functions $f$ with square integrable first derivative. For $s > 1$ one can show that $\mathcal{H}_s$ is the $s$-fold tensor product of the spaces $\mathcal{H}_1$, i.e.,

$$\mathcal{H}_s = \underbrace{\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_1}_{s\text{-fold}} = \operatorname{clos\,span} \left\{ \boldsymbol{x} \mapsto \prod_{j=1}^{s} f_j(x_j) \ : \ f_j \in \mathcal{H}_1 \right\}$$

where $\boldsymbol{x} = (x_1, \ldots, x_s)$ and where the closure clos is taken with respect to the norm $\|\cdot\|_{s,2}$.

We have

$$\int_{[0,1]^s} \sqrt{K_s(\boldsymbol{y}, \boldsymbol{y})} \, \mathrm{d}\boldsymbol{y} = \left( \int_0^1 \sqrt{2 - y} \, \mathrm{d}y \right)^s = \left( \frac{2(\sqrt{8} - 1)}{3} \right)^s < \infty,$$

and so $K_s$ satisfies condition **C**. Thus it follows from (3.9), that

$$e(f, \mathcal{P}) = \langle f, h \rangle_s$$

for all $f \in \mathcal{H}_s$ and all $N$-element point sets $\mathcal{P}$ in $[0,1)^s$, where $h$ is the representer of the integration error in $\mathcal{H}_s$. We have

$$\begin{aligned}h(\boldsymbol{x}) &= \int_{[0,1]^s} K_s(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} - \frac{1}{N}\sum_{n=0}^{N-1} K_s(\boldsymbol{x},\boldsymbol{x}_n)\\ &= \prod_{j=1}^{s}\left[\int_0^1 (1+\min(1-x_j,1-y_j))\,\mathrm{d}y_j\right]\\ &\qquad - \frac{1}{N}\sum_{n=0}^{N-1}\prod_{j=1}^{s}(1+\min(1-x_j,1-x_{n,j}))\\ &= \prod_{j=1}^{s}\frac{3-x_j^2}{2} - \frac{1}{N}\sum_{n=0}^{N-1}\prod_{j=1}^{s}(1+\min(1-x_j,1-x_{n,j})).\end{aligned}$$

For $\mathfrak{u} \subseteq [s]$ we have

$$\begin{aligned}\frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} h(\boldsymbol{x}_{\mathfrak{u}},\boldsymbol{1}) &= (-1)^{|\mathfrak{u}|}\left(\prod_{j\in\mathfrak{u}} x_j - \frac{1}{N}\sum_{n=0}^{N-1}\prod_{j\in\mathfrak{u}}\chi_{[0,x_j)}(x_{n,j})\right)\\ &= (-1)^{|\mathfrak{u}|}\left(\prod_{j\in\mathfrak{u}} x_j - \frac{1}{N}\sum_{n=0}^{N-1}\chi_{[\boldsymbol{0},(\boldsymbol{x}_{\mathfrak{u}},\boldsymbol{1}))}(\boldsymbol{x}_n)\right)\\ &= (-1)^{|\mathfrak{u}|+1}\Delta_{\mathcal{P},N}(\boldsymbol{x}_{\mathfrak{u}},\boldsymbol{1}),\end{aligned}$$

where $\boldsymbol{x}_n = (x_{n,1},\ldots,x_{n,s})$ and where $\Delta_{\mathcal{P},N}$ denotes the local discrepancy of the $N$-element point set $\mathcal{P}$. Note that $\Delta_{\mathcal{P},N}(\boldsymbol{x}_\emptyset,\boldsymbol{1}) = \Delta_{\mathcal{P},N}(\boldsymbol{1}) = 0$.

From these considerations we obtain a formula for the integration error in $\mathcal{H}_s$ which is known as *Hlawka's identity* or as *Zaremba's identity*.

**Theorem 3.10 (Hlawka's identity, Zaremba's identity).** *For $f \in \mathcal{H}_s$ and $\mathcal{P}$ in $[0,1)^s$ we have*

$$e(f,\mathcal{P}) = \sum_{\emptyset\neq\mathfrak{u}\subseteq[s]} (-1)^{|\mathfrak{u}|}\int_{[0,1]^{|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} f}{\partial \boldsymbol{x}_{\mathfrak{u}}}(\boldsymbol{x}_{\mathfrak{u}},\boldsymbol{1})\Delta_{\mathcal{P},N}(\boldsymbol{x}_{\mathfrak{u}},\boldsymbol{1})\,\mathrm{d}\boldsymbol{x}_{\mathfrak{u}}.$$

Taking absolute values on both sides of the equality in Theorem 3.10 and using the estimate

$$|\Delta_{\mathcal{P},N}(\boldsymbol{x}_{\mathfrak{u}},\boldsymbol{1})| \le \sup_{\boldsymbol{x}\in[0,1]^s}|\Delta_{\mathcal{P},N}(\boldsymbol{x})| = D_N^*(\mathcal{P}),$$

we immediately end up with the desired special case of the Koksma-Hlawka inequality. For a function $f : [0,1]^s \to \mathbb{R}$ for which all partial mixed derivatives up to order one in each variable are continuous on $[0,1]^s$ we define the norm

$$\|f\|_{s,1}=\sum_{\mathfrak{u}\subseteq[s]}\int_{[0,1]^{|\mathfrak{u}|}}\left|\frac{\partial^{|\mathfrak{u}|}f}{\partial\boldsymbol{x}_{\mathfrak{u}}}(\boldsymbol{x}_{\mathfrak{u}},\mathbf{1})\right|\,\mathrm{d}\boldsymbol{x}_{\mathfrak{u}}.$$

**Theorem 3.11 (Koksma-Hlawka inequality).** *Let $\mathcal{P}$ be an $N$-element point set in $[0,1)^s$. Then for all functions $f$ on $[0,1]^s$ with $\|f\|_{s,1}<\infty$ we have*

$$|e(f,\mathcal{P})|\le\|f\|_{s,1}D_N^*(\mathcal{P}).$$

This fundamental error estimate separates the influence of the function $f$ and of the integration nodes $\mathcal{P}$ on the integration error. The Koksma-Hlawka inequality in its original, more general form states that $|e(f,\mathcal{P})|\le V(f)D_N^*(\mathcal{P})$, where $V(f)$ denotes the variation in the sense of Hardy and Krause. We do not give the definition of the latter notion here, but we mention that if all partial mixed derivatives of $f$ up to order one in each variable are continuous on $[0,1]^s$, then

$$V(f)=\sum_{\emptyset\neq\mathfrak{u}\subseteq[s]}\int_{[0,1]^{|\mathfrak{u}|}}\left|\frac{\partial^{|\mathfrak{u}|}f}{\partial\boldsymbol{x}_{\mathfrak{u}}}(\boldsymbol{x}_{\mathfrak{u}},\mathbf{1})\right|\,\mathrm{d}\boldsymbol{x}_{\mathfrak{u}}.$$

## 3.5 Further Reading and Exercises

### Further Reading

This chapter is mainly based on [18, Chapter 2], to which we refer for more detailed information. A very recent overview can be found in the survey by Dick, Kuo, and Sloan [22]. For an introduction to the theory of reproducing kernel Hilbert spaces we refer to the paper [4] of Aronszajn. The foundations of the theory of Hilbert spaces and of absolutely continuous functions can be found, for example, in [82]. Hickernell [35] was the first who introduced reproducing kernel Hilbert spaces in the context of QMC. A proof of the Koksma-Hlawka inequality in its general form can be found in the original paper by Hlawka [42] and in the book by Kuipers and Niederreiter [56].

### Exercises

3.1 Let $x_0,x_1,\ldots,x_{N-1}\in[0,1)$ and let $\tau$ be an arbitrary real number. Construct two continuous functions $f,g:[0,1]\to\mathbb{R}$ such that $f(x_n)=g(x_n)$ for all $n=0,1,\ldots,N-1$, but $\int_0^1 f(x)\,\mathrm{d}x=\tau+\int_0^1 g(x)\,\mathrm{d}x$. Can you find $f,g$ that are infinitely smooth, but still have this property?

3.2 Let

$$\mathcal{B}_r=\{p(x)=a_0+a_1x+\cdots+a_rx^r\ :\ a_0,\ldots,a_r\in\mathbb{R},\ x\in[0,1]\}$$

be the space of all polynomials on $[0,1]$ of degree at most $r\in\mathbb{N}_0$. Define the inner product of $p(x)=a_0+a_1x+\cdots+a_rx^r$ and $q(x)=b_0+b_1x+\cdots+b_rx^r$ in $\mathcal{B}_r$ by

$$\langle p, q\rangle = a_0 b_0 + \cdots + a_r b_r.$$

Determine the reproducing kernel $K(x, y)$ for this space and show directly that $K$ is symmetric, uniquely determined, and positive semidefinite.

3.3 Let

$$\mathcal{C}_r = \{a_0 + a_1 \exp(2\pi \mathtt{i} x) + \cdots + a_r \exp(2\pi \mathtt{i} r x) : a_0, \ldots, a_r \in \mathbb{C},\ x \in [0, 1]\}$$

be the space of all trigonometric polynomials on $[0, 1]$ of degree at most $r \in \mathbb{N}_0$. Define the inner product of $f(x) = a_0 + a_1 \exp(2\pi \mathtt{i} x) + \cdots + a_r \exp(2\pi \mathtt{i} r x)$ and $g(x) = b_0 + b_1 \exp(2\pi \mathtt{i} x) + \cdots + b_r \exp(2\pi \mathtt{i} r x)$ in $\mathcal{C}_r$ by

$$\langle f, g\rangle = a_0 \overline{b}_0 + \cdots + a_r \overline{b}_r.$$

Determine the reproducing kernel $K(x, y)$ for this space and show directly that $K$ is symmetric, uniquely determined, and positive semidefinite.

3.4 Show that the function $K(x, y) = \min(1 - x, 1 - y)$ is the reproducing kernel for the space

$$\mathcal{H} = \{f : [0, 1] \to \mathbb{R} : f \text{ absolutely continuous, } \|f\| < \infty, \text{ and } f(1) = 0\},$$

where $\|f\| = \langle f, f\rangle^{1/2}$ and $\langle f, g\rangle = \int_0^1 f'(x) g'(x)\, \mathrm{d}x$.

3.5 For an integer $b \geq 2$ let $\omega_b = \exp(2\pi \mathtt{i}/b)$. Let $k \in \mathbb{N}_0$ with $b$-adic expansion $k = \kappa_0 + \kappa_1 b + \cdots + \kappa_{a-1} b^{a-1}$. The *kth b-adic Walsh function* ${}_b\mathrm{wal}_k : \mathbb{R} \to \mathbb{C}$, periodic with period one, is, defined for $x \in [0, 1)$, as

$${}_b\mathrm{wal}_k(x) = \omega_b^{\kappa_0 \xi_1 + \kappa_1 \xi_2 + \cdots + \kappa_{a-1} \xi_a},$$

where $x = \xi_1 b^{-1} + \xi_2 b^{-2} + \xi_3 b^{-3} + \cdots$ is the $b$-adic expansion of $x$ (unique in the sense that infinitely many of the digits $\xi_i$ must be different from $b - 1$).

The system $\{{}_b\mathrm{wal}_k : k \in \mathbb{N}_0\}$ is a complete orthonormal basis of $L_2([0, 1])$ and it is called the *b-adic Walsh function system*.

For $\alpha > 1$ we define the Hilbert space $\mathcal{H}_{\mathrm{wal},b,\alpha} \subseteq L_1([0, 1])$ consisting of all functions $f$ with absolutely convergent Walsh series

$$f(x) = \sum_{k=0}^{\infty} \hat{f}_{\mathrm{wal}}(h)\, {}_b\mathrm{wal}_k(x), \quad \text{where} \quad \hat{f}_{\mathrm{wal}}(h) = \int_0^1 f(x)\, \overline{{}_b\mathrm{wal}(x)}\, \mathrm{d}x,$$

and with finite norm $\|f\|_{\mathrm{wal},b,\alpha} = \langle f, f\rangle_{\mathrm{wal},b,\alpha}^{1/2}$, where the inner product is

$$\langle f, g\rangle_{\mathrm{wal},b,\alpha} = \sum_{k=0}^{\infty} r_{\mathrm{wal},b,\alpha}(h) \hat{f}(h) \overline{\hat{g}(h)},$$

with $r_{\mathrm{wal},b,\alpha}(0) = 1$ and, $r_{\mathrm{wal},b,\alpha}(k) = b^{-\alpha(1 + \lfloor \log_b k \rfloor)}$, for $k \in \mathbb{N}$.

Show that $\mathcal{H}_{\mathrm{wal},b,\alpha}$ is a reproducing kernel Hilbert space with kernel

$$K_{\mathrm{wal}}(x, y) = \sum_{k=0}^{\infty} r_{\mathrm{wal},b,\alpha}(k)\, {}_b\mathrm{wal}_k(x)\, \overline{{}_b\mathrm{wal}_k(y)} \quad \text{for} \quad x, y \in [0, 1).$$

3.6 Show that the worst-case error of QMC integration in $\mathcal{B}_r$ from Exercise 3.2 based on a point set $\mathcal{P} = \{x_0, \ldots, x_{N-1}\}$ in $[0, 1)$ is given by

$$e^2(\mathcal{B}_r,\mathcal{P}) = \sum_{l=0}^{r}\left(\frac{1}{l+1} - \frac{1}{N}\sum_{n=0}^{N-1} x_n^l\right)^2.$$

Assume that $\mathcal{P}$ is given by $x_n = n/N$ for $n = 0,\ldots,N-1$. Show that then $e(\mathcal{B}_r,\mathcal{P}) \asymp_r 1/N$.

3.7 Show that the worst-case error of QMC integration in $\mathcal{C}_r$ from Exercise 3.3 based on a point set $\mathcal{P} = \{x_0,\ldots,x_{N-1}\}$ in $[0,1)$ is given by

$$e^2(\mathcal{C}_r,\mathcal{P}) = \sum_{l=1}^{r}\left|\frac{1}{N}\sum_{n=0}^{N-1}\exp(2\pi \mathtt{i} l x_n)\right|^2.$$

Let $N > r$ be a prime number and assume that $\mathcal{P}$ is given by $x_n = n/N$ for $n = 0,\ldots,N-1$. Show that then $e(\mathcal{C}_r,\mathcal{P}) = 0$.

3.8 Let $K_1(x,y) = 1 + \min(1-x, 1-y)$. Show that the local discrepancy of a point set $\mathcal{P} = \{x_0,\ldots,x_{N-1}\}$ satisfies

$$\Delta_{\mathcal{P},N}(y) = \frac{\mathrm{d}}{\mathrm{d}y}\left(\int_0^1 K_1(x,y)\,\mathrm{d}x - \frac{1}{N}\sum_{n=0}^{N-1} K_1(x_n,y)\right) \tag{3.12}$$

for all $y \notin \mathcal{P}$, so that in particular (3.12) holds for almost all $y \in [0,1]$.

3.9 Let $\mathcal{H}$ be the reproducing kernel Hilbert space with kernel

$$K(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{s}\min(1-x_j, 1-y_j),$$

for $\boldsymbol{x} = (x_1,\ldots,x_s)$ and $\boldsymbol{y} = (y_1,\ldots,y_s)$ in $[0,1]^s$ and the inner product

$$\langle f,g\rangle = \int_{[0,1]^s}\frac{\partial^s f}{\partial \boldsymbol{x}}(\boldsymbol{x})\frac{\partial^s g}{\partial \boldsymbol{x}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Show that for every $N$-element point set $\mathcal{P}$ in $[0,1)^s$ we have $e(\mathcal{H},\mathcal{P}) = L_{2,N}(\mathcal{P})$.

3.10 Use Theorem 3.9 to compute $e(\mathcal{H},\mathcal{P})$ from Exercise 3.9 and show in this way the formula from Exercise 2.7.

3.11 Let $\mathcal{H}_s$ be the reproducing kernel Hilbert space with reproducing kernel (3.10) and inner product (3.11). Show that

$$\begin{aligned} e^2(\mathcal{H}_s,\mathcal{P}) &= \sum_{\emptyset\neq\mathfrak{u}\subseteq[s]}(L_{2,N}(\mathcal{P}_{\mathfrak{u}}))^2 \\ &= \frac{4^s}{3^s} - \frac{2}{N}\sum_{n=1}^{N}\prod_{j=1}^{s}\frac{3-x_{n,j}^2}{2} + \frac{1}{N^2}\sum_{n,m=1}^{N}\prod_{j=1}^{s}[1+\min(1-x_{n,j},1-x_{m,j})], \end{aligned}$$

where $\mathcal{P}_{\mathfrak{u}}$ stands for the projection of the points in $\mathcal{P}$ onto the coordinates in $\mathfrak{u}$ and $L_{2,N}(\mathcal{P}_{\mathfrak{u}})$ stands for the $L_2$ discrepancy of $\mathcal{P}_{\mathfrak{u}}$. *Remark:* Note that this result is the multi-dimensional version of (3.7).

3.12 Let $\mathcal{H}$ be a Hilbert space of functions $f : [0,1]^s \to \mathbb{C}$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$, and with a reproducing kernel $K$.

For every $i = 1,\ldots,k$ let $\mathcal{P}_i$ be an $N_i$-element point sets in $[0,1)^s$. Let $\mathcal{P}$ be the superposition of the $\mathcal{P}_i$, where repeated elements are allowed. Let $N = N_1 + \cdots + N_k$. Show the so-called *triangle inequality for the worst-case error:*

$$e(\mathcal{H},\mathcal{P}) \le \sum_{i=1}^{k} \frac{N_i}{N} e(\mathcal{H},\mathcal{P}_i).$$

3.13 Let $K$ be a reproducing kernel. Show that condition **C** implies that

$$\int_{[0,1]^{2s}} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} < \infty.$$

3.14 Let $\mathcal{H}$ be a Hilbert space of functions $f : [0,1] \to \mathbb{C}$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$, and with a reproducing kernel $K$ that satisfies condition **C**. The so-called *initial error* is defined as

$$e(\mathcal{H},0) = \sup_{f\in\mathcal{H},\|f\|\le 1} \left| \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \right|.$$

Show that

$$e(\mathcal{H},0) = \int_{[0,1]^{2s}} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}.$$

*Remark:* The initial error is often used as a reference value for the worst-case error.

3.15 Let $\mathcal{H}$ be a Hilbert space of functions $f : [0,1] \to \mathbb{C}$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$, and with a reproducing kernel $K$ satisfying condition **C**.

Rather than QMC rules, we now consider more general so-called *linear integration rules* of the form

$$Q_{\mathcal{P},\boldsymbol{w}}(f) = \sum_{n=0}^{N-1} w_n f(\boldsymbol{x}_n),$$

where $\boldsymbol{w} = (w_0,\ldots,w_{N-1}) \in \mathbb{R}^N$ and $\mathcal{P} = \{\boldsymbol{x}_0,\ldots,\boldsymbol{x}_{N-1}\} \subseteq [0,1)^s$ are given. Note that a QMC rule is obtained by choosing $\boldsymbol{w} = (N^{-1},\ldots,N^{-1})$. The worst-case integration error for a linear rule is defined as

$$e(\mathcal{H},\mathcal{P},\boldsymbol{w}) = \sup_{f\in\mathcal{H},\|f\|\le 1} \left| \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - Q_{\mathcal{P},\boldsymbol{w}}(f) \right|.$$

Show that

$$e^2(\mathcal{H},\mathcal{P},\boldsymbol{w}) = \int_{[0,1]^2} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} - 2\sum_{n=0}^{N-1} w_n \int_{[0,1]^s} K(\boldsymbol{x}_n,\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$
$$+ \sum_{n,m=0}^{N-1} w_n w_m K(\boldsymbol{x}_n,\boldsymbol{x}_m).$$

3.16 Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel $K$ satisfying $\int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$. As integration nodes $\mathcal{P}$ choose realisations of $N$ independent and uniformly distributed random variables $X_0,\ldots,X_{N-1}$ in $[0,1]^s$. Show that the QMC mean square worst-case error $\mathbb{E}[e^2(\mathcal{H},\mathcal{P})]$ is given by

$$\mathbb{E}[e^2(\mathcal{H},\mathcal{P})] = \frac{1}{N}\left[\int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{[0,1]^{2s}} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\right].$$

# Lattice Point Sets

# 4

## 4.1 Definition and Discrepancy Estimates

We have shown in Proposition 2.6 that the infinite sequence $(\{n\boldsymbol{\alpha}\})_{n\in\mathbb{N}_0}$ is uniformly distributed modulo one under a certain condition on the vector $\boldsymbol{\alpha} \in \mathbb{R}^s$. In this chapter we consider "finite" versions of such sequences which are referred to as lattice point sets.

**Definition 4.1**

Let $s, N \in \mathbb{N}$, $N \geq 2$, and let $\boldsymbol{g} \in \mathbb{Z}^s$. The point set $\mathcal{P}(\boldsymbol{g}, N) = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$, with

$$\boldsymbol{x}_n := \left\{\frac{n}{N}\boldsymbol{g}\right\} \quad \text{for} \quad n = 0, 1, \ldots, N-1,$$

where the fractional part $\{\cdot\}$ is applied component-wise, is called a *lattice point set*. The vector $\boldsymbol{g}$ is called the *generating vector* of the lattice point set.

An example of a two-dimensional lattice point set consisting of $N = 987$ elements is illustrated in Fig. 4.1.

Every element of a lattice point set $\mathcal{P}(\boldsymbol{g}, N) = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ with generating vector $\boldsymbol{g} \in \mathbb{Z}^s$ is of the form $\boldsymbol{x}_n = \{(1/N)\boldsymbol{y}_n\}$ with $\boldsymbol{y}_n = n\boldsymbol{g} \in \mathbb{Z}^s$. Hence, for a discrepancy estimate we can apply Theorem 2.30. In this way we obtain

$$D_N(\mathcal{P}(\boldsymbol{g}, N)) \leq 1 - \left(1 - \frac{1}{N}\right)^s + \sum_{\boldsymbol{h}\in C_s^*(N)} \frac{1}{r(\boldsymbol{h}, N)} \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} n \boldsymbol{h} \cdot \boldsymbol{g}/N) \right|.$$

Using the formula for geometric sums we obtain the following important property of lattice point sets:

$$\sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} n \boldsymbol{h} \cdot \boldsymbol{g}/N) = \begin{cases} N & \text{if } \boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \pmod{N}, \\ 0 & \text{if } \boldsymbol{h} \cdot \boldsymbol{g} \not\equiv 0 \pmod{N}. \end{cases} \tag{4.1}$$

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_4,

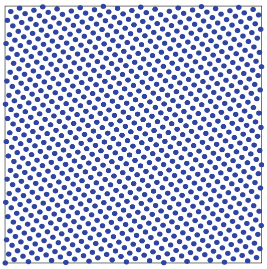

**Fig. 4.1** A two-dimensional lattice point set with $N = 987$ elements and generating vector $\boldsymbol{g} = (1{,}610)$

**Definition 4.2**

The set

$$\mathcal{L}(\boldsymbol{g}, N) = \{\boldsymbol{h} \in \mathbb{Z}^s \ : \ \boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \pmod N\}$$

is called the *dual lattice* of the lattice point set $\mathcal{P}(\boldsymbol{g}, N)$.

Hence the above discrepancy estimate for lattice point sets reduces to

$$D_N(\mathcal{P}(\boldsymbol{g}, N)) \le 1 - \left(1 - \frac{1}{N}\right)^s + \sum_{\boldsymbol{h} \in C_s^*(N) \cap \mathcal{L}(\boldsymbol{g}, N)} \frac{1}{r(\boldsymbol{h}, N)}. \tag{4.2}$$

For $h \in \mathbb{Z}$ we define $r_1(h) = \max(1, |h|)$ and for $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{Z}^s$ we define $r_1(\boldsymbol{h}) = \prod_{j=1}^s r_1(h_j)$. Then for any $\boldsymbol{h} \in C_s^*(N)$ we have $r(\boldsymbol{h}, N) \ge 2r_1(\boldsymbol{h})$. This inequality follows from the fact that $\sin(\pi t) \ge 2t$ for $0 \le t \le 1/2$. Now we define

$$R_N(\boldsymbol{g}) := \sum_{\boldsymbol{h} \in C_s^*(N) \cap \mathcal{L}(\boldsymbol{g}, N)} \frac{1}{r_1(\boldsymbol{h})}. \tag{4.3}$$

Summing up we get the following estimate for the discrepancy of lattice point sets.

**Proposition 4.3.** *The discrepancy of a lattice point set $\mathcal{P}(\boldsymbol{g}, N)$ in $[0, 1)^s$ with $\boldsymbol{g} \in \mathbb{Z}^s$ and $N \ge 2$ satisfies*

$$D_N(\mathcal{P}(\boldsymbol{g}, N)) \le \frac{s}{N} + \frac{1}{2} R_N(\boldsymbol{g}).$$

Now, for given $N$, we aim at finding a generating vector $\boldsymbol{g}$ with low value for the quantity $R_N$. Such generating vectors yield a low discrepancy for the corresponding lattice point set.

One way of finding generating vectors with small value of $R_N$ is a computer search. Note that it follows from the definition of $R_N$ that it is enough to search for $\boldsymbol{g}$ within the finite set $\{0, \ldots, N-1\}^s$. Nevertheless, apart from small values of $N$ and $s$, an exhaustive search for $\boldsymbol{g}$ is practically not feasible, since one still has to check

$N^s$ integer vectors. At this point we use an idea that was first used by N.M. Korobov in the 1960s and which was later re-invented by I.H. Sloan and A.V. Reztsov in 2002. This basic idea is to construct a generating vector component-by-component (CBC). That is, for given $N$, one starts with a sufficiently good one-dimensional generator $(g_1)$. To this generator one appends a second component $g_2$ which is chosen as to minimize a desired figure of merit, in our case the quantity $R_N$. In a next step, one appends to the now two-dimensional generator $(g_1, g_2)$ a third component $g_3$ which is again chosen so as to minimize the desired figure of merit. This procedure is repeated until one obtains an $s$-dimensional generating vector. In each of the $s$ steps the search space $\{0, \dots, N-1\}$ has cardinality $N$ and hence the overall search space for the component-by-component method reduces to a size of $sN$. Therefore, this provides a feasible way of finding a generating vector which is (hopefully) "good" with respect to the involved figure of merit.

It is obvious that we may exclude the element 0 from the search space. Let in the following $G_N := \{1, \dots, N-1\}$.

**Algorithm 4.4 (CBC algorithm).** Let $s, N \in \mathbb{N}$.

1. Choose $g_1 = 1$.
2. For $d = 2, 3, \dots, s$, choose $g_d \in G_N$ to minimize $R_N((g_1, \dots, g_{d-1}, z))$ as a function of $z \in G_N$.

Having formulated a construction algorithm, there is still the question whether the resulting generating vector is good in the sense that it yields a sufficiently small value for the figure of merit $R_N$, and hence also for the discrepancy of the corresponding lattice point set. This question is answered in the affirmative in Theorem 4.5. For technical reasons we restrict ourselves to integers $N$ which are prime numbers.

**Theorem 4.5.** *Let $N$ be a prime number. If the lattice point $\boldsymbol{g} = (g_1, \dots, g_s)$ is constructed using Algorithm 4.4, then for all $d \in \{1, \dots, s\}$*

$$R_N(\boldsymbol{g}_d) \leq \frac{1}{N-1}(1 + S_N)^d,$$

*where $S_N = \sum_{h \in C_1^*(N)} |h|^{-1}$ and where $\boldsymbol{g}_d = (g_1, \dots, g_d)$.*

*Proof.* Since $N$ is a prime number it follows that $R_N(z) = 0$ for all $z \in G_N$, which, in particular, applies for $g_1 = 1$. Let $d \in \mathbb{N}$ and assume that

$$R_N(\boldsymbol{g}) \leq \frac{1}{N-1}(1 + S_N)^d,$$

where $\boldsymbol{g} = (g_1, \dots, g_d)$. Now, with some abuse of notation, we consider the $(d+1)$-dimensional lattice point $(\boldsymbol{g}, g_{d+1}) := (g_1, \dots, g_d, g_{d+1})$.

According to Algorithm 4.4, $g_{d+1}$ is a minimizer of the function $R_N((\boldsymbol{g},\cdot))$ over $G_N$. Therefore,

$$\begin{aligned}R_N((\boldsymbol{g},g_{d+1})) &\le \frac{1}{N-1}\sum_{g_{d+1}=1}^{N-1} \sum_{\substack{(\boldsymbol{h},h_{d+1})\in C_{d+1}^*(N)\\ \boldsymbol{h}\cdot\boldsymbol{g}+h_{d+1}g_{d+1}\equiv 0 \pmod N}} \frac{1}{r_1(\boldsymbol{h})}\frac{1}{r_1(h_{d+1})}\\ &= \sum_{(\boldsymbol{h},h_{d+1})\in C_{d+1}^*(N)} \frac{1}{r_1(\boldsymbol{h})}\frac{1}{r_1(h_{d+1})}\frac{1}{N-1} \sum_{\substack{g_{d+1}\in G_N\\ \boldsymbol{h}\cdot\boldsymbol{g}+h_{d+1}g_{d+1}\equiv 0 \pmod N}} 1,\end{aligned}$$

where we just interchanged the order of summation. Decomposing the range of summation according to $C_{d+1}^*(N) = (C_d^*(N)\times\{\boldsymbol{0}\}) \cup (C_d(N)\times C_1^*(N))$ we obtain

$$\begin{aligned}&R_N((\boldsymbol{g},g_{d+1}))\\ &\le R_N(\boldsymbol{g}) + \sum_{\boldsymbol{h}\in C_d(N)}\frac{1}{r_1(\boldsymbol{h})} \sum_{h_{d+1}\in C_1^*(N)} \frac{1}{r_1(h_{d+1})}\frac{1}{N-1} \sum_{\substack{g_{d+1}\in G_N\\ h_{d+1}g_{d+1}\equiv -\boldsymbol{h}\cdot\boldsymbol{g} \pmod N}} 1.\end{aligned}$$

Now we use that $N$ is a prime number. In this case we have that the congruence $h_{d+1}g_{d+1} \equiv -\boldsymbol{h}\cdot\boldsymbol{g} \pmod N$ has exactly one solution $g_{d+1}\in G_N$ whenever $\boldsymbol{h}\cdot\boldsymbol{g}\not\equiv 0 \pmod N$ and no solution in $G_N$ whenever $\boldsymbol{h}\cdot\boldsymbol{g}\equiv 0 \pmod N$. This leads to

$$\begin{aligned}R_N((\boldsymbol{g},g_{d+1})) &\le R_N(\boldsymbol{g}) + \frac{1}{N-1}\sum_{\boldsymbol{h}\in C_d(N)}\frac{1}{r_1(\boldsymbol{h})}\sum_{h_{d+1}\in C_1^*(N)}\frac{1}{r_1(h_{d+1})}\\ &= R_N(\boldsymbol{g}) + \frac{S_N}{N-1}\sum_{\boldsymbol{h}\in C_d(N)}\frac{1}{r_1(\boldsymbol{h})}\\ &= R_N(\boldsymbol{g}) + \frac{S_N}{N-1}(1+S_N)^d\\ &\le \frac{1}{N-1}(1+S_N)^d + \frac{S_N}{N-1}(1+S_N)^d\\ &= \frac{1}{N-1}(1+S_N)^{d+1},\end{aligned}$$

where from line three to four of this displayed formula we applied the induction hypothesis for $R_N(\boldsymbol{g})$. Hence the desired result follows by induction on $d$. □

Using a straightforward estimate we have

$$S_N \le 2\sum_{h=1}^{\lfloor N/2\rfloor}\frac{1}{h} \le 2\left(1+\int_1^{N/2}\frac{\mathrm{d}t}{t}\right) = 2(\log N + 1 - \log 2). \qquad (4.4)$$

Hence Algorithm 4.4 yields lattice points $\boldsymbol{g} \in G_N^s$ with

$$\begin{aligned} R_N(\boldsymbol{g}) &\leq \frac{1}{N-1}\left(1+2(\log N+1-\log 2)\right)^s \\ &\leq \frac{2^{s+1}}{N}(\log N+1)^s = O_s\left(\frac{(\log N)^s}{N}\right). \end{aligned}$$

This estimate is best possible in the order of magnitude in $N$, as was shown by G. Larcher in 1987, and stated in the following theorem.

**Theorem 4.6 (Larcher).** *For every dimension $s \geq 2$ there exists a $c_s > 0$, depending only on $s$, such that for all $\boldsymbol{g} \in \mathbb{Z}^s$ and every $N \in \mathbb{N}$ we have*

$$R_N(\boldsymbol{g}) > c_s \frac{\ell(\log(N/\ell))^s}{N}$$

*with $\ell = \gcd(g_1, g_2, \ldots, g_s, N)$.*

The sophisticated proof of this result is based on the theory of continued fractions, which is beyond the scope of this introductory book.

Combining Proposition 4.3 and Theorem 4.5 we obtain the following estimate for the discrepancy of lattice point sets whose generating vector is constructed by Algorithm 4.4.

**Corollary 4.7.** *Let $N$ be a prime number. If the lattice point $\boldsymbol{g} = (g_1, \ldots, g_s)$ is constructed with Algorithm 4.4, then, for all $d \in \{1, \ldots, s\}$,*

$$D_N(\mathcal{P}(\boldsymbol{g}_d, N)) \leq \frac{d}{N} + \frac{2^d}{N}(\log N+1)^d,$$

*where $\boldsymbol{g}_d = (g_1, \ldots, g_d)$.*

Hence the CBC algorithm yields lattice points $\boldsymbol{g} \in G_N^s$ for which the corresponding lattice point set $\mathcal{P}(\boldsymbol{g}, N)$ satisfies

$$D_N(\mathcal{P}(\boldsymbol{g}, N)) = O_s\left(\frac{(\log N)^s}{N}\right).$$

Using a different approach, it was shown by V.A. Bykovskii in 2012 that for every $s, N \in \mathbb{N}$, $s \geq 2$ and $N \geq 3$, there exists a generating vector $\boldsymbol{g} \in \mathbb{Z}^s$ for which

$$D_N(\mathcal{P}(\boldsymbol{g}, N)) = O_s\left(\frac{(\log N)^{s-1}\log\log N}{N}\right). \tag{4.5}$$

For dimension $s = 2$ this result was already obtained by G. Larcher in 1986. It is currently the best result for the discrepancy of lattice point sets. In contrast to Corollary 4.7, this is a pure existence result.

## 4.2 The Fast CBC Construction

In the last section we learned that, for given $N$ and $s$, we can construct a generating vector $\boldsymbol{g}$ with low discrepancy component-by-component. So far we did not think about the construction cost of the CBC algorithm. Is the CBC algorithm reasonably fast to construct generating vectors also for large $N$ and large dimensions $s$? We will answer this question in the affirmative in this section.

First, let us think about the computation of the figure of merit $R_N$ which is required in each step of the algorithm. Using (4.1) we have

$$\begin{aligned} R_N(\boldsymbol{g}) &= \sum_{\boldsymbol{h} \in C_s^*(N) \cap \mathcal{L}(\boldsymbol{g}, N)} \frac{1}{r_1(\boldsymbol{h})} \\ &= \sum_{\boldsymbol{h} \in C_s^*(N)} \frac{1}{r_1(\boldsymbol{h})} \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} n \boldsymbol{h} \cdot \boldsymbol{g}/N) \\ &= -1 + \frac{1}{N} \sum_{n=0}^{N-1} \prod_{j=1}^{s} \left[ \sum_{h \in C(N)} \frac{1}{r_1(h)} \exp(2\pi \mathtt{i} n h g_j / N) \right] \\ &= -1 + \frac{1}{N} \sum_{n=0}^{N-1} \prod_{j=1}^{s} \phi\left(\frac{n g_j}{N}\right), \end{aligned}$$

where

$$\phi(x) = \sum_{h \in C(N)} \frac{1}{r_1(h)} \exp(2\pi \mathtt{i} h x).$$

Observe that $\phi\,(ng/N)$ for $n = 0, 1, \ldots, N-1$ and $g \in G_N$ takes on at most $N$ different values, $\phi\,(ng/N) = \phi(k/N)$ for some $k \in \{0, 1, \ldots, N-1\}$. The possible values for $\phi$ can be precomputed and stored, and the computational cost for this is at most $O(N^2)$.

Once we have available the values $\phi(k/N)$ for $k \in \{0, 1, \ldots, N-1\}$, we can compute the quantity $R_N(\boldsymbol{g})$ for a vector $\boldsymbol{g} \in G_N^s$ at a cost of at most $O(sN)$ elementary operations. Algorithm 4.4 needs at most $sN$ evaluations of the quantity $R_N$ in order to find the minimizer in $G_N$ for all dimensions $d = 1, 2, \ldots, s$. Therefore, we find that the construction cost of Algorithm 4.4 is at most of order $O(s^2 N^2)$. The order in which $s$ appears can in fact be easily reduced to 1: After each step of the CBC construction we may store the values $\prod_{j=1}^{d} \phi(\frac{n g_j}{N})$, for all $n \in$

$\{0,1,\ldots,N-1\}$, which are needed for the computation of $R_N((g_1,\ldots,g_d,z))$, $z\in\{0,1,\ldots,N-1\}$ in the next step. Thus the computation of the values $\prod_{j=1}^{d+1}\phi(\frac{ng_j}{N})$ for all $n\in\{0,1,\ldots,N-1\}$ needs only $N$ multiplications for every value of $g_{d+1}$, so that the total cost for computing the next component is $O(N^2)$. Note that updating the values $\prod_{j=1}^{d}\phi(\frac{ng_j}{N})$ to $\prod_{j=1}^{d+1}\phi(\frac{ng_j}{N})$, for all $n\in\{0,1,\ldots,N-1\}$ only requires another $N$ operations. Therefore, the cost of computing a generating vector using Algorithm 4.4 is $O((s+1)N^2+sN)=O(sN^2)$.

The quadratic occurrence of $N$ in the cost means that the CBC algorithm in the form of Algorithm 4.4 can only be used for moderate values of $N$. However, we are interested in lattice point sets with a large number of points. Thus, we need to reduce the factor $N^2$ in the construction cost to get an applicable construction method also for large values of $N$.

A breakthrough regarding this problem was obtained by D. Nuyens and R. Cools in 2006 using fast Fourier transform (FFT) methods for the construction of lattice point sets. This way it is possible to construct, for a given prime number $N$, an $s$-dimensional generating vector $\boldsymbol{g}$ in $O(sN\log N)$ operations, compared to $O(s^2N^2)$ operations for the usual CBC algorithm. In the following we explain this so-called *fast CBC construction.*

First, the values $\phi(k/N)$ for $k\in\{0,1,\ldots,N-1\}$ can be computed using only $O(N\log N)$ operations. Consider the vector

$$\boldsymbol{\phi}_N=\left(\phi\left(\frac{0}{N}\right),\phi\left(\frac{1}{N}\right),\ldots,\phi\left(\frac{N-1}{N}\right)\right)^\top,$$

where

$$\begin{aligned}\phi\left(\frac{k}{N}\right)&=\sum_{h\in C(N)}\frac{1}{r_1(h)}\exp\left(2\pi\mathtt{i}\frac{hk}{N}\right)\\&=\sum_{h=0}^{\frac{N-1}{2}}\frac{1}{r_1(h)}\exp\left(2\pi\mathtt{i}\frac{hk}{N}\right)+\sum_{h=\frac{N+1}{2}}^{N-1}\frac{1}{r_1(h-N)}\exp\left(2\pi\mathtt{i}\frac{hk}{N}\right)\end{aligned}$$

for $k=0,1,\ldots,N-1$ (recall that $N$ is an odd number).

For $m\in\mathbb{N}$, let $\omega_m=\exp(2\pi\mathtt{i}/m)$ and let $F_m:=\frac{1}{\sqrt{m}}\left(\omega_m^{kl}\right)_{k,l=0}^{m-1}$ be the Fourier matrix of order $m$. Note that $F_m$ is symmetric and that $F_m\overline{F}_m=I_m$, where $I_m$ denotes the $m\times m$ identity matrix and $\overline{F}_m$ denotes the complex conjugate of the matrix $F_m$.

With this notation we can write

$$\boldsymbol{\phi}_N=\sqrt{N}F_N\boldsymbol{x},$$

where

$$\boldsymbol{x} = \left( r_1(0)^{-1}, \dots, r_1\left(\frac{N-1}{2}\right)^{-1}, r_1\left(\frac{N-1}{2}\right)^{-1}, \dots, r_1(1)^{-1} \right)^\top .$$

Using FFT the matrix-vector product $F_N \boldsymbol{x}$ for the computation of $\boldsymbol{\phi}_N$ requires only $O(N \log N)$ operations. The procedure of the FFT will be explained later on in this section (see Theorem 4.10). Note that for the storage of the vector $\boldsymbol{\phi}_N$ we require a memory space of size $O(N)$.

Now we turn to the actual CBC algorithm. Using Algorithm 4.4 we construct, component-by-component, a generating vector $\boldsymbol{g} = (g_1, \dots, g_s) \in G_N^s$ such that for all $d = 1, \dots, s$, the quantity $R_N((g_1, \dots, g_{d-1}, z))$ is minimized with respect to $z \in G_N$ for fixed $g_1, \dots, g_{d-1}$.

Thus, assume that the components $g_1, \dots, g_{d-1}$ have already been constructed, and that the values $\eta_{d-1}(n) = \prod_{j=1}^{d-1} \phi(\frac{n g_j}{N})$ have been computed for all $n \in \{0, 1, \dots, N-1\}$ (in the case $d = 1$ we set $\eta_0(n) = 1$).

We have to find $z \in G_N$ which minimizes

$$R_N(z) := R_N((g_1, \dots, g_{d-1}, z)) = -1 + \frac{1}{N} \sum_{n=0}^{N-1} \phi\left(\frac{nz}{N}\right) \eta_{d-1}(n).$$

For $n = 0$ we have $\phi\left(\frac{ng}{N}\right) = \phi(0)$, which does not depend on $z$. Hence we can write

$$R_N(z) = -1 + \frac{\phi(0)^d}{N} + \frac{1}{N} \sum_{n=1}^{N-1} \phi\left(\frac{nz}{N}\right) \eta_{d-1}(n). \tag{4.6}$$

In formula (4.6) only the terms under the sum depend on $z$, so minimizing $R_N(z)$ is equivalent to minimizing

$$T_N(z) = \sum_{n=1}^{N-1} \phi\left(\frac{nz}{N}\right) \eta_{d-1}(n). \tag{4.7}$$

Now a key observation is that formula (4.7) does express the column vector $\boldsymbol{T}_N = (T_N(1), \dots, T_N(N-1))^\top$ as the product of the $(N-1) \times (N-1)$ matrix

$$\Omega_N := \left( \phi\left(\frac{nz}{N}\right) \right)_{\substack{z=1,\dots,N-1 \\ n=1,\dots,N-1}}$$

with the vector $\boldsymbol{\eta}_{d-1} = (\eta_{d-1}(1), \dots, \eta_{d-1}(N-1))^\top$:

$$\boldsymbol{T}_N = \Omega_N \boldsymbol{\eta}_{d-1}.$$

Now, while the cost of general matrix-vector multiplication is quadratic in the dimension, it can by much lower for special matrices. Fortunately, it turns out that

$\Omega_N$ is such a special matrix, and we will show in the remainder of this section that the cost for computing $\boldsymbol{T}_N = \Omega_N \boldsymbol{\eta}_{d-1}$ is $O(N \log N)$.

**Algorithm 4.8.**

```
1. Compute the vector $\boldsymbol{\phi}_N$.
2. Set $\boldsymbol{\eta}_0 = \mathbf{1} = (1, \ldots, 1)^\top$
3. For $d = 1, \ldots, s$ do:
    (i) Compute $\boldsymbol{T}_N = \Omega_N \boldsymbol{\eta}_{d-1}$, where $\boldsymbol{T} = (T_N(1), \ldots, T_N(N-1))^\top$;
    (ii) Set $g_d := \operatorname{argmin}_{z \in G_N} T_N(z)$;
    (iii) Set $\eta_d(n) := \eta_{d-1}(n)\phi\left(\frac{n g_d}{N}\right)$ for all $n \in \{1, \ldots, N-1\}$;
    (iv) If $d < s$, increase $d$ by 1, otherwise exit;
```

It is now evident that, provided $\boldsymbol{T}_N = \Omega_N \boldsymbol{\eta}_{d-1}$ can be computed in $O(N \log N)$ operations, the total cost of computing $\boldsymbol{g} = (g_1, \ldots, g_s)$ using Algorithm 4.8 is $O(sN \log N)$.

Now let us investigate the matrix-vector product $\Omega_N \boldsymbol{\eta}$. The entries of $\Omega_N$ are of the form $\phi\left(\frac{nz}{N}\right)$, where the product of the nonzero integers $z$ and $n$ has to be evaluated modulo $N$. Since $N$ is a prime number, there exists a primitive root $q$ modulo $N$, that is, there is a $q \in \{1, \ldots, N-1\}$ such that $\{q^k \pmod{N} : k = 0, \ldots, N-2\} = \{1, \ldots, N-1\}$. Thus, any product of nonzero integers $z$ and $n$ can be written – modulo $N$ – as a power of $q$. Now we aim at permuting the rows of $\Omega_N$ by the positive powers of the primitive root $q$ and the columns by the negative powers of $q$.

We describe this procedure, which is often called *Rader transform*, in detail. Let $q$ be a primitive root modulo $N$. Define an $(N-1) \times (N-1)$ matrix $\Pi(q) = (\pi_{k,l}(q))_{k,l=1,\ldots,N-1}$ by

$$\pi_{k,l}(q) = \begin{cases} 1 & \text{if } k \equiv q^l \pmod{N}, \\ 0 & \text{otherwise.} \end{cases}$$

Since $q$ is a primitive root modulo $N$, it follows that each row and each column of $\Pi(q)$ has exactly one entry which is 1 and the remaining entries are 0. Further, $\Pi(q)\Pi(q)^\top = I_{N-1}$, the $(N-1) \times (N-1)$ identity matrix. In fact, $\Pi(q)$ is a permutation matrix. That is, for any $(N-1)\times(N-1)$ matrix $C$, $\Pi(q)C$ just changes the order of the rows of $C$ and $C\,\Pi(q)$ only changes the order of the columns of $C$.

Let $C = (c_{k,l})_{k,l=1,\ldots,N-1}$, where

$$C = \Pi(q)\Omega_N \Pi(q^{-1}).$$

Then

$$c_{k,l} = \sum_{u,v=1}^{N-1} \pi_{u,k}(q)\phi\left(\frac{uv}{N}\right)\pi_{v,l}(q^{-1}) = \phi\left(\frac{q^k q^{-l}}{N}\right).$$

Let $c_r = \phi\,(q^r/N)$. Note that $c_r = c_{r'}$ for all $r, r' \in \mathbb{Z}$ with $r \equiv r' \pmod{N-1}$, since $q^{N-1} \equiv 1 \pmod{N}$. Then $c_{k,l} = c_{k-l}$ and therefore we obtain

$$C = \begin{pmatrix} c_0 & c_{-1} & \dots & \dots & c_2 & c_1 \\ c_1 & c_0 & c_{-1} & \dots & \dots & c_2 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ c_{-2} & \dots & \dots & c_1 & c_0 & c_{-1} \\ c_{-1} & c_{-2} & \dots & \dots & c_1 & c_0 \end{pmatrix}. \tag{4.8}$$

Matrices of the form $C$ are called *circulant*. In general, a circulant matrix $C_m = \mathrm{circ}(\boldsymbol{c})$ of order $m$ is an $m \times m$ matrix defined by the $m$ elements of a vector $\boldsymbol{c} = (c_0, c_1, \dots, c_{m-1})^\top$ as

$$C_m = \begin{pmatrix} c_0 & c_{m-1} & \dots & \dots & c_2 & c_1 \\ c_1 & c_0 & c_{m-1} & \dots & \dots & c_2 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ c_{m-2} & \dots & \dots & c_1 & c_0 & c_{m-1} \\ c_{m-1} & c_{m-2} & \dots & \dots & c_1 & c_0 \end{pmatrix}.$$

For such a matrix we set $c_{k'} = c_k$ for all $k, k' \in \mathbb{Z}$ such that $k \equiv k' \pmod{m}$.

Note that the circulant matrix $C_m$ is fully determined by its first column $\boldsymbol{c}$. Such matrices have a similarity transform which has the Fourier matrix as its eigenvectors. We shall prove this in the next lemma. Before we do so, recall that the Fourier matrix of order $m$ is defined as $F_m = \frac{1}{\sqrt{m}}\left(\omega_m^{kl}\right)_{k,l=0}^{m-1}$ with $\omega_m = \exp(2\pi\mathtt{i}/m)$. Let $\mathrm{diag}(a_1, \dots, a_m)$ be the $m \times m$ diagonal matrix $(A_{i,j})_{i,j=1}^m$ with $A_{i,i} = a_i$ for $i = 1, \dots, m$ and $A_{i,j} = 0$ for $i \neq j$.

**Lemma 4.9.** *A circulant matrix $C_m = \mathrm{circ}(\boldsymbol{c})$ of order $m$, with first column $\boldsymbol{c} = (c_0, c_{m-1}, c_{m-2}, \dots, c_1)^\top$, has a similarity transform $C_m = \overline{F}_m D F_m$, with a diagonal matrix $D = \mathrm{diag}(p_{\boldsymbol{c}}(1), p_{\boldsymbol{c}}(\omega_m), \dots, p_{\boldsymbol{c}}(\omega_m^{m-1}))$, where $p_{\boldsymbol{c}}(z) := c_0 + c_1 z + \dots + c_{m-1} z^{m-1}$.*

*Proof.* Let $D = (d_{k,l})_{k,l=0,\dots,m-1}$ be given by $D = F_m C_m \overline{F}_m$. Then

$$d_{k,l} = \frac{1}{m} \sum_{u,v=0}^{m-1} \omega_m^{ku} c_{u-v} \omega_m^{-lv} = \frac{1}{m} \sum_{u=0}^{m-1} \omega_m^{u(k-l)} \sum_{v=0}^{m-1} c_{u-v} \omega_m^{l(u-v)}.$$

We have $\sum_{v=0}^{m-1} c_{u-v}\omega_m^{l(u-v)} = p_{\boldsymbol{c}}(\omega_m^l)$, and therefore

$$d_{k,l} = p_{\boldsymbol{c}}(\omega_m^l)\frac{1}{m}\sum_{u=0}^{m-1}\omega_m^{u(l-k)}.$$

The result now follows since $\frac{1}{m}\sum_{u=0}^{m-1}\omega_m^{u(l-k)} = 1$ if $l = k$ and 0 otherwise. □

Thus, we have shown that

$$\Omega_N = \Pi(q)^\top \overline{F}_{N-1} D F_{N-1} \Pi(q^{-1})^\top,$$

where $\Pi(q)^\top$, $\Pi(q^{-1})^\top$ are permutation matrices, $F_{N-1}$ is a Fourier matrix, $\overline{F}_{N-1}$ its complex conjugate, and $D$ is a diagonal matrix.

For any vector $\boldsymbol{x} = (x_1, \dots, x_{N-1})^\top$ with $x_j \in \mathbb{C}$ for $j = 1, \dots, N-1$ the matrix-vector multiplications $\Pi(g)^\top \boldsymbol{x}$, $D\boldsymbol{x}$, and $\Pi(g^{-1})^\top \boldsymbol{x}$ can be done in at most $O(N)$ operations. Hence it only remains to show that $F_{N-1}\boldsymbol{x}$ and $\overline{F}_{N-1}\boldsymbol{x}$ can be computed in $O(N \log N)$ operations. Since $\overline{F}_{N-1}\boldsymbol{x} = \overline{F_{N-1}\overline{\boldsymbol{x}}}$, it is enough to show that $F_{N-1}\boldsymbol{x}$ can be computed in $O(N \log N)$ operations. Again this can be done using the FFT.

We illustrate this in the following. The Fourier matrix considered here is of size $N-1$ (for the computation of the vector $\boldsymbol{\phi}_N$ it is of size $N$, but this case can be handled analogously). The FFT is most efficient for Fourier matrices whose size is a power of 2. Hence we first aim at transforming the matrix-vector product of size $N-1$ into a related one of size $2^{h+1}$.

Note that the matrix $C$ given in (4.8) is a $(N-1) \times (N-1)$ matrix. Let $h \in \mathbb{N}$ such that $2^{h-1} < N - 1 \le 2^h$ and let $k \in \{0, \dots, 2^{h-1} - 1\}$ with $N - 1 + k = 2^h$. We extend $C$ by $k$ rows and $k$ columns to obtain the $2^h \times 2^h$ matrix

$$T = \begin{pmatrix}
c_0 & c_{-1} & \dots & \dots & c_2 & c_1 & 0 & 0 & \dots & 0 \\
c_1 & c_0 & c_{-1} & \dots & \dots & c_2 & c_1 & 0 & \dots & \ddots \\
\vdots & \ddots & \ddots & \ddots & & \vdots & c_2 & c_1 & \ddots & \ddots \\
\vdots & & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots \\
c_{-2} & \dots & \dots & c_1 & c_0 & c_{-1} & \vdots & \vdots & \ddots & \\
c_{-1} & c_{-2} & \dots & \dots & c_1 & c_0 & c_{-1} & c_{-2} & \ddots & \ddots \\
0 & c_{-1} & c_{-2} & \dots & \dots & c_1 & c_0 & c_{-1} & \ddots & \ddots \\
0 & 0 & c_{-1} & c_{-2} & \dots & \dots & c_1 & c_0 & \ddots & \ddots \\
\vdots & \ddots & \ddots & & & & \ddots & \ddots & \ddots & \ddots \\
0 & \ddots & \ddots & \ddots & & & \ddots & \ddots & c_1 & c_0
\end{pmatrix}.$$

$T$ is no longer a circulant matrix, but it is still a Toeplitz matrix. To obtain a circulant matrix again, let

$$R = \begin{pmatrix} 0 & \dots & 0 & c_{-1} & \dots & \dots & c_2 & c_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & 0 & \ddots & 0 & c_{-1} & \ddots & \ddots \\ c_1 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & c_{-1} \\ \vdots & \ddots & c_1 & \ddots & \ddots & \ddots & \ddots & 0 \\ c_{-2} & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ c_{-1} & c_{-2} & \ddots & \ddots & c_1 & 0 & \ddots & 0 \end{pmatrix}$$

of size $2^h \times 2^h$. Then

$$C' = \begin{pmatrix} T & R \\ R & T \end{pmatrix}$$

is a circulant matrix of size $2^{h+1} \times 2^{h+1}$.

Let $\boldsymbol{x} = (x_1, \dots, x_{N-1})^\top$ be a complex column vector and assume we want to multiply the $(N-1) \times (N-1)$ matrix $C$ given by (4.8) by $\boldsymbol{x}$. Then we can do so by multiplying the matrix $C'$ by the vector

$$\boldsymbol{x}' = (x_1, \dots, x_{N-1}, \underbrace{0, \dots, 0}_{k+2^h \text{ zeros}})^\top$$

of length $2^{h+1}$. Let $\boldsymbol{y}' = (y_1, \dots, y_{2^{h+1}})^\top = C'\boldsymbol{x}'$. Then $\boldsymbol{y} = (y_1, \dots, y_{N-1})^\top = C\boldsymbol{x}$. Hence we can use Lemma 4.9 with $n = 2^{h+1}$ (rather than $n = N-1$). This simplifies the FFT algorithm. The following result was shown by J.W. Cooley and J.W. Tukey in 1965.

**Theorem 4.10 (Cooley and Tukey).** *Let $F_{2^{h+1}} = 2^{-(h+1)/2}(\omega_{2^{h+1}}^{kl})_{k,l=0,\dots,2^{h+1}-1}$ be a Fourier matrix. Let $\boldsymbol{u} = (u_0, \dots, u_{2^{h+1}-1})^\top$ be a given complex vector of length $2^{h+1}$. Then the matrix-vector product $F_{2^{h+1}}\boldsymbol{u}$ can be computed in $O((h+1)2^{h+1})$ operations.*

*Proof.* Let $\boldsymbol{z} = (z_0, \dots, z_{2^{h+1}-1})^\top = F_{2^{h+1}}\boldsymbol{u}$, where $\boldsymbol{u} = (u_0, \dots, u_{2^{h+1}-1})^\top$. Then

$$z_k = 2^{-\frac{h+1}{2}} \sum_{l=0}^{2^{h+1}-1} \omega_{2^{h+1}}^{kl} u_l \quad \text{for } k = 0, \dots, 2^{h+1} - 1.$$

We compute the above sum recursively. Let $k = \kappa_0 + \kappa_1 2 + \dots + \kappa_h 2^h$ and $l = \lambda_0 + \lambda_1 2 + \dots + \lambda_h 2^h$, with dyadic digits $\kappa_0, \dots, \kappa_h, \lambda_0, \dots, \lambda_h \in \{0, 1\}$. Put

$$G_0(\lambda_0,\dots,\lambda_{h-1},\kappa_0) = \sum_{\lambda_h=0}^{1} \omega_2^{\kappa_0\lambda_h} u_{\lambda_0+\lambda_1 2+\cdots+\lambda_h 2^h}, \tag{4.9}$$

and, for $r \in \{1,\dots,h\}$, put

$$\begin{aligned} &G_r(\lambda_0,\dots,\lambda_{h-r-1},\kappa_0,\dots,\kappa_r) \\ &= \sum_{\lambda_{h-r}=0}^{1} \omega_{2^{r+1}}^{(\kappa_0+\cdots+\kappa_r 2^r)\lambda_{h-r}} G_{r-1}(\lambda_0,\dots,\lambda_{h-r},\kappa_0,\dots,\kappa_{r-1}). \end{aligned} \tag{4.10}$$

Now we show by induction on $r$ that for all $r = 1,\dots,h$ we have

$$\begin{aligned} &G_r(\lambda_0,\dots,\lambda_{h-r-1},\kappa_0,\dots,\kappa_r) \\ &= \sum_{\lambda_{h-r}=0}^{1} \cdots \sum_{\lambda_h=0}^{1} \omega_{2^{r+1}}^{(\kappa_0+\cdots+\kappa_r 2^r)(\lambda_{h-r}+\cdots+\lambda_h 2^r)} u_{\lambda_0+\cdots+\lambda_h 2^h}. \end{aligned} \tag{4.11}$$

For $r = 1$ we have

$$G_1(\lambda_0,\dots,\lambda_{h-2},\kappa_0,\kappa_1) = \sum_{\lambda_{h-1}=0}^{1} \omega_4^{(\kappa_0+\kappa_1 2)\lambda_{h-1}} \sum_{\lambda_h=0}^{1} \omega_2^{\kappa_0\lambda_h} u_{\lambda_0+\cdots+\lambda_h 2^h}.$$

Now note that $\omega_2^{\kappa_0\lambda_h} = \omega_4^{2\kappa_0\lambda_h} = \omega_4^{2\kappa_0\lambda_h+4\kappa_1\lambda_h} = \omega_4^{(\kappa_0+2\kappa_1)2\lambda_h}$, and hence

$$G_1(\lambda_0,\dots,\lambda_{h-2},\kappa_0,\kappa_1) = \sum_{\lambda_{h-1}=0}^{1} \sum_{\lambda_h=0}^{1} \omega_4^{(\kappa_0+\kappa_1 2)(\lambda_{h-1}+\lambda_h 2)} u_{\lambda_0+\cdots+\lambda_h 2^h}.$$

Thus (4.11) holds for $r = 1$. Assume that (4.11) holds for the index $r$. Then

$$\begin{aligned} &G_{r+1}(\lambda_0,\dots,\lambda_{h-r-2},\kappa_0,\dots,\kappa_{r+1}) \\ &= \sum_{\lambda_{h-r-1}=0}^{1} \omega_{2^{r+2}}^{(\kappa_0+\cdots+\kappa_{r+1}2^{r+1})\lambda_{h-r-1}} G_r(\lambda_0,\dots,\lambda_{h-r-1},\kappa_0,\dots,\kappa_r) \\ &= \sum_{\lambda_{h-r-1}=0}^{1} \omega_{2^{r+2}}^{(\kappa_0+\cdots+\kappa_{r+1}2^{r+1})\lambda_{h-r-1}} \\ &\qquad \times \sum_{\lambda_{h-r}=0}^{1} \cdots \sum_{\lambda_h=0}^{1} \omega_{2^{r+1}}^{(\kappa_0+\cdots+\kappa_r 2^r)(\lambda_{h-r}+\cdots+\lambda_h 2^r)} u_{\lambda_0+\cdots+\lambda_h 2^h}. \end{aligned}$$

Since

$$\begin{aligned}\omega_{2^{r+1}}^{(\kappa_0+\cdots+\kappa_r 2^r)(\lambda_{h-r}+\cdots+\lambda_h 2^r)} &= \omega_{2^{r+1}}^{(\kappa_0+\cdots+\kappa_r 2^r+\kappa_{r+1}2^{r+1})(\lambda_{h-r}+\cdots+\lambda_h 2^r)}\\ &= \omega_{2^{r+2}}^{(\kappa_0+\cdots+\kappa_{r+1}2^{r+1})(\lambda_{h-r}2+\cdots+\lambda_h 2^{r+1})},\end{aligned}$$

we obtain

$$\begin{aligned}&G_{r+1}(\lambda_0,\ldots,\lambda_{h-r-2},\kappa_0,\ldots,\kappa_{r+1})\\ &= \sum_{\lambda_{h-r-1}=0}^{1}\cdots\sum_{\lambda_h=0}^{1}\omega_{2^{r+2}}^{(\kappa_0+\cdots+\kappa_{r+1}2^{r+1})(\lambda_{h-r-1}+\cdots+\lambda_h 2^{r+1})} u_{\lambda_0+\cdots+\lambda_h 2^h}\end{aligned}$$

and hence (4.11) is shown.

In particular, (4.11) leads with the choice $r = h$ to the desired formula

$$z_k = 2^{-\frac{h+1}{2}} G_h(\kappa_0,\ldots,\kappa_h),$$

for all $k = \kappa_0 + \cdots + \kappa_h 2^h \in \{0,\ldots,2^{h+1}-1\}$.

For the computation of the $z_k$ we compute $G_r$ recursively. For $r = 0,\ldots,h$ compute

$$G_r(\lambda_0,\ldots,\lambda_{m-r-1},\kappa_0,\ldots,\kappa_r)$$

for all $\lambda_0,\ldots,\lambda_{h-r-1},\kappa_0,\ldots,\kappa_r \in \{0,1\}$ using (4.9) and the recursion formula (4.10). For each $r$ this requires $O(2^{h+1})$ operations.

Overall we require $O((h+1)2^{h+1})$ operations to compute $G_h$. Hence, we can compute $\boldsymbol{z}$ in $O((h+1)2^{h+1})$ operations. □

Recall that we have chosen $h$ such that $2^{h-1} < N - 1 \leq 2^h$ and hence $h < 1 + \log_2 N$. Thus we obtain the following result.

**Corollary 4.11.** *Let $\Omega_N$ be defined as above and let $\boldsymbol{x}$ be a complex column vector of length $N-1$. Then the matrix-vector product $\Omega_N\boldsymbol{x}$ can be computed in $O(N\log N)$ operations.*

Therefore Algorithm 4.8 requires only $O(sN\log N)$ operations by using $O(N)$ memory space (compared to $O(sN^2)$ operations). This is a significant speed-up compared to a straightforward implementation of Algorithm 4.8. Only through this reduction of the construction cost does the CBC algorithm become applicable for the generation of lattice point sets with reasonably large cardinality.

## 4.3 Numerical Integration in Korobov Spaces

If we use a lattice point set as constructed by means of Algorithm 4.4 as underlying sample points of a QMC rule for functions $f : [0,1]^s \to \mathbb{R}$ with bounded variation in the sense of Hardy and Krause, then Corollary 4.7 together with the inequality of Koksma and Hlawka guarantee a convergence rate of order $O((\log N)^s/N)$. For the Koksma-Hlawka inequality no additional information on the smoothness of the integrand was required. On the other hand, for functions of higher smoothness we would certainly expect a better convergence rate of the integration error. This, however, is not reflected in the error estimate via the Koksma-Hlawka inequality.

Originally, lattice point sets were introduced for the integration of smooth and one-periodic functions. Such functions will be considered in the following.

**Definition 4.12**

A QMC-rule that uses a lattice point set as underlying sample nodes is called a *lattice rule*.

We will see that lattice rules are particularly suitable for the integration of periodic functions of sufficient smoothness. We define reproducing kernel Hilbert spaces of smooth one-periodic functions and analyze the worst-case error of lattice rules in these function spaces.

We begin with the one-dimensional case. The smoothness in this function space will be described by the so-called smoothness parameter $\alpha > 1$, which controls the decay of the Fourier coefficients of the functions. Let $\gamma > 0$ be a real number. The value of $\gamma$ does not have any influence on the one-dimensional analysis and may be chosen equal to one in that case. However, it will play a crucial role in the multi-dimensional case, where every dimension will be assigned its own $\gamma$. We define the Hilbert space $\mathcal{H}_{\alpha,\gamma} \subseteq L_1([0,1])$ consisting of all one-periodic functions $f$ with absolutely convergent Fourier series

$$f(x) = \sum_{h \in \mathbb{Z}} \hat{f}(h) \exp(2\pi \mathrm{i} hx)$$

with Fourier coefficients

$$\hat{f}(h) = \int_0^1 f(x) \exp(-2\pi \mathrm{i} hx) \,\mathrm{d}x$$

and with finite norm $\|f\|_{\alpha,\gamma} = \langle f, f \rangle_{\alpha,\gamma}^{1/2}$, where the inner product is defined by

$$\langle f, g \rangle_{\alpha,\gamma} = \sum_{h \in \mathbb{Z}} r_{\alpha,\gamma}(h) \hat{f}(h) \overline{\hat{g}(h)}$$

and

$$r_{\alpha,\gamma}(h) = \begin{cases} 1 & \text{if } h = 0, \\ \gamma^{-1}|h|^{\alpha} & \text{if } h \neq 0. \end{cases}$$

**Definition 4.13**

The space $\mathcal{H}_{\alpha,\gamma}$ is called the *Korobov space of smoothness* $\alpha$ (with weight $\gamma$).

We now show that the reproducing kernel of $\mathcal{H}_{\alpha,\gamma}$ is given by

$$K_{\alpha,\gamma}(x, y) = \sum_{h\in\mathbb{Z}} \frac{1}{r_{\alpha,\gamma}(h)} \exp(2\pi \mathrm{i} h(x - y)). \tag{4.12}$$

First, note that since $\alpha > 1$, the series in (4.12) is uniformly convergent, so $K_{\alpha,\gamma}(\cdot, y)$ is integrable and

$$\begin{aligned}\hat{K}_{\alpha,\gamma}(\cdot, y)(h) &= \int_0^1 K_{\alpha,\gamma}(x, y) \exp(-2\pi \mathrm{i} hx)\, \mathrm{d}x \\ &= \sum_{k\in\mathbb{Z}} \frac{1}{r_{\alpha,\gamma}(h)} \exp(-2\pi \mathrm{i} ky) \int_0^1 \exp(2\pi \mathrm{i}(k - h)x)\, \mathrm{d}x \\ &= \frac{\exp(-2\pi \mathrm{i} hy)}{r_{\alpha,\gamma}(h)}.\end{aligned}$$

Thus

$$\|K_{\alpha,\gamma}(\cdot, y)\|_{\alpha}^2 = \langle K_{\alpha,\gamma}(\cdot, y), K_{\alpha,\gamma}(\cdot, y)\rangle_{\alpha,\gamma} = \sum_{h\in\mathbb{Z}} r_{\alpha,\gamma}(h) \frac{1}{r_{\alpha,\gamma}^2(h)} = 1 + 2\gamma\zeta(\alpha),$$

where $\zeta(\alpha) = \sum_{h=1}^{\infty} h^{-\alpha}$ denotes the Riemann zeta function. For $\alpha > 1$ we have $\zeta(\alpha) < \infty$, and therefore we obtain $K_{\alpha,\gamma}(\cdot, y) \in \mathcal{H}_{\alpha,\gamma}$. Furthermore,

$$\langle f, K_{\alpha,\gamma}(\cdot, y)\rangle_{\alpha,\gamma} = \sum_{h\in\mathbb{Z}} r_{\alpha,\gamma}(h) \hat{f}(h) \frac{\exp(2\pi \mathrm{i} hy)}{r_{\alpha,\gamma}(h)} = f(y)$$

and hence also the reproducing property of $K_{\alpha,\gamma}$ holds. This shows that $K_{\alpha,\gamma}$ is indeed the reproducing kernel of $\mathcal{H}_{\alpha,\gamma}$.

The smoothness parameter $\alpha$ is linked to the actual smoothness of functions in the sense that a periodic function $f$ which is sufficiently often continuously differentiable belongs to the space $\mathcal{H}_{\alpha,\gamma}$.

**Proposition 4.14.** *Let $\alpha \geq 2$ be an integer and assume that $f : [0, 1] \to \mathbb{R}$ is one-periodic and $f \in C^{\alpha}([0, 1])$. Then $f \in \mathcal{H}_{\alpha,\gamma}$.*

*Proof.* The function $f$ and its derivatives $f^{(1)}, f^{(2)}, \ldots, f^{(\alpha-1)}$ are continuously differentiable and one-periodic, and the derivative $f^{(\alpha)}$ is continuous and one-periodic. Let $h \neq 0$ be an integer. Integration by parts then yields

$$\begin{aligned}\hat{f}(h) &= \int_0^1 f(x)\exp(-2\pi \mathtt{i} hx)\,\mathrm{d}x\\ &= f(x)\frac{\exp(-2\pi \mathtt{i} hx)}{-2\pi \mathtt{i} h}\bigg|_0^1 + \frac{1}{2\pi \mathtt{i} h}\int_0^1 f^{(1)}(x)\exp(-2\pi \mathtt{i} hx)\,\mathrm{d}x\\ &= \frac{1}{2\pi \mathtt{i} h}\int_0^1 f^{(1)}(x)\exp(-2\pi \mathtt{i} hx)\,\mathrm{d}x\\ &= \frac{1}{2\pi \mathtt{i} h}\widehat{f^{(1)}}(h).\end{aligned}$$

Now we can repeat this with $\widehat{f^{(1)}}(h)$ in place of $\hat{f}(h)$ and obtain

$$\hat{f}(h) = \frac{1}{(2\pi \mathtt{i} h)^2}\widehat{f^{(2)}}(h).$$

After $\alpha$ steps we finally obtain

$$\hat{f}(h) = \frac{1}{(2\pi \mathtt{i} h)^\alpha}\widehat{f^{(\alpha)}}(h).$$

Consequently,

$$|\hat{f}(h)| \le \frac{1}{(2\pi|h|)^\alpha}\int_0^1 |f^{(\alpha)}(x)|\,\mathrm{d}x = \frac{c}{|h|^\alpha},$$

where $c := \frac{1}{(2\pi)^\alpha}\int_0^1 |f^{(\alpha)}(x)|\,\mathrm{d}x < \infty$, since $f^{(\alpha)}$ is continuous. Thus we obtain

$$\|f\|_{\alpha,\gamma}^2 = \langle f, f\rangle_{\alpha,\gamma} = \sum_{h\in\mathbb{Z}} r_{\alpha,\gamma}(h)|\hat{f}(h)|^2 \le |\hat{f}(0)|^2 + 2\gamma^{-1}c\zeta(\alpha) < \infty$$

and this implies $f \in \mathcal{H}_{\alpha,\gamma}$. □

On the other hand, every function in $\mathcal{H}_{\alpha,\gamma}$ has some smoothness which will be shown in Proposition 4.17 for arbitrary dimension $s \in \mathbb{N}$.

Now we turn to the $s$-dimensional case where $s \in \mathbb{N}$. Let $\boldsymbol{\gamma} = (\gamma_j)_{j\in\mathbb{N}}$ be a sequence of positive weights. We consider the $s$-fold tensor product $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ of the spaces $\mathcal{H}_{\alpha,\gamma_1}, \mathcal{H}_{\alpha,\gamma_2}, \ldots, \mathcal{H}_{\alpha,\gamma_s}$, i.e.,

$$\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}} = \mathcal{H}_{\alpha,\gamma_1} \otimes \mathcal{H}_{\alpha,\gamma_2} \otimes \cdots \otimes \mathcal{H}_{\alpha,\gamma_s} = \operatorname{clos\,span}\left\{\boldsymbol{x} \mapsto \prod_{j=1}^s f_j(x_j) \;:\; f_j \in \mathcal{H}_{\alpha,\gamma_j}\right\},$$

where $\boldsymbol{x} = (x_1, \ldots, x_s)$ and where the closure clos is taken with respect to the norm induced by the inner product

$$\langle f, g\rangle_{s,\alpha,\boldsymbol{\gamma}} = \sum_{\boldsymbol{h}\in\mathbb{Z}^s} r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})\hat{f}(\boldsymbol{h})\overline{\hat{g}(\boldsymbol{h})},$$

and where $r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h}) = \prod_{j=1}^{s} r_{\alpha,\gamma_j}(h_j)$. The reproducing kernel $K_{s,\alpha,\boldsymbol{\gamma}}$ of $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ is the $s$-fold product of the one-dimensional kernels $K_{\alpha,\gamma_j}$, for $j = 1, \ldots, s$, i.e.,

$$K_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{s} K_{\alpha,\gamma_j}(x_j, y_j),$$

where $\boldsymbol{x} = (x_1, \ldots, x_s)$ and $\boldsymbol{y} = (y_1, \ldots, y_s)$. Then we have

$$\begin{aligned} K_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{h_1,\ldots,h_s\in\mathbb{Z}} \frac{\exp(2\pi \mathtt{i} h_1(x_1 - y_1))\cdots\exp(2\pi \mathtt{i} h_s(x_s - y_s))}{r_{\alpha,\gamma_1}(h_1)\cdots r_{\alpha,\gamma_s}(h_s)} \\ &= \sum_{\boldsymbol{h}\in\mathbb{Z}^s} \frac{\exp(2\pi \mathtt{i} \boldsymbol{h}\cdot(\boldsymbol{x} - \boldsymbol{y}))}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})}. \end{aligned}$$

The proof that $K_{s,\alpha,\boldsymbol{\gamma}}$ is the reproducing kernel of $\mathcal{H}_{\alpha,s,\boldsymbol{\gamma}}$ follows along the same lines as in the one-dimensional case.

**Definition 4.15**

The space $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ is called a *weighted Korobov space of smoothness $\alpha$ with weights $\boldsymbol{\gamma}$.*

We present a multivariate version of Proposition 4.14, but without a proof. For $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_s) \in \mathbb{N}_0^s$ denote by

$$D^{\boldsymbol{\beta}} f = \frac{\partial^{|\boldsymbol{\beta}|}}{\partial x_1^{\beta_1}\, \partial x_2^{\beta_2}\, \cdots\, \partial x_s^{\beta_s}} f$$

the operator of partial differentiation, where $|\boldsymbol{\beta}| = \beta_1 + \beta_2 + \cdots + \beta_s$.

**Proposition 4.16.** *Let $\alpha \geq 2$ be an integer and assume that $f : [0, 1]^s \to \mathbb{R}$ is one-periodic in each of its s variables and that $D^{\boldsymbol{\beta}} f$ exists and is continuous for all $\boldsymbol{\beta} \in \{0, \ldots, \alpha\}^s$. Then $f \in \mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$.*

On the other hand, every function in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ has some smoothness.

**Proposition 4.17.** *If $f \in \mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$, then $D^{\boldsymbol{\beta}} f$ exists and is continuous for all $\boldsymbol{\beta} \in \{0, 1, \ldots, \lceil \frac{\alpha-1}{2} \rceil - 1\}^s$.*

*Proof.* We have

$$D^{\boldsymbol{\beta}} f(\boldsymbol{x}) = \sum_{\boldsymbol{h}\in\mathbb{Z}^s} \left[ \hat{f}(\boldsymbol{h})\,(2\pi \mathtt{i})^{|\boldsymbol{\beta}|} \prod_{j=1}^{s} h_j^{\beta_j} \right] \exp(2\pi \mathtt{i} \boldsymbol{h}\cdot\boldsymbol{x}),$$

where, by convention, we take $0^0 = 1$. The last series is convergent as long as $\max_{1\le j\le s} \beta_j < \frac{\alpha-1}{2}$. Indeed,

$$\begin{aligned} |D^{\boldsymbol{\beta}} f(\boldsymbol{x})| &= \left| \sum_{\boldsymbol{h}\in\mathbb{Z}^s} \left[\hat{f}(\boldsymbol{h}) r_{\alpha,\boldsymbol{\gamma}}^{1/2}(\boldsymbol{h})\right] \left[ r_{\alpha,\boldsymbol{\gamma}}^{-1/2}(\boldsymbol{h})(2\pi\mathtt{i})^{|\boldsymbol{\beta}|} \prod_{j=1}^{s} h_j^{\beta_j} \right] \exp(2\pi\mathtt{i}\boldsymbol{h}\cdot\boldsymbol{x}) \right| \\ &\le \|f\|_{s,\alpha,\boldsymbol{\gamma}} \left[ \sum_{\boldsymbol{h}\in\mathbb{Z}^s} (2\pi)^{2|\boldsymbol{\beta}|} \prod_{j=1}^{s} \frac{|h_j|^{2\beta_j}}{r_{\alpha,\gamma_j}(h_j)} \right]^{1/2} \\ &= \|f\|_{s,\alpha,\boldsymbol{\gamma}} \left[ (2\pi)^{2|\boldsymbol{\beta}|} \prod_{j=1}^{s} \sum_{h_j\in\mathbb{Z}} \frac{|h_j|^{2\beta_j}}{r_{\alpha,\gamma_j}(h_j)} \right]^{1/2} \\ &= \|f\|_{s,\alpha,\boldsymbol{\gamma}} \left[ (2\pi)^{2|\boldsymbol{\beta}|} \prod_{j=1}^{s} 2\gamma_j \sum_{h_j=1}^{\infty} \frac{h_j^{2\beta_j}}{h_j^{\alpha}} \right]^{1/2} \\ &= (2\pi)^{|\boldsymbol{\beta}|} 2^{s/2} \|f\|_{s,\alpha,\boldsymbol{\gamma}} \left[ \prod_{j=1}^{s} \gamma_j \zeta(\alpha - 2\beta_j) \right]^{1/2} < \infty. \end{aligned}$$

□

Now we study QMC integration in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ for $\alpha > 1$. We have

$$K_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{y},\boldsymbol{y}) = \prod_{j=1}^{s} (1 + 2\gamma_j \zeta(\alpha)) < \infty.$$

Hence condition **C** is satisfied and we can therefore use the formula for the worst-case error from Theorem 3.9. We have

$$\begin{aligned} \int_{[0,1]^s}\int_{[0,1]^s} K_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} &= \prod_{j=1}^{s} \int_0^1\int_0^1 K_{\alpha,\gamma_j}(x_j,y_j)\,\mathrm{d}x_j\,\mathrm{d}y_j \\ = \prod_{j=1}^{s} \left( \sum_{h\in\mathbb{Z}} \frac{1}{r_{\alpha,\gamma_j}(h)} \int_0^1\int_0^1 \exp(2\pi\mathtt{i}h(x_j-y_j))\,\mathrm{d}x_j\,\mathrm{d}y_j \right) &= 1. \end{aligned}$$

In the same way one can show that for any fixed $\boldsymbol{x}_n \in [0,1)^s$

$$\int_{[0,1]^s} K_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{x}_n, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = 1.$$

Now from Theorem 3.9 we obtain for any $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ in $[0,1)^s$

$$\begin{aligned} e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}) &= -1 + \frac{1}{N^2} \sum_{n,m=0}^{N-1} \sum_{\boldsymbol{h} \in \mathbb{Z}^s} \frac{\exp(2\pi \mathtt{i} \boldsymbol{h} \cdot (\boldsymbol{x}_n - \boldsymbol{x}_m))}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})} \\ &= \sum_{\boldsymbol{h} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\}} \frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})} \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} \boldsymbol{h} \cdot \boldsymbol{x}_n) \right|^2 . \end{aligned} \tag{4.13}$$

For lattice point sets $\mathcal{P}(\boldsymbol{g}, N)$, i.e., $\boldsymbol{x}_n = \{n\boldsymbol{g}/N\}$, where $\boldsymbol{g} \in \mathbb{Z}^s$, the above formula yields

$$e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N)) = \sum_{\boldsymbol{h} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\}} \frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})} \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathtt{i} n \boldsymbol{h} \cdot \boldsymbol{g}/N) \right|^2 .$$

Now, using property (4.1) and the notion $\mathcal{L}(\boldsymbol{g}, N)$ of the dual lattice of $\mathcal{P}(\boldsymbol{g}, N)$ (Definition 4.2) we obtain the following result.

**Theorem 4.18.** *The worst-case error of a lattice rule in the weighted Korobov space $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ is given by*

$$e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N)) = \sum_{\boldsymbol{h} \in \mathcal{L}(\boldsymbol{g}, N) \setminus \{\boldsymbol{0}\}} \frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})}.$$

▶ **Remark 4.19** If $\alpha \geq 2$ is an even integer, then the Bernoulli polynomial $B_\alpha$ of degree $\alpha$ has the Fourier expansion

$$B_\alpha(x) = \frac{(-1)^{(\alpha+2)/2} \alpha!}{(2\pi)^\alpha} \sum_{\substack{h \in \mathbb{Z} \\ h \neq 0}} \frac{\exp(2\pi \mathtt{i} h x)}{|h|^\alpha} \quad \text{for all } x \in [0,1).$$

Hence in this case we obtain

$$e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N)) = -1 + \frac{1}{N} \sum_{k=0}^{N-1} \prod_{j=1}^{s} \left(1 + \gamma_j \frac{(-1)^{(\alpha+2)/2} (2\pi)^\alpha}{\alpha!} B_\alpha\left(\left\{\frac{k g_j}{N}\right\}\right)\right),$$

so that $e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N))$ can be calculated in $O(Ns)$ operations.

## 4.4 Existence Results for the Unweighted Case

In this section we restrict ourselves to the unweighted case where all weights are equal to 1. In this case we omit the weights in our notations from the previous section and just write $\mathcal{H}_{s,\alpha}$, $K_{s,\alpha}$, $r_\alpha$, etc. Note that in the unweighted case we have

$$r_\alpha(h) = \max(1, |h|^\alpha) = r_1(h)^\alpha.$$

First we show a result which brings the quantity $R_N$ as defined in (4.3) in relation to the worst-case error of a lattice rule in the Korobov space.

**Lemma 4.20.** *Let $\alpha > 1$ and let $N$ be a prime number. Then for all $\boldsymbol{g} \in G_N^s$*

$$e^2(\mathcal{H}_{s,\alpha}, \mathcal{P}(\boldsymbol{g}, N)) \le (1 + 2\zeta(\alpha))^s \left(\frac{1}{N^\alpha} + R_N(\boldsymbol{g})^\alpha\right).$$

*Proof.* By Theorem 4.18,

$$e^2(\mathcal{H}_{s,\alpha}, \mathcal{P}(\boldsymbol{g}, N)) = \sum_{\boldsymbol{h} \in \mathcal{L}(\boldsymbol{g}, N) \setminus \{\boldsymbol{0}\}} \frac{1}{r_\alpha(\boldsymbol{h})} =: \Sigma_1 + \Sigma_2,$$

where $\Sigma_1$ is the sum over all $\boldsymbol{h}$ such that $\boldsymbol{h} = N\boldsymbol{k}$ with $\boldsymbol{k} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\}$. In this case we obviously have $\boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \pmod{N}$. In $\Sigma_2$ we sum over the remaining $\boldsymbol{h}$ from $\mathcal{L}(\boldsymbol{g}, N) \setminus \{\boldsymbol{0}\}$.

For $\boldsymbol{h} = N\boldsymbol{k}$ with $\boldsymbol{k} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\}$ we have

$$\frac{1}{r_\alpha(\boldsymbol{h})} = \frac{1}{r_\alpha(N\boldsymbol{k})} \le \frac{1}{N^\alpha} \frac{1}{r_\alpha(\boldsymbol{k})},$$

and so

$$\Sigma_1 \le \frac{1}{N^\alpha} \sum_{\boldsymbol{k} \in \mathbb{Z}^s} \frac{1}{r_\alpha(\boldsymbol{k})} = \frac{1}{N^\alpha} \left(\sum_{k \in \mathbb{Z}} \frac{1}{r_\alpha(k)}\right)^s = \frac{1}{N^\alpha} (1 + 2\zeta(\alpha))^s.$$

The remaining $\boldsymbol{h}$ which appear in $\Sigma_2$ can be uniquely represented in the form $\boldsymbol{h} = \boldsymbol{h}^* + N\boldsymbol{k}$, where $\boldsymbol{k} \in \mathbb{Z}^s$ and $\boldsymbol{h}^* = (h_1^*, \dots, h_s^*) \in C_s^*(N)$ with $\boldsymbol{h}^* \cdot \boldsymbol{g} \equiv 0 \pmod{N}$. Hence

$$\Sigma_2 = \sum_{\boldsymbol{k} \in \mathbb{Z}^s} \sum_{\substack{\boldsymbol{h}^* \in C_s^*(N) \\ \boldsymbol{h}^* \cdot \boldsymbol{g} \equiv 0 \pmod{N}}} \frac{1}{r_\alpha(\boldsymbol{h}^* + N\boldsymbol{k})}.$$

Now we show that

$$r_1(\boldsymbol{h}^* + N\boldsymbol{k}) \geq r_1(\boldsymbol{h}^*)r_1(\boldsymbol{k}). \tag{4.14}$$

To this end it suffices to verify that

$$\max(1, |h_j^* + Nk_j|) \geq \max(1, |h_j^*|) \max(1, |k_j|)$$

for all $j = 1, 2, \ldots, s$. If $h_j^* = 0$ or $k_j = 0$, then this inequality is obviously satisfied. If $h_j^* \neq 0$ and $k_j \neq 0$, then

$$|h_j^* + Nk_j| \geq N|k_j| - |h_j^*| \geq N|k_j| - \frac{N}{2} = \frac{N}{2}(2|k_j| - 1) \geq |h_j^*| \cdot |k_j|,$$

and the inequality follows as well. Hence (4.14) is shown.

From (4.14) we obtain $r_\alpha(\boldsymbol{h}^* + N\boldsymbol{k}) \geq r_\alpha(\boldsymbol{h}^*)r_\alpha(\boldsymbol{k})$. Using this estimate we find that

$$\begin{aligned}\Sigma_2 &\leq \sum_{\boldsymbol{k}\in\mathbb{Z}^s} \frac{1}{r_1(\boldsymbol{k})^\alpha} \sum_{\substack{\boldsymbol{h}^*\in C_s^*(N) \\ \boldsymbol{h}^*\cdot\boldsymbol{g}\equiv 0 \pmod{N}}} \frac{1}{r_1(\boldsymbol{h}^*)^\alpha} \\ &= (1 + 2\zeta(\alpha))^s \sum_{\substack{\boldsymbol{h}^*\in C_s^*(N) \\ \boldsymbol{h}^*\cdot\boldsymbol{g}\equiv 0 \pmod{N}}} \frac{1}{r_1(\boldsymbol{h}^*)^\alpha} \\ &\leq (1 + 2\zeta(\alpha))^s R_N(\boldsymbol{g})^\alpha.\end{aligned}$$

Now the result follows by adding the estimates for $\Sigma_1$ and $\Sigma_2$. □

It should be pointed out that the quantity $R_N(\boldsymbol{g})$ does not depend on the smoothness parameter $\alpha$. Assume now that $\boldsymbol{g} \in G_N^s$ is constructed with Algorithm 4.4. Then it follows from Theorem 4.5 in combination with (4.4) that

$$R_N(\boldsymbol{g}) \leq 2^{s+1}\frac{(1 + \log N)^s}{N}.$$

Inserting the bound on $R_N(\boldsymbol{g})$ into the bound from Lemma 4.20 yields the following result, which goes back to work by E. Hlawka.

**Theorem 4.21.** *Let $\alpha > 1$ and let $N$ be a prime number. Assume that $\boldsymbol{g} \in G_N^s$ is constructed by Algorithm 4.4. Then*

$$e^2(\mathcal{H}_{s,\alpha}, \mathcal{P}(\boldsymbol{g}, N)) \leq (1 + 2\zeta(\alpha))^s \frac{1 + 2^{\alpha(s+1)}(1 + \log N)^{\alpha s}}{N^\alpha}.$$

Theorem 4.21 shows the existence of lattice rules for which the convergence rate for the worst-case error in the Korobov space $\mathcal{H}_{s,\alpha}$ is of order

$$e(\mathcal{H}_{s,\alpha}, \mathcal{P}(\boldsymbol{g}, N)) = O_{s,\alpha}\left(\frac{(\log N)^{\alpha s/2}}{N^{\alpha/2}}\right).$$

This bound reflects the smoothness of the problem. Higher smoothness $\alpha$ leads to improved convergence rates for the worst-case integration error. Furthermore, as Algorithm 4.4 is independent of the parameter $\alpha$, it is clear that the output vector $\boldsymbol{g}$ is independent of $\alpha$ as well. Hence the lattice rule that is constructed by Algorithm 4.4 adjusts itself to the smoothness of a given integrand.

We next present a lower bound on the integration error for numerical integration in the Korobov space which shows that the result from Theorem 4.21 is, up to some $(\log N)$-powers, best possible in the order of magnitude in $N$. For reasons of simplicity we restrict ourselves to the QMC case and remark that in fact the following result holds even for more general quadrature rules.

**Theorem 4.22.** *Let $\mathcal{P}$ be an arbitrary $N$-element point set in $[0,1)^s$. For any $\alpha > 1$ we have*

$$e^2(\mathcal{H}_{s,\alpha}, \mathcal{P}) \geq C(s,\alpha)\frac{(\log N)^{s-1}}{N^\alpha},$$

*where $C(s,\alpha) > 0$ depends on $\alpha$ and $s$, but not on $N$.*

The proof of this theorem is technical and can be skipped by beginners.

*Proof.* Let $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ be an arbitrary $N$-element point set in $[0,1)^s$. Let $f : \mathbb{R} \to [0,1]$ be an infinitely differentiable function such that $f(x) > 0$ for $x \in (0,1)$ and $f^{(r)}(x) = 0$ for $x \in \mathbb{R} \setminus (0,1)$ for all $0 \leq r \leq a := \lceil \alpha/2 \rceil + 1$. For instance choose $f(x) = x^{a+1}(1-x)^{a+1}$ for $x \in (0,1)$, and $f(x) = 0$ otherwise. For $m \in \mathbb{N}_0$ let $f_m(x) = f(2^{m+2}x)$ and for $\boldsymbol{m} = (m_1, \ldots, m_s) \in \mathbb{N}_0^s$ put

$$f_{\boldsymbol{m}}(\boldsymbol{x}) = \prod_{j=1}^s f_{m_j}(x_j),$$

where $\boldsymbol{x} = (x_1, \ldots, x_s)$. Let $\|\boldsymbol{m}\|_1 = m_1 + \cdots + m_s$ and $I(f) = \int_0^1 f(y)\,\mathrm{d}y$. Then

$$\hat{f}_{\boldsymbol{m}}(\boldsymbol{0}) = \prod_{j=1}^s \int_0^1 f(2^{m_j+2}x)\,\mathrm{d}x = \prod_{j=1}^s \left(\frac{1}{2^{m_j+2}}\int_0^1 f(y)\,\mathrm{d}y\right) = \frac{1}{2^{\|\boldsymbol{m}\|_1+2s}} I^s(f). \tag{4.15}$$

Let $t$ be such that $2N \leq 2^t < 4N$, let $F(\boldsymbol{y}) = \sum_{n=0}^{N-1} f_{\boldsymbol{m}}(\boldsymbol{x}_n - \boldsymbol{y})$, and let

$$B_{\boldsymbol{m}} = \{\boldsymbol{y} \in [0,1]^s \ : \ F(\boldsymbol{y}) = 0\}.$$

Notice that the support of $f_{\boldsymbol{m}}(\boldsymbol{x}_n - \boldsymbol{y})$ (as a function of $\boldsymbol{y}$) is contained in the interval $\prod_{j=1}^{s}(x_{n,j} - 2^{-m_j-2}, x_{n,j})$ and hence the support of $F(\boldsymbol{y})$ is contained in $\bigcup_{n=0}^{N-1}\prod_{j=1}^{s}(x_{n,j} - 2^{-m_j-2}, x_{n,j})$. Therefore the volume of the support of $F$ is at most $N2^{-\|\boldsymbol{m}\|_1}$. Thus for all $\boldsymbol{m}$ such that $\|\boldsymbol{m}\|_1 = t$ we have

$$\lambda_s(B_{\boldsymbol{m}}) \geq 1 - \frac{N}{2^{\|\boldsymbol{m}\|_1}} = 1 - \frac{N}{2^t} > \frac{1}{4}.$$

For the QMC rule $Q_{N,s}$ based on $\mathcal{P}$

$$\begin{aligned}
&Q_{N,s}(f_{\boldsymbol{m}}(\cdot - \boldsymbol{y})) - \hat{f}_{\boldsymbol{m}}(\boldsymbol{0})\\
&= \sum_{\boldsymbol{h}\in\mathbb{Z}^s\setminus\{\boldsymbol{0}\}} \left(\hat{f}_{\boldsymbol{m}}(\boldsymbol{h})\frac{1}{N}\sum_{n=0}^{N-1}\exp(2\pi\mathrm{i}\boldsymbol{h}\cdot\boldsymbol{x}_n)\right)\exp(-2\pi\,\mathrm{i}\boldsymbol{h}\cdot\boldsymbol{y})
\end{aligned}$$

and, for $\boldsymbol{y} \in B_{\boldsymbol{m}}$, $Q_{N,s}(f_{\boldsymbol{m}}(\cdot - \boldsymbol{y})) = \frac{1}{N}F(\boldsymbol{y}) = 0$. Therefore,

$$\begin{aligned}
\lambda_s(B_{\boldsymbol{m}})|\hat{f}_{\boldsymbol{m}}(\boldsymbol{0})|^2 &= \int_{B_{\boldsymbol{m}}} |Q_{N,s}(f_{\boldsymbol{m}}(\cdot - \boldsymbol{y})) - \hat{f}_{\boldsymbol{m}}(\boldsymbol{0})|^2\,\mathrm{d}\boldsymbol{y}\\
&\leq \int_{[0,1]^s} |Q_{N,s}(f_{\boldsymbol{m}}(\cdot - \boldsymbol{y})) - \hat{f}_{\boldsymbol{m}}(\boldsymbol{0})|^2\,\mathrm{d}\boldsymbol{y}\\
&= \sum_{\boldsymbol{h}\in\mathbb{Z}^s\setminus\{\boldsymbol{0}\}} |\hat{f}_{\boldsymbol{m}}(\boldsymbol{h})|^2 \left|\frac{1}{N}\sum_{n=0}^{N-1}\exp(2\pi\mathrm{i}\boldsymbol{h}\cdot\boldsymbol{x}_n)\right|^2, \qquad (4.16)
\end{aligned}$$

where for the last line we applied Parseval's identity.

We have

$$\hat{f}_m(h) = \int_0^1 f(2^{m+2}x)\exp(-2\pi\mathrm{i}hx)\,\mathrm{d}x = \frac{1}{2^{m+2}}\hat{f}(h2^{-m-2}).$$

Since, by assumption, $f$ is infinitely differentiable, integration by parts shows that for any $m \in \mathbb{N}_0$ we have

$$|\hat{f}_m(h)| \leq \frac{1}{2^{m+2}}|\hat{f}(h2^{-m-2})| \leq C_a\frac{1}{2^{m+2}}\min\left(1, \frac{2^{a(m+2)}}{h^a}\right),$$

where $C_a > 0$ depends only on $a$ and $f$. Then for $\boldsymbol{m}$ with $\|\boldsymbol{m}\|_1 = t$ we have

$$\begin{aligned}
|\hat{f}_{\boldsymbol{m}}(\boldsymbol{h})| &\leq C(a,s)\prod_{j=1}^{s}\left(\frac{1}{2^{m_j}}\min\left(1, \frac{2^{am_j}}{r_a(h_j)}\right)\right)\\
&= C(a,s)2^{(\alpha/2-1)t}\prod_{j=1}^{s}\left(\frac{1}{2^{\alpha m_j/2}}\min\left(1, \frac{2^{am_j}}{r_a(h_j)}\right)\right).
\end{aligned}$$

Taking the square and summing over all choices of $\boldsymbol{m}$ with $\|\boldsymbol{m}\|_1 = t$ we obtain

$$\begin{aligned}\sum_{\substack{\boldsymbol{m}\in\mathbb{N}_0^s\\ \|\boldsymbol{m}\|_1=t}} |\hat{f}_{\boldsymbol{m}}(\boldsymbol{h})|^2 &\le 2^{(\alpha-2)t} C^2(a,s) \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_0^s\\ \|\boldsymbol{m}\|_1=t}} \prod_{j=1}^{s} \left( \frac{1}{2^{\alpha m_j}} \min\left(1, \frac{2^{2am_j}}{r_a^2(h_j)}\right)\right) \\ &\le 2^{(\alpha-2)t} C^2(a,s) \prod_{j=1}^{s} \left( \sum_{m=0}^{\infty} \frac{1}{2^{\alpha m}} \min\left(1, \frac{2^{2am}}{r_a^2(h_j)}\right)\right). \qquad (4.17)\end{aligned}$$

The last sum can now be estimated by

$$\begin{aligned}\sum_{m=0}^{\infty} \frac{1}{2^{\alpha m}} \min\left(1, \frac{2^{2am}}{r_a^2(h_j)}\right) &= \sum_{0\le m\le(\log_2 r_a(h_j))/a} \frac{2^{(2a-\alpha)m}}{r_a^2(h_j)} + \sum_{m>(\log_2 r_a(h_j))/a} \frac{1}{2^{\alpha m}} \\ &\le \frac{r_{2a-\alpha}(h_j) 2^{2a-\alpha} - 1}{2^{2a-\alpha}-1} \frac{1}{r_{2a}(h_j)} + \frac{2^{\alpha}}{2^{\alpha}-1} \frac{1}{r_{\alpha}(h_j)} \\ &\le \frac{1}{r_{\alpha}(h_j)} \left(1 + \frac{2^{\alpha}}{2^{\alpha}-1}\right) \le \frac{3}{r_{\alpha}(h_j)}. \qquad (4.18)\end{aligned}$$

Thus, combining (4.17) and (4.18), we have

$$\frac{C_1(a,s)}{2^{(\alpha-2)t}} \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_0^s\\ \|\boldsymbol{m}\|_1=t}} |\hat{f}_{\boldsymbol{m}}(\boldsymbol{h})|^2 \le \frac{1}{r_{\alpha}(\boldsymbol{h})} \qquad (4.19)$$

with some suitable $C_1(a,s) > 0$ depending only on $a$ and $s$. Using the formula for the worst-case error from (4.13), in conjunction with (4.19), (4.16) and (4.15) we finally get

$$\begin{aligned} e^2(\mathcal{H}_{s,\alpha}, \mathcal{P}) &= \sum_{\boldsymbol{h}\in\mathbb{Z}^s\setminus\{\boldsymbol{0}\}} \frac{1}{r_{\alpha}(\boldsymbol{h})} \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathrm{i} \boldsymbol{h}\cdot \boldsymbol{x}_n)\right|^2 \\ &\ge \frac{C_1(a,s)}{2^{(\alpha-2)t}} \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_0^s\\ \|\boldsymbol{m}\|_1=t}} \sum_{\boldsymbol{h}\in\mathbb{Z}^s\setminus\{\boldsymbol{0}\}} |\hat{f}_{\boldsymbol{m}}(\boldsymbol{h})|^2 \left| \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi \mathrm{i} \boldsymbol{h}\cdot \boldsymbol{x}_n)\right|^2 \\ &\ge \frac{C_1(a,s)}{2^{(\alpha-2)t}} \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_0^s\\ \|\boldsymbol{m}\|_1=t}} \lambda_s(B_{\boldsymbol{m}}) |\hat{f}_{\boldsymbol{m}}(\boldsymbol{0})|^2 \\ &\ge C_2(a,s) \frac{2^{2t}}{N^{\alpha}} \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_0^s\\ \|\boldsymbol{m}\|_1=t}} 2^{-2t-4s} I^{2s}(f) \end{aligned}$$

$$\geq C_3(a,s)\frac{1}{N^\alpha}\binom{t+s-1}{s-1} \geq C_4(a,s)\frac{t^{s-1}}{N^\alpha} \geq C_5(a,s)\frac{(\log N)^{s-1}}{N^\alpha},$$

since $t \geq \log_2 N$, with suitable $C_j(a,s) > 0$ depending only on $a$ and $s$. □

In this section we have shown that the best convergence rate of QMC rules in the Korobov space of smoothness $\alpha$ is $O(N^{-\alpha/2})$, up to some $\log N$ factors, and this rate can be achieved with a lattice rule. However, it is not clear how long we have to wait to see this nice asymptotic behavior, especially for large dimension $s$. In practical applications $s$ can be huge, for example, in the hundreds or even thousands for problems from mathematical finance. The dependence of the worst-case error on the dimension $s$ is the subject of tractability, which is the topic of the next section.

## 4.5 Tractability

To systematically analyse the dependence of the worst-case integration error on the dimension one considers the following quantities. For an arbitrary reproducing kernel Hilbert space $\mathcal{H}(K)$ we denote by $e(\mathcal{H}(K),\mathcal{P})$ the worst-case error of a QMC rule based on the point set $\mathcal{P}$ in $[0,1)^s$. The *Nth minimal (worst-case) error* is given by

$$e(N,s) := \inf_{\mathcal{P}} e(\mathcal{H}(K),\mathcal{P}),$$

where the infimum is taken over all $N$-element point sets $\mathcal{P}$ in $[0,1)^s$. For $N = 0$ we set the QMC rule to be zero, and define the *initial error* by

$$e(0,s) = \sup_{\substack{f\in\mathcal{H}(K)\\ \|f\|_{\mathcal{H}(K)}\leq 1}} \left| \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \right|.$$

The initial error is used as a reference value.

When studying QMC rules, we do not only want to control how the error depends on $N$, but also how it depends on the dimension $s$, which is of particular importance for high-dimensional problems (see also Chap. 6 for a discussion of this topic). To this end, we define, for $\varepsilon \in (0,1]$ and $s \in \mathbb{N}$, the *information complexity* by

$$N_{\mathcal{H}(K)}(\varepsilon,s) = \min\{N \in \mathbb{N} \,:\, e(N,s) \leq \varepsilon\, e(0,s)\}$$

i.e., as the minimal number of information evaluations needed to reduce the initial error by a factor of $\varepsilon$.

The subject of tractability deals with the question in which way the information complexity depends on $\varepsilon^{-1}$ and $s$. Roughly speaking, tractability means that the information complexity lacks a certain disadvantageous dependence on $\varepsilon^{-1}$ and $s$. Usually, one is interested in situations where $N_{\mathcal{H}(K)}(\varepsilon, s)$ depends at most polynomially on $s$ and $\varepsilon^{-1}$, and frequently the following notions of tractability are studied.

**Definition 4.23**

We say that we have:

- The *curse of dimensionality* if there exist positive numbers $C$, $\tau$ and $\varepsilon_0$ such that

$$N_{\mathcal{H}(K)}(\varepsilon, s) \geq C(1+\tau)^s \quad \text{for all } \varepsilon \leq \varepsilon_0 \text{ and infinitely many } s.$$

- *Polynomial* (QMC) *tractability* if there exist non-negative numbers $C, \tau_1, \tau_2$ such that

$$N_{\mathcal{H}(K)}(\varepsilon, s) \leq C s^{\tau_1} \varepsilon^{-\tau_2} \quad \text{for all } s \in \mathbb{N},\ \varepsilon \in (0,1).$$

- *Strong polynomial* (QMC) *tractability* if there exist non-negative numbers $C$ and $\tau$ such that

$$N_{\mathcal{H}(K)}(\varepsilon, s) \leq C \varepsilon^{-\tau} \quad \text{for all } s \in \mathbb{N},\ \varepsilon \in (0,1).$$

The *exponent $\tau^*$ of strong polynnomial tractability* is defined as the infimum over all $\tau$ for which strong polynomial tractability holds.

▶ **Remark 4.24** Usually in tractability theory one studies the class of arbitrary algorithms using $N$ information evaluations rather than only QMC rules as we do here. For this reason we added in the above notation the term "(QMC)". However, as we only deal with QMC rules, we will omit this term in the following.

It is known that many multivariate problems defined over standard spaces of functions suffer from the curse of dimensionality, as, for example, integration of Lipschitz functions (cf. Example 1.1), of functions from the unweighted Korobov space $\mathcal{H}_{s,\alpha}$, of monotone functions, of convex functions, of smooth functions, etc. The reason for this disadvantageous behaviour may be found in the fact that for standard spaces all variables and groups of variables are equally important. As a way out, I.H. Sloan and H. Woźniakowski suggested to consider weighted spaces, in which the importance of successive variables and groups of variables is monitored by corresponding weights, to vanquish the curse of dimensionality and obtain polynomial or even strong polynomial tractability, depending on the decay of the weights.

For this reason we now switch again to the weighted setting and study tractability for the weighted Korobov space $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$.

**Theorem 4.25.** *For any prime number $N$, any dimension $s$, and any $\lambda \in (1/\alpha, 1]$ there exists a generating vector $\boldsymbol{g} \in G_N^s$ such that*

$$e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}},\mathcal{P}(\boldsymbol{g},N)) \le \frac{1}{(N-1)^{1/\lambda}}\left(-1+\prod_{j=1}^{s}(1+2\gamma_j^{\lambda}\zeta(\lambda\alpha))\right)^{1/\lambda}.$$

*Proof.* Let $\mathrm{av}(N,s)$ denote the average squared worst-case error over all $\boldsymbol{g}\in G_N^s$,

$$\mathrm{av}(N,s)=\frac{1}{(N-1)^s}\sum_{\boldsymbol{g}\in G_N^s} e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}},\mathcal{P}(\boldsymbol{g},N)).$$

Then we have

$$\begin{aligned}\mathrm{av}(N,s)&=\frac{1}{(N-1)^s}\sum_{\boldsymbol{g}\in G_N^s}\sum_{\boldsymbol{h}\in\mathcal{L}(\boldsymbol{g},N)\setminus\{\boldsymbol{0}\}}\frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})}\\&=\frac{1}{(N-1)^s}\sum_{\boldsymbol{h}\in\mathbb{Z}^s\setminus\{\boldsymbol{0}\}}\frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})}\#\{\boldsymbol{g}\in G_N^s\ :\ \boldsymbol{h}\cdot\boldsymbol{g}\equiv 0\pmod N\}.\end{aligned}$$

Let $\boldsymbol{h}=(h_1,\ldots,h_s)\in\mathbb{Z}^s\setminus\{\boldsymbol{0}\}$. Assume first that $h_1\neq 0$. Then, for arbitrary $g_2,\ldots,g_s\in G_N$, the condition $\boldsymbol{h}\cdot\boldsymbol{g}\equiv 0 \pmod N$ is equivalent to

$$h_1g_1\equiv-(h_2g_2+\cdots+h_sg_s)\pmod N.$$

Since $N$ is prime, there exists at most one $g_1\in G_N$ which satisfies this congruence. Consequently,

$$\#\{\boldsymbol{g}\in G_N^s\ :\ \boldsymbol{h}\cdot\boldsymbol{g}\equiv 0\pmod N\}\le(N-1)^{s-1}.$$

The same argument applies if $h_j\neq 0$ for any $j\in\{1,2,\ldots,s\}$. Now it follows that

$$\mathrm{av}(N,s)\le\frac{1}{N-1}\sum_{\boldsymbol{h}\in\mathbb{Z}^s\setminus\{\boldsymbol{0}\}}\frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})}=\frac{1}{N-1}\left(-1+\prod_{j=1}^{s}(1+2\gamma_j\zeta(\alpha))\right).$$

However, since the average square worst-case error over $G_N^s$ satisfies this bound, there must exist at least one particular $\boldsymbol{g}\in G_N^s$ for which the worst-case error satisfies this bound as well. Hence we have shown the existence of $\boldsymbol{g}\in G_N^s$ for which we have

$$e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}},\mathcal{P}(\boldsymbol{g},N))\le\frac{1}{N-1}\left(-1+\prod_{j=1}^{s}(1+2\gamma_j\zeta(\alpha))\right). \tag{4.20}$$

Now we use a popular trick to improve the convergence rate of this existence result. To this end we require, for any $\lambda\in(0,1]$ and non-negative reals $a_k$, the inequality

$$\left(\sum_k a_k\right)^\lambda \leq \sum_k a_k^\lambda \tag{4.21}$$

which can be proved as follows: We have $0 \leq a_j / \left(\sum_k a_k\right) \leq 1$ and hence, since $\lambda \in (0, 1]$,

$$\frac{a_j}{\sum_k a_k} \leq \left(\frac{a_j}{\sum_k a_k}\right)^\lambda .$$

Summation over all $j$ implies

$$1 = \frac{\sum_j a_j}{\sum_k a_k} \leq \frac{\sum_j a_j^\lambda}{\left(\sum_k a_k\right)^\lambda}$$

which finally yields (4.21).

Inequality (4.21) applied to the formula for the square worst-case error yields

$$e^{2\lambda}(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N)) = \left(\sum_{\boldsymbol{h} \in \mathcal{L}(\boldsymbol{g},N)\setminus\{\boldsymbol{0}\}} \frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})}\right)^\lambda \leq \sum_{\boldsymbol{h} \in \mathcal{L}(\boldsymbol{g},N)\setminus\{\boldsymbol{0}\}} \frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})^\lambda},$$

where we have to restrict $\lambda$ to $\lambda \in (1/\alpha, 1]$ in order to guarantee the convergence of the series $\sum_{h \in \mathbb{Z}} r_{\alpha,\gamma}(h)^{-\lambda}$. For $h \in \mathbb{Z} \setminus \{0\}$ we have $r_{\alpha,\gamma}(h)^\lambda = \gamma^{-\lambda}|h|^{\lambda\alpha} = r_{\lambda\alpha,\gamma^\lambda}(h)$. Therefore,

$$r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})^\lambda = r_{\lambda\alpha,\boldsymbol{\gamma}^\lambda}(\boldsymbol{h}),$$

where $\boldsymbol{\gamma}^\lambda = (\gamma_j^\lambda)_{j \in \mathbb{N}}$, and we obtain

$$e^{2\lambda}(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N)) \leq \sum_{\boldsymbol{h} \in \mathcal{L}(\boldsymbol{g},N)\setminus\{\boldsymbol{0}\}} \frac{1}{r_{\lambda\alpha,\boldsymbol{\gamma}^\lambda}(\boldsymbol{h})} = e^2(\mathcal{H}_{s,\lambda\alpha,\boldsymbol{\gamma}^\lambda}, \mathcal{P}(\boldsymbol{g}, N)).$$

Applying the existence result (4.20) to $e^2(\mathcal{H}_{s,\lambda\alpha,\boldsymbol{\gamma}^\lambda}, \mathcal{P}(\boldsymbol{g}, N))$ we finally obtain the existence of $\boldsymbol{g} \in G_N^s$ such that

$$e^{2\lambda}(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N)) \leq \frac{1}{N-1}\left(-1 + \prod_{j=1}^s (1 + 2\gamma_j^\lambda \zeta(\lambda\alpha))\right).$$

Note that $\zeta(\lambda\alpha) < \infty$ since $\lambda \in (1/\alpha, 1]$. Hence the desired result follows. □

Since $\lambda$ can be chosen arbitrarily close to $1/\alpha$, we obtain a convergence rate of $O(N^{-\frac{\alpha}{2}+\delta})$ for $\delta > 0$. This is in accordance with Theorem 4.21. The result of Theorem 4.25 can even be made explicit with a CBC algorithm.

If we estimate the bound from Theorem 4.25 further we obtain the following result for the $N$th minimal error in the weighted Korobov space $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$:

**Corollary 4.26.** *Let $N$ be a prime number, let $s \in \mathbb{N}$, $\alpha > 1$ and let $e(N,s)$ be the $N$th minimal error in the weighted Korobov space $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$.*

1. *For some $\lambda \in (1/\alpha, 1]$ assume*
$$\Gamma_\lambda := \sum_{j=1}^{\infty} \gamma_j^\lambda < \infty. \tag{4.22}$$
*Then*
$$e(N,s) \le \frac{2^{1/(2\lambda)}}{N^{1/(2\lambda)}} \exp\left(\frac{\zeta(\lambda\alpha)}{\lambda}\Gamma_\lambda\right) \quad \text{for all } s \in \mathbb{N}.$$
*This bound is independent of the dimension $s$ and can be achieved by a lattice rule.*
2. *Assume that*
$$A := \limsup_{s\to\infty} \frac{\sum_{j=1}^{s}\gamma_j}{\log s} < \infty. \tag{4.23}$$
*Then for any $\delta > 0$ there exists a $c_\delta > 0$ such that*
$$e(N,s) \le \frac{c_\delta s^{\zeta(\alpha)(A+\delta)}}{\sqrt{N}} \quad \text{for all } s \in \mathbb{N}.$$
*This bound depends only polynomially on the dimension $s$ and can be achieved by a lattice rule.*
3. *We have $e(0,s) = 1$.*

*Proof.* To begin with, for the bound from Theorem 4.25 we have

$$\begin{aligned} e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P}(\boldsymbol{g}, N)) &\le \frac{2^{1/\lambda}}{N^{1/\lambda}} \exp\left(\frac{1}{\lambda}\log\left(\prod_{j=1}^{s}(1+2\gamma_j^\lambda\zeta(\lambda\alpha))\right)\right) \\ &= \frac{2^{1/\lambda}}{N^{1/\lambda}} \exp\left(\frac{1}{\lambda}\sum_{j=1}^{s}\log(1+2\gamma_j^\lambda\zeta(\lambda\alpha))\right) \\ &\le \frac{2^{1/\lambda}}{N^{1/\lambda}} \exp\left(\frac{2\zeta(\lambda\alpha)}{\lambda}\sum_{j=1}^{s}\gamma_j^\lambda\right), \end{aligned} \tag{4.24}$$

where we used the estimate $\log(1+x) \le x$. Now assertion 1 follows immediately from (4.24) and the condition (4.22) on the weights.

We turn to assertion 2: Since $A < \infty$, there exists, for any $\delta > 0$, a positive $s_\delta$ such that

$$\sum_{j=1}^{s} \gamma_j \leq (A+\delta)\log s \quad \text{for all } s \geq s_\delta.$$

From (4.24) with $\lambda = 1$ we obtain for $s \geq s_\delta$ that

$$e(N,s) \leq \sqrt{\frac{2}{N}}\, s^{\zeta(\alpha)\frac{\sum_{j=1}^{s}\gamma_j}{\log s}} \leq \sqrt{\frac{2}{N}}\, s^{\zeta(\alpha)(A+\delta)}.$$

Hence there exists a $c_\delta > 0$ such that $e(N,s) \leq \frac{c_\delta s^{\zeta(\alpha)(A+\delta)}}{\sqrt{N}}$ for all $s \in \mathbb{N}$.

Finally, to show that $e(0,s) = 1$ is left as an exercise (cf. Exercise 4.7). □

If $\sum_{j=1}^{\infty} \gamma_j < \infty$, then there exists a $\lambda \in (1/\alpha, 1]$ which satisfies (4.22). Put $C_{\alpha,\lambda,\boldsymbol{\gamma}} := 2\exp(2\zeta(\lambda\alpha)\Gamma_\lambda)$. For $\varepsilon > 0$, let $N$ be the smallest prime larger or equal to $\lceil C_{\alpha,\lambda,\boldsymbol{\gamma}}\varepsilon^{-2\lambda}\rceil =: M$. Then $e(N,s) \leq \varepsilon$ and hence

$$N_{\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}}(\varepsilon,s) \leq N < 2M = 2\lceil C_{\alpha,\lambda,\boldsymbol{\gamma}}\varepsilon^{-2\lambda}\rceil,$$

where we used Bertrand's postulate which tells us that $M \leq N < 2M$. Hence multivariate integration in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ is strongly polynomially tractable with $\varepsilon$-exponent at most $2\lambda_0$, where $\lambda_0$ is the infimum over all $\lambda \in (1/\alpha, 1]$ such that (4.22) holds.

In the same way one can show that polynomial tractability holds under the condition (4.23). The proof for this result is left as an exercise (cf. Exercise 4.8).

Now we turn to necessary conditions for (strong) polynomial tractability. Let $\gamma_j' = \min(\gamma_j, 1/(2\zeta(\alpha)))$ and $\boldsymbol{\gamma}' = (\gamma_j')_{j\in\mathbb{N}}$. From $\gamma_j' \leq \gamma_j$ we obtain $\|f\|_{s,\alpha,\boldsymbol{\gamma}'} \geq \|f\|_{s,\alpha,\boldsymbol{\gamma}}$ and hence

$$\{f \in \mathcal{H}_{s,\alpha,\boldsymbol{\gamma}'} \,:\, \|f\|_{s,\alpha,\boldsymbol{\gamma}'} \leq 1\} \subseteq \{f \in \mathcal{H}_{s,\alpha,\boldsymbol{\gamma}} \,:\, \|f\|_{s,\alpha,\boldsymbol{\gamma}} \leq 1\}.$$

This implies that integration in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}'}$ is no harder than integration in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$, that is, $e(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}'},\mathcal{P}) \leq e(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}},\mathcal{P})$ for any $N$-element point set in $[0,1)^s$. Furthermore, according to the definition of $\boldsymbol{\gamma}'$, we have that $K_{s,\alpha,\boldsymbol{\gamma}'}$ is non-negative.

Now let us estimate $e(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}},\mathcal{P})$ from below. We have

$$\begin{aligned} e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}},\mathcal{P}) \geq & e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}'},\mathcal{P}) = -1 + \frac{1}{N^2}\sum_{h,i=0}^{N-1} K_{s,\alpha,\boldsymbol{\gamma}'}(\boldsymbol{x}_h,\boldsymbol{x}_i) \\ \geq & -1 + \frac{1}{N^2}\sum_{h=0}^{N-1} K_{s,\alpha,\boldsymbol{\gamma}'}(\boldsymbol{x}_h,\boldsymbol{x}_h) \end{aligned}$$

$$= -1 + \frac{1}{N^2}\sum_{h=0}^{N-1}\prod_{j=1}^{s}\Big(1+2\gamma_j'\zeta(\alpha)\Big)$$

$$= -1 + \frac{1}{N}\prod_{j=1}^{s}\Big(1+2\gamma_j'\zeta(\alpha)\Big). \tag{4.25}$$

We summarize the result in the following lemma.

**Lemma 4.27.** *For any $N$-element point set $\mathcal{P}$ in $[0,1)^s$ we have*

$$e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}},\mathcal{P}) \geq -1 + \frac{1}{N}\prod_{j=1}^{s}\big(1+\min(2\zeta(\alpha)\gamma_j,1)\big)$$

*and hence*

$$N_{\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}}(s,\varepsilon) \geq \frac{\prod_{j=1}^{s}\big(1+\min(2\zeta(\alpha)\gamma_j,1)\big)}{1+\varepsilon^2}.$$

Suppose there exists a $\gamma_* > 0$ such that $\gamma_j \geq \gamma_*$ for all $j \in \mathbb{N}$. Then

$$N_{\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}}(s,\varepsilon) \geq \frac{(1+\min(2\zeta(\alpha)\gamma_*,1))^s}{1+\varepsilon^2},$$

which implies the curse of dimensionality. This shows that $\lim_{j\to\infty}\gamma_j = 0$ is a necessary condition for tractability.

Now suppose that $\lim_{j\to\infty}\gamma_j = 0$, but $\sum_{j=1}^{s}\gamma_j = \infty$. Then for growing $s$ we have

$$\prod_{j=1}^{s}\big(1+\min(2\zeta(\alpha)\gamma_j,1)\big) \geq 1 + \sum_{j=1}^{s}\min(2\zeta(\alpha)\gamma_j,1) \to \infty.$$

Hence $\lim_{s\to\infty} N_{\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}}(s,\varepsilon) = \infty$, which means that we cannot have strong polynomial tractability. Thus $\sum_{j=1}^{\infty}\gamma_j < \infty$ is a necessary condition for strong polynomial tractability.

Finally, suppose that $\lim_{j\to\infty}\gamma_j = 0$ and $\limsup_{s\to\infty}\sum_{j=1}^{s}\gamma_j/\log s = \infty$. Note that for $x \in [0, 2\zeta(\alpha)\sup_j \gamma_j]$ we have $\log(1+x) \geq cx$ for some $c > 0$. Hence

$$\log\prod_{j=1}^{s}\big(1+\min(2\zeta(\alpha)\gamma_j,1)\big) = \sum_{j=1}^{s}\log\big(1+\min(2\zeta(\alpha)\gamma_j,1)\big) \geq c\sum_{j=1}^{s}\min(2\zeta(\alpha)\gamma_j,1)$$

and therefore

$$\prod_{j=1}^{s}\left(1+\min(2\zeta(\alpha)\gamma_j,1)\right)\ge \exp\left(c\sum_{j=1}^{s}\min(2\zeta(\alpha)\gamma_j,1)\right)=s^{\frac{c}{\log s}\sum_{j=1}^{s}\min(2\zeta(\alpha)\gamma_j,1)}.$$

This implies that $N_{\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}}(s,\varepsilon)$ goes to infinity faster than any power of $s$ and we cannot have polynomial tractability. Hence we have shown that (4.23) is a necessary condition for polynomial tractability.

Summing up we have shown the following result.

**Corollary 4.28.** *Numerical integration in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ is*

1. *strongly polynomially tractable if and only if $\sum_{j=1}^{\infty}\gamma_j < \infty$. If $\lambda_0$ is the infimum over all $\lambda \in (1/\alpha, 1]$ such that (4.22) holds, then the $\varepsilon$-exponent $\tau^*$ of strong polynomial tractability is at most $2\lambda_0$;*
2. *polynomially tractable if and only if (4.23) holds.*

► **Remark 4.29** The above if and only if conditions for (strong) polynomial tractability also hold for more general algorithms than for QMC rules as considered here. This is clear for the sufficient conditions. For the necessary conditions this was proved by F.J. Hickernell and H. Woźniakowski in 2001.

## 4.6 Further Reading and Exercises

### Further Reading

Standard references for the theory of lattice point sets and lattice rules are the books of Niederreiter [69] and Sloan and Joe [85], both warmly recommended for further reading. Further books dealing with lattice rules are the ones by Hua and Wang [46] and by Korobov [55], the latter being available in Russian language only. For technical reasons we often restricted ourselves to prime $N$ within this section. We remark that most existence results also hold for arbitrary integers $N \ge 2$ and refer to [69] for further information. Theorem 4.5 was first shown by Joe in [51]. The proof of Theorem 4.6 can be found in [58]. The currently best existence result (4.5) for lattice point sets with low discrepancy was proved by Larcher [57] for $s = 2$ and by Bykovskii [11] for general $s$. The CBC construction of lattice point sets was introduced by Korobov [55] and later re-invented by Sloan and Reztsov [86]. More information on the fast CBC algorithm can be found in [78] and implementations of the fast CBC algorithm can be found in [77, 79]. Lattice rules for functions which are not necessarily periodic are discussed in the papers by Hickernell [36] and by Dick, Nuyens and Pillichshammer [23]. Our exposition of the proof of Theorem 4.22 is based on the proof of [90, Lemma 3.1] by Temlyakov. A proof of Theorem 4.22 for linear algorithms can be found in [20], see also Bakhvalov [6] and Temlyakov [90]. More information on tractability for the weighted Korobov space can be found in the paper by Sloan and Woźniakowski [88]. The current state of the art concerning tractability is summarized in the three volumes [74–76] by Novak and Woźniakowski (in particular, [75, Chapter 16] is devoted to multivariate integration for Korobov spaces). The curse of dimensionality for the integration of monotone functions and of convex functions was shown by Hinrichs, Novak and Woźniakowski [41] and for smooth functions by Hinrichs, Novak, Ullrich and Woźniakowski [39, 40]. The necessity of the conditions for (strong) polynomial tractability from Corollary 4.28 for arbitrary algorithms using $N$ function evaluations was shown by Hickernell and Woźniakowski [37].

A software tool called “Lattice Builder” (see [60]) for constructing lattice rules is available at: https://github.com/mungerd/latbuilder#readme

## Exercises

4.1 Consider lattice points of the form $\boldsymbol{v}_s(g) = (g_1, g_2, \ldots, g_s) \in \mathbb{Z}^s$, where $g_i \equiv g^{i-1} \pmod N$ with $g \in G_N$. Such generating vectors are called *Korobov vectors*. Define $R_N$ as in (4.3) and let $N$ be a prime number. Show that

$$\frac{1}{N-1} \sum_{g \in G_N} R_N(\boldsymbol{v}_s(g)) \le \frac{s-1}{N-1}(1+S_N)^s, \quad \text{where } S_N := \sum_{h \in C_1^*(N)} |h|^{-1}.$$

4.2 Let $N$ be a prime number. Show that for all $\varepsilon \in [0,1)$ there exist more than $\varepsilon |G_N|$ elements $g \in G_N$, such that

$$D_N(\mathcal{P}(\boldsymbol{v}_s(g), N)) \le \frac{s}{N} + \frac{1}{1-\varepsilon} \frac{s-1}{N-1}(1-S_N)^s.$$

4.3 Let $g : [0,1] \to \mathbb{R}$, $g(x) = x^2 - \frac{x}{2}$ if $x \in [0, 1/2]$ and $g(x) = \frac{3x}{2} - x^2 - \frac{1}{2}$ if $x \in [1/2, 1]$. Show that $g \in \mathcal{H}_{2,\gamma}$ but $g \notin C^2([0,1])$.

4.4 Show that the QMC mean square worst-case error in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ is

$$\mathbb{E}[e^2(\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathcal{P})] = \frac{1}{N}\left[-1 + \prod_{j=1}^{s}(1 + 2\gamma_j \zeta(\alpha))\right].$$

4.5 Prove the formula for the Fourier expansion of the Bernoulli polynomials given in Remark 4.19 for the second Bernoulli polynomial $B_2(x) = x^2 - x + \frac{1}{6}$.

4.6 Use the Fourier expansion of $B_\alpha(x)$ from Remark 4.19 to show the formula for $e^2(\mathcal{H}_{s,\alpha}, \mathcal{P}(\boldsymbol{g}, N))$ for all for even $\alpha \ge 2$ as stated in the same remark.

4.7 Use Exercise 3.14 to show that the initial error $e(0, s)$ in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ equals one.

4.8 Show that integration in $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$ is polynomially tractable if (4.23) holds.

4.9 Besides polynomial and strong polynomial tractability, there are many other notions of tractability such as, e.g., *weak tractability*. We say that we have weak tractability if the information complexity $N_{\mathcal{H}(K)}(\varepsilon, s)$ satisfies

$$\lim_{\varepsilon^{-1}+s \to \infty} \frac{\log N_{\mathcal{H}(K)}(\varepsilon, s)}{\varepsilon^{-1}+s} = 0.$$

Show that weak tractability means that $N_{\mathcal{H}(K)}(\varepsilon, s)$ is asymptotically much smaller than $q^{\varepsilon^{-1}+s}$ for any $q > 1$. *Remark:* Problems that are not weakly tractable are said to be *intractable*. This means that the information complexity depends exponentially on $\varepsilon^{-1}$ or $s$, whereas the curse of dimensionality means that the information complexity depends exponentially on $s$.

4.10 Consider integration in the weighted Korobov space $\mathcal{H}_{s,\alpha,\boldsymbol{\gamma}}$. Show that a sufficient condition for weak tractability is $\lim_{s\to\infty} \frac{1}{s} \sum_{j=1}^{s} \gamma_j = 0$. *Remark:* This condition is also necessary.

# $(t, m, s)$-Nets and $(t, s)$-Sequences 5

## 5.1 Definition and Existence

We are interested in point sets with very low star discrepancy. This means that we aim on finding point sets $\mathcal{P}$ for which the absolute local discrepancy,

$$|\Delta_{\mathcal{P},N}(\boldsymbol{y})| = \left| \frac{A([\boldsymbol{0}, \boldsymbol{y}), \mathcal{P}, N)}{N} - \lambda_s([\boldsymbol{0}, \boldsymbol{y})) \right|,$$

is as small as possible for all $\boldsymbol{y} \in (0,1]^s$. Our strategy for achieving this goal is to discretize the problem and to investigate point sets $\mathcal{P}$ for which

$$A(J, \mathcal{P}, N) = N\lambda_s(J) \tag{5.1}$$

for all $J$ from a sufficiently large class of intervals $J$, and hope that this already implies a low star discrepancy of $\mathcal{P}$. If (5.1) holds, then we say that $\mathcal{P}$ is *fair* with respect to $J$. It is obvious that a given point set $\mathcal{P}$ cannot be fair with respect to all intervals since, for example, for any $N$-element point set $\mathcal{P}$ in $[0,1)^s$ we have $D_N^*(\mathcal{P}) \geq (2N)^{-1}$.

Now we try to find classes of intervals with the property that fairness of a point set $\mathcal{P}$ with respect to all intervals from this class implies low star discrepancy of $\mathcal{P}$.

**1. Attempt:** On first sight, it is near at hand to consider the class of intervals of the form

$$J = \prod_{j=1}^{s} \left[ \frac{a_j}{b}, \frac{a_j+1}{b} \right), \tag{5.2}$$

where $a_j \in \{0, 1, \ldots, b-1\}$ for $j = 1, \ldots, s$. Assume that we are given a fair point set $\mathcal{P}$ with respect to these intervals. This means that each of the intervals of the form (5.2) contains the same number of elements from $\mathcal{P}$, and hence the

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_5,

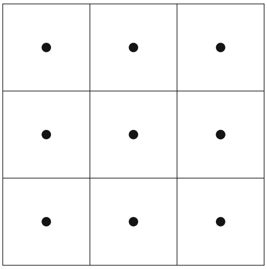

**Fig. 5.1** The point set $\Gamma_{3,2}$ with fair intervals

cardinality of $\mathcal{P}$ is a multiple of $b^s$. Here we observe that the regular lattice $\Gamma_{b,s}$ also satisfies these requirements (see Fig. 5.1). But the discrepancy of $\Gamma_{b,s}$ is of order $1/b$, which is rather large. Hence fairness with respect to the class of intervals of the form (5.2) does not necessarily lead to small discrepancy. This means that this class is in some sense too small to achieve our goal.

**2. Attempt:** We have seen that we have to demand fairness for a larger and, in some sense, finer class of intervals to find point sets with low discrepancy. This class will consist of intervals of a special shape which we define now:

**Definition 5.1**

Let $b \in \mathbb{N}$, $b \geq 2$. An *elementary interval* in base $b$ is an interval of the form

$$J = \prod_{j=1}^{s} \left[ \frac{a_j}{b^{d_j}}, \frac{a_j + 1}{b^{d_j}} \right),$$

where $d_j \in \mathbb{N}_0$ and $a_j \in \{0, 1, \ldots, b^{d_j} - 1\}$ for all $j = 1, 2, \ldots, s$.

Now we consider finite point sets which are fair with respect to all elementary intervals of prescribed volume. This leads to the following definition:

**Definition 5.2**

Let $m, s, b \in \mathbb{N}$, $b \geq 2$. A $(0, m, s)$-*net in base* $b$ is a $b^m$-element point set in $[0, 1)^s$ which is fair with respect to all $s$-dimensional elementary intervals in base $b$ having volume $b^{-m}$.

Let us provide an example.

## Example 5.3

A $(0, 4, 2)$-net in base 2 is a $2^4 = 16$-element point set $\mathcal{P}$ in $[0, 1)^2$ for which every 2-dimensional elementary interval in base 2 of area $2^{-4} = 1/16$ contains exactly one element of $\mathcal{P}$ (see Fig. 5.2).

We will show later that $(0, m, s)$-nets are in fact point sets with low discrepancy. Before we do so, we should be concerned with the existence of $(0, m, s)$-nets, i.e.,

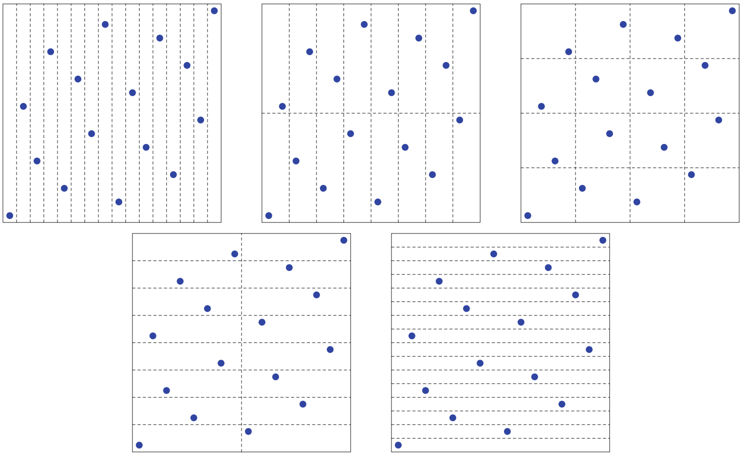

**Fig. 5.2** A $(0, 4, 2)$-net in base 2; every 2-dimensional elementary interval in base 2 of area $2^{-4}$ contains exactly one point

with the question for which parameters $m, s, b \in \mathbb{N}$, where $b \geq 2$, a $(0, m, s)$-net in base $b$ exists. For the moment we will only show a necessary condition. As we will see, this condition is independent of the choice of $m$, but depends on a relation between $s$ and $b$. A sufficient condition, and even an explicit construction, for the case of prime power bases $b$ will be provided in Sect. 5.4.

We prove a series of lemmas which finally lead to the necessary condition for the existence of $(0, m, s)$-nets in base $b$. In the first lemma we show that from a given $(0, m, s)$-net in base $b$ one can always construct a $(0, m, r)$-net in base $b$ of dimension $r \leq s$. For $\boldsymbol{x} = (x_1, \ldots, x_s)$ and for $r \in \{1, \ldots, s\}$ let $\pi_r(\boldsymbol{x}) := (x_1, \ldots, x_r)$ be the projection of $\boldsymbol{x}$ onto the first $r$ components of $\boldsymbol{x}$. For a point set $\mathcal{P}$, the projection $\pi_r(\mathcal{P})$ is defined element-wise.

**Lemma 5.4.** *Let $\mathcal{P}$ be a $(0, m, s)$-net in base $b$. Then for every $r \in \{1, \ldots, s\}$ the point set $\pi_r(\mathcal{P})$ is a $(0, m, r)$-net in base $b$.*

*Proof.* Let $J'$ be an $r$-dimensional elementary interval in base $b$ of volume $b^{-m}$. Then the interval $J = J' \times [0, 1)^{s-r}$ is an $s$-dimensional elementary interval in base $b$ which also has volume $b^{-m}$. By the $(0, m, s)$-net property, $J$ contains exactly one element of $\mathcal{P}$ and therefore $J'$ contains exactly one element of $\mathcal{P}'$ as well. □

**Lemma 5.5.** *If a $(0, m, s)$-net in base $b$ with $m \geq 2$ exists, then there exists a $(0, 2, s)$-net in base $b$.*

*Proof.* Let $\mathcal{P}$ be a $(0, m, s)$-net in base $b$. Then the interval

$$J = \left[0, \frac{1}{b^{m-2}}\right) \times \prod_{i=2}^{s}[0,1) = \bigcup_{k=0}^{b^2-1}\left(\left[\frac{k}{b^m}, \frac{k+1}{b^m}\right) \times [0,1)^{s-1}\right)$$

contains exactly $b^2$ elements of $\mathcal{P}$. We denote these elements by $z_0, \ldots, z_{b^2-1}$. Multiplying the first component of each of these $b^2$ points by $b^{m-2}$ gives a point set $\{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{b^2-1}\}$ in $[0,1)^s$. We show that this point set is a $(0, 2, s)$-net in base $b$. To this end we have to prove that every elementary interval of the form

$$E = \prod_{j=1}^{s}\left[\frac{a_j}{b^{d_j}}, \frac{a_j+1}{b^{d_j}}\right)$$

of volume $b^{-2}$ contains exactly one element of the set $\{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{b^2-1}\}$. The number of indices $n \in \{0, 1, \ldots, b^2-1\}$ satisfying $\boldsymbol{y}_n \in E$ equals the number of indices $n \in \{0, 1, \ldots, b^2-1\}$, for which $z_n \in E'$, where

$$E' = \underbrace{\left[\frac{a_1}{b^{d_1}b^{m-2}}, \frac{a_1+1}{b^{d_1}b^{m-2}}\right)}_{\subseteq[0,b^{-m+2})} \times \prod_{j=2}^{s}\left[\frac{a_j}{b^{d_j}}, \frac{a_j+1}{b^{d_j}}\right).$$

This, however, is exactly the number of points of $\mathcal{P}$ which belong to the interval $E'$. Since $\mathcal{P}$ is a $(0, m, s)$-net in base $b$ and since $E'$ is an elementary interval of volume $b^{-m}$, this number is exactly one and the proof is finished. □

**Lemma 5.6.** *A* $(0, 2, b+2)$*-net in base b cannot exist.*

*Proof.* Assume that there exists a $(0, 2, b+2)$-net $\mathcal{P} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{b^2-1}\}$ in base $b$. Every point $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,b+2})$ in $\mathcal{P}$ corresponds to a $(b+2)$-dimensional vector

$$\boldsymbol{x}_n \leftrightarrow \begin{pmatrix} a_n^{(1)} \\ \vdots \\ a_n^{(b+2)} \end{pmatrix},$$

where $a_n^{(j)} = \lfloor b x_{n,j} \rfloor \in \{0, 1, \ldots, b-1\}$ is chosen so that $\boldsymbol{x}_n$ belongs to the interval $\prod_{j=1}^{b+2}[a_n^{(j)}/b, (a_n^{(j)}+1)/b)$. Hence the $b^2$ elements of $\mathcal{P}$ correspond to the following $(b+2) \times b^2$ array:

$$\begin{array}{cccc} \boldsymbol{x}_0 & \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_{b^2-1} \\ \updownarrow & \updownarrow & & \updownarrow \\ a_0^{(1)} & a_1^{(1)} & \ldots & a_{b^2-1}^{(1)} \\ \vdots & \vdots & & \vdots \\ a_0^{(b+2)} & a_1^{(b+2)} & \ldots & a_{b^2-1}^{(b+2)} \end{array} \tag{5.3}$$

We prove that this array satisfies the following “orthogonality property”: For any two rows of this array, say

$$\begin{array}{cccc} a_0^{(i)} & a_1^{(i)} & \ldots & a_{b^2-1}^{(i)}, \\ a_0^{(j)} & a_1^{(j)} & \ldots & a_{b^2-1}^{(j)}, \end{array}$$

where $i \neq j$, the $b^2$ two-dimensional columns

$$\begin{pmatrix} a_k^{(i)} \\ a_k^{(j)} \end{pmatrix}_{k=0,\ldots,b^2-1}$$

attain any possible value $\binom{A}{B}$, with $A, B \in \{0, \ldots, b-1\}$, exactly once.

Assume to the contrary that some $\binom{A}{B}$ appears more than once. Then the elementary interval

$$[0,1)^{i-1} \times \left[\frac{A}{b}, \frac{A+1}{b}\right) \times [0,1)^{j-i-1} \times \left[\frac{B}{b}, \frac{B+1}{b}\right) \times [0,1)^{b+2-j}$$

of volume $b^{-2}$ does not contain exactly one element of $\mathcal{P}$, which contradicts the $(0, 2, b+2)$-net property in base $b$ of $\mathcal{P}$. Hence the orthogonality property is proven.

In particular, any possible value $A \in \{0, \ldots, b-1\}$ must occur in any row of (5.3) exactly $b$-times. This, however, cannot be satisfied for all possible pairs of rows of (5.3) as we show now:

Assume that any two of the $b+2$ rows of (5.3) satisfy the orthogonality property. We may assume that in array (5.3) the values of the first column all equal 1, since a permutation of the values $A \in \{0, \ldots, b-1\}$ in a single row of the array does not affect the orthogonality property.

Then, by the orthogonality property, in any of the remaining $b^2 - 1$ columns, 1 can occur at most once. But in each row, 1 must occur $b$-times, and hence we would require a place for the $(b-1)(b+2)$ remaining 1’s in these $b^2 - 1$ columns. Since $(b-1)(b+2) = b^2 + b - 2 > b^2 - 1$, we have a contradiction. □

Now we come to the desired necessary condition for the existence of $(0, m, s)$-nets in base $b$.

**Theorem 5.7 (Niederreiter).** *Let $s, b \in \mathbb{N}$, $b \geq 2$. Assume that for $m \in \mathbb{N}$ with $m \geq 2$ there exists a $(0, m, s)$-net in base $b$. Then $s \leq b + 1$.*

*Proof.* Assume that a $(0, m, s)$-net in base $b$ where $m \geq 2$ and $s \geq b + 2$ exists. Then, according to Lemmas 5.4 and 5.5, there exists a $(0, 2, b + 2)$-net in base $b$, which contradicts Lemma 5.6. □

In other words, a $(0, m, s)$-net in base $b$ with $m \geq 2$ cannot exist as long as $s \geq b + 2$. For example, there is no $(0, m, s)$-net in base 2 for $m \geq 2$ and $s \geq 4$.

As $(0, m, s)$-nets in base $b$ do not exist for all parameters $s, b \in \mathbb{N}$, $b \geq 2$, we now weaken the condition from their definition.

**Definition 5.8**

Let $m, s, b \in \mathbb{N}$, $b \geq 2$, and let $t \in \{0, \ldots, m\}$. A *$(t, m, s)$-net in base $b$* is a $b^m$-element point set $\mathcal{P}$ in $[0, 1)^s$ which is fair with respect to every $s$-dimensional elementary interval in base $b$ having volume $b^{-m+t}$. The parameter $t$ is called the *quality parameter* of the net. Moreover, $\mathcal{P}$ is called a *strict* $(t, m, s)$-net in base $b$, if $t$ is the smallest number $u \in \{0, \ldots, m\}$ such that $\mathcal{P}$ is a $(u, m, s)$-net in base $b$.

Some remarks on the definition of $(t, m, s)$-nets in base $b$ are appropriate:

**Remark 5.9**

1. An $s$-dimensional elementary interval in base $b$ of volume $b^{-m+t}$ is an interval of the form

$$E = \prod_{i=1}^{s} \left[ \frac{a_i}{b^{d_i}}, \frac{a_i + 1}{b^{d_i}} \right),$$

   where $d_1 + \cdots + d_s = m - t$. Every such interval contains exactly $b^t$ elements of a $(t, m, s)$-net in base $b$. The larger $t$ is, the larger are the intervals under consideration and the smaller is the class of considered intervals. For such intervals we have $A(E, \mathcal{P}, b^m) = b^t = b^m \lambda_s(E)$.
2. Every $b^m$-element point set in $[0, 1)^s$ is a $(t, m, s)$-net in base $b$ for some $t \in \{0, \ldots, m\}$. In the worst case it is a $(m, m, s)$-net in base $b$.
3. Every $(0, m, s)$-net in base $b$ is strict by definition.
4. Every elementary interval in base $b$ of volume $b^{-m+t}$ is a disjoint union of $b^t$ elementary intervals in base $b$ of volume $b^{-m}$. Hence, every $(0, m, s)$-net in base $b$ is also a $(t, m, s)$-net in base $b$ for every $t \in \{0, \ldots, m\}$.
5. Similarly, every $(t, m, s)$-net in base $b$ is also a $(u, m, s)$-net in base $b$ for every $u \in \{t, \ldots, m\}$.
6. From Lemma 5.5 we know that from a $(0, m, s)$-net in base $b$ with $m \geq 2$ one can construct a $(0, 2, s)$-net in base $b$. More generally, but in the same way, from

a $(t, m, s)$-net in base $b$ one can construct a $(t, u, s)$-net in base $b$ for every $u \in \{t, \dots, m\}$.

7. From Lemma 5.4 we know that from a $(0, m, s)$-net in base $b$ a $(0, m, r)$-net in base $b$ with $r \in \{1, \dots, s\}$ can be constructed. More generally, but in the same way, from a $(t, m, s)$-net in base $b$ one can construct a $(t, m, r)$-net in base $b$ for any $r \in \{1, \dots, s\}$.
8. It can be shown that, for all $s, b \in \mathbb{N}$, $b \geq 2$, there exists a number $t_b(s) \in \mathbb{N}_0$ such that for every $m \geq t_b(s)$ there exists a $(t_b(s), m, s)$-net in base $b$. It is a remarkable result by H. Niederreiter and C.P. Xing that $t_b(s) \asymp_b s$ asymptotically for $s \to \infty$.

## Example 5.10

The regular lattice $\Gamma_{n,s}$ as defined in (2.7) with $n = b^L$ (i.e., with $b^{sL}$ elements) is a strict $(sL - L, sL, s)$-net in base $b$.

*Proof.* The structure of the regular grid $\Gamma_{b^L,s}$ is such that every interval of the form

$$M_{\boldsymbol{k}} = \prod_{j=1}^{s} \left[ \frac{k_j}{b^L}, \frac{k_j + 1}{b^L} \right),$$

where $\boldsymbol{k} = (k_1, \dots, k_s) \in \{0, \dots, b^L - 1\}^s$, contains exactly one element. Let

$$J = \prod_{j=1}^{s} \left[ \frac{a_j}{b^{d_j}}, \frac{a_j + 1}{b^{d_j}} \right)$$

be an elementary interval of volume $b^{-L}$. Then we can write $J$ as a disjoint union of intervals $M_{\boldsymbol{k}}$ in the following way:

$$J = \bigcup_{k_1 = b^{L-d_1} a_1}^{b^{L-d_1}(a_1+1)-1} \cdots \bigcup_{k_s = b^{L-d_s} a_s}^{b^{L-d_s}(a_s+1)-1} M_{\boldsymbol{k}} \quad \text{with } d_1 + \dots + d_s = L.$$

Since each of the intervals $M_{\boldsymbol{k}}$ contains exactly one element of the regular lattice, it follows that $J$ contains exactly $b^{sL-L}$ elements of the regular lattice. This shows that the regular lattice with $b^{sL}$ elements is an $(sL - L, sL, s)$-net in base $b$.

It remains to show strictness. To this end we consider the elementary interval

$$J' = \left[ 0, \frac{1}{b^{L+1}} \right) \times [0, 1)^{s-1}.$$

This interval has volume $b^{-(L+1)}$ and contains no elements of the regular lattice. Therefore $A(J', \Gamma_{b^L,s}, b^{sL}) - b^{sL} \lambda_s(B) = -b^{(s-1)L-1} \neq 0$. □

A disadvantage of nets in base $b$ is that the number of points is restricted to a power of $b$. At first glance one could argue that we can always choose $b$ arbitrarily large and $m = 1$, which would mean that there is no restriction at all. However, it is intuitively obvious (and is supported by the discrepancy estimates in Sect. 5.2) that the structure of a $(t,m,s)$-net in base $b$ becomes strong only if $m$ is large compared to $b$. Hence, for a given number $N$ of points, it is sometimes better to realize the point set with a small base $b$, i.e., a larger value for $m$, and with a suboptimal quality parameter $t$, instead of choosing a large base $b$ (e.g., $b = N$) (and therefore a small $m$, for instance $m = 1$), with optimal quality parameter.

To overcome this problem one considers infinite sequences which have a certain "net structure" in the sense that the elements of certain finite subsequences form $(t,m,s)$-nets in base $b$. This leads to the definition of $(t,s)$-sequences.

**Definition 5.11**

Let $s, b \in \mathbb{N}$, $b \geq 2$, and let $t \in \mathbb{N}_0$. An infinite sequence $(\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$ of points in $[0,1)^s$ is called a $(t,s)$-*sequence in base* $b$ if for all integers $m > t$ and $k \geq 0$, the point set consisting of the points $\boldsymbol{x}_{kb^m}, \boldsymbol{x}_{kb^m+1}, \ldots, \boldsymbol{x}_{kb^m+b^m-1}$ forms a $(t,m,s)$-net in base $b$. The parameter $t$ is called the *quality parameter* of the $(t,s)$-sequence.

**Definition 5.12**

A $(t,s)$-sequence in base $b$ with $t \geq 1$ is called a *strict* $(t,s)$-*sequence in base* $b$ if it is not a $(t-1,s)$-sequence in base $b$. A $(0,s)$-sequence is called strict by definition.

We give a first example.

## Example 5.13

The van der Corput sequence in base $b$ is a $(0,1)$-sequence in base $b$. This follows from the proof of Proposition 2.10, where we have shown that for every $m \in \mathbb{N}_0$ every elementary interval in base $b$ of length $b^{-m}$ contains exactly one of $b^m$ consecutive elements of the van der Corput sequence in base $b$.

We have shown in Proposition 2.10 that the van der Corput sequence is uniformly distributed modulo one. This result can be generalised in the following sense.

**Theorem 5.14.** *A $(t,s)$-sequence in base $b$ is uniformly distributed modulo one.*

For the proof of this result we require the following lemma.

**Lemma 5.15.** *Let $\mathcal{P}$ be a $(t,m,s)$-net in base $b$. Let $J_1,\ldots,J_r$ be disjoint $s$-dimensional elementary intervals in base $b$ of volume $b^{t-m}$ and let $J = J_1 \cup \ldots \cup J_r$. Then*

$$A(J,\mathcal{P},b^m) = b^m \lambda_s(J).$$

*Proof.* For every $J_l$ we have $A(J_l,\mathcal{P},b^m) = b^t = b^m\lambda_s(J_l)$, and hence

$$A(J,\mathcal{P},b^m) = \sum_{l=1}^{r} A(J_l,\mathcal{P},b^m) = b^m \sum_{l=1}^{r} \lambda_s(J_l) = b^m\lambda_s(J).$$

□

*Proof of Theorem 5.14.* Let $\mathcal{S} = (\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ be a $(t,s)$-sequence in base $b$. Further, let $J := \prod_{j=1}^{s}[\alpha_j,\beta_j)$ with $0 \le \alpha_j < \beta_j \le 1$, be an arbitrary subinterval of $[0,1)^s$, and let $\varepsilon > 0$ be given.

Let $r \in \mathbb{N}$ be fixed and such that $2sb^{-r} < \varepsilon/2$, let $l = rs$ and let $m$ be fixed such that $m - t \ge l$. Let $A_j, B_j \in \{0,1,\ldots,b^r - 1\}$ be such that

$$\tfrac{A_j}{b^r} \le \alpha_j < \tfrac{A_j+1}{b^r} \quad \text{and} \quad \tfrac{B_j}{b^r} \le \beta_j < \tfrac{B_j+1}{b^r} \quad \text{for } j = 1,\ldots,s.$$

Then for

$$J_1 := \prod_{j=1}^{s}\left[\frac{A_j+1}{b^r},\frac{B_j}{b^r}\right) \quad \text{and} \quad J_2 := \prod_{j=1}^{s}\left[\frac{A_j}{b^r},\frac{B_j+1}{b^r}\right)$$

we have

$$J_1 \subseteq J \subseteq J_2 \subseteq [0,1)^s,$$

and, by Lemma 2.16, $\lambda_s(J_2 \setminus J_1) \le 2sb^{-r}$. Furthermore,

$$J_1 = \bigcup_{a_1=A_1+1}^{B_1-1} \cdots \bigcup_{a_s=A_s+1}^{B_s-1} \prod_{j=1}^{s}\left[\frac{a_j}{b^r},\frac{a_j+1}{b^r}\right),$$

i.e., $J_1$ is a union of at most $b^l$ elementary intervals of volume $b^{-rs} = b^{-l}$. In the same way, $J_2$ is a union of at most $b^l$ elementary intervals of volume $b^{-rs} = b^{-l}$. Hence, by Lemma 5.15 and since $m - t \ge l$, the point sets $\{\boldsymbol{x}_{kb^m},\boldsymbol{x}_{kb^m+1},\ldots,\boldsymbol{x}_{kb^m+b^m-1}\}$ are fair with respect to the intervals $J_1$ and $J_2$. Therefore, for all $N \in \mathbb{N}$, we have

$$\begin{aligned}A(J,\mathcal{S},N) - N\lambda_s(J) &\leq A(J_2,\mathcal{S},N) - N\lambda_s(J_2) + N\frac{2s}{b^r}\\ &\leq A(J_2,\mathcal{S},\lfloor Nb^{-m}\rfloor b^m) - \lfloor Nb^{-m}\rfloor\, b^m\lambda_s(J_2) + b^m + N\frac{2s}{b^r}\\ &= b^m + N\frac{2s}{b^r}\end{aligned}$$

and

$$\begin{aligned}A(J,\mathcal{S},N) - N\lambda_s(J) &\geq A(J_1,\mathcal{S},N) - N\lambda_s(J_1) - N\frac{2s}{b^r}\\ &\geq A(J_1,\mathcal{S},\lceil Nb^{-m}\rceil b^m) - \lceil Nb^{-m}\rceil\, b^m\lambda_s(J_1) - b^m - N\frac{2s}{b^r}\\ &= -b^m - N\frac{2s}{b^r},\end{aligned}$$

so that

$$\left|\frac{A(J,\mathcal{S},N)}{N} - \lambda_s(J)\right| \leq \frac{b^m}{N} + \frac{2s}{b^r} < \varepsilon$$

for $N$ large enough. Hence the result follows. □

As for $(0, m, s)$-nets in base $b$, it is clear that a $(0, s)$-sequence in base $b$ cannot exist for all dimensions $s$. A necessary condition for the existence of $(0, s)$-sequences in base $b$ can be obtained from Theorem 5.7.

**Theorem 5.16 (Niederreiter).** *Let $s, b \in \mathbb{N}$, $b \geq 2$. The existence of a $(0, s)$-sequence in base b implies $s \leq b$.*

For the proof of Theorem 5.16 we need the following lemma, which shows that if a $(t, s)$-sequence in base $b$ exists, then, for every $m \geq t$, also a $(t, m, s + 1)$-net in base $b$ exists.

**Lemma 5.17.** *Let $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ be a $(t, s)$-sequence in base $b$. Then, for every $m$, the point set $\{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{b^m-1}\}$ with $\boldsymbol{y}_k := (kb^{-m}, \boldsymbol{x}_k)$, is a $(t, m, s + 1)$-net in base $b$.*

*Proof.* Let $J = \prod_{j=1}^{s+1}\left[\frac{A_j}{b^{d_j}}, \frac{A_j+1}{b^{d_j}}\right)$ be an elementary interval of volume $b^{-m+t}$. Then $\boldsymbol{y}_k \in J$ if and only if

$$\frac{k}{b^m} \in \left[\frac{A_1}{b^{d_1}}, \frac{A_1+1}{b^{d_1}}\right) \quad \text{and} \quad \boldsymbol{x}_k \in \prod_{j=2}^{s+1} \left[\frac{A_j}{b^{d_j}}, \frac{A_j+1}{b^{d_j}}\right).$$

The first condition implies

$$A_1 b^{m-d_1} \leq k < A_1 b^{m-d_1} + b^{m-d_1}.$$

Since $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ is a $(t,s)$-sequence in base $b$, the points $\boldsymbol{x}_{A_1 b^{m-d_1}+l}$ for $l = 0,\ldots,b^{m-d_1}-1$, form a $(t, m-d_1, s)$-net in base $b$. The interval

$$\prod_{j=2}^{s+1} \left[\frac{A_j}{b^{d_j}}, \frac{A_j+1}{b^{d_j}}\right)$$

has volume $b^{-d_2-\cdots-d_{s+1}} = b^{-m+d_1+t}$ and therefore contains exactly $b^t$ of the points $\boldsymbol{x}_{A_1 b^{m-d_1}+l}$ for $l = 0,\ldots,b^{m-d_1}-1$. Consequently $J$ contains exactly $b^t$ of the $\boldsymbol{y}_k$'s and the result follows. □

*Proof of Theorem 5.16.* Assume that there exists a $(0,s)$-sequence in base $b$. Then, according to Lemma 5.17, there also exists a $(0, m, s+1)$-net in base $b$. Then it follows from Theorem 5.7 that $s+1 \leq b+1$, and hence $s \leq b$. □

## 5.2 Star Discrepancy of $(t, m, s)$-Nets and $(t, s)$-Sequences

In this section we will show that $(t,m,s)$-nets and $(t,s)$-sequences defined above have in fact the desired property of having low discrepancy.

First we consider $(t,m,s)$-nets in base $b$, for which the following fundamental discrepancy estimate holds.

**Theorem 5.18.** *For the star discrepancy of a $(t,m,s)$-net $\mathcal{P}$ in base $b$ we have*

$$D_{b^m}^*(\mathcal{P}) \leq \frac{1}{b^{m-t}} \sum_{k=0}^{s-1} \binom{m-t}{k} (b-1)^k.$$

For the proof of this estimate we require the following lemma.

**Lemma 5.19.** *For $s, m \in \mathbb{N}$ we have*

$$\sum_{\substack{k_1,\ldots,k_s\in\mathbb{N}\\ k_1+\cdots+k_s\leq m}} 1 = \binom{m}{s}.$$

*Proof.* We have

$$\sum_{\substack{k_1,\ldots,k_s\in\mathbb{N}\\ k_1+\cdots+k_s\leq m}} 1 = \sum_{l=s}^{m} \sum_{\substack{k_1,\ldots,k_s\in\mathbb{N}\\ k_1+\cdots+k_s=l}} 1.$$

For $l \geq s$ the number of solutions of the equation $k_1 + \cdots + k_s = l$ with $k_i \in \mathbb{N}$ equals the number of solutions of the equation $k_1 + \cdots + k_s = l - s$ with $k_i \in \mathbb{N}_0$, which in turn equals $\binom{l-1}{s-1}$. consequently,

$$\sum_{\substack{k_1,\ldots,k_s\in\mathbb{N}\\ k_1+\cdots+k_s\leq m}} 1 = \sum_{l=s}^{m} \binom{l-1}{s-1} = \binom{m}{s},$$

where the last equality can be shown by induction on $m$. □

*Proof of Theorem 5.18.* We only consider the case $t = 0$; the general case can be proven along the same lines by considering the same $b$-adic expansions, only up to $m - t$ instead of $m$.

Thus, let $\mathcal{P}$ be a $(0, m, s)$-net in base $b$. Consider an interval $B = [0, \alpha^{(1)}) \times \cdots \times [0, \alpha^{(s)})$ with

$$\alpha^{(j)} = \frac{a_1^{(j)}}{b} + \frac{a_2^{(j)}}{b^2} + \cdots + \frac{a_m^{(j)}}{b^m} + \cdots$$

for $j = 1, 2, \ldots, s$. The fundamental idea of the proof is to approximate the interval $B$ from the interior and exterior by sets $B_s$ and $B_s'$, respectively, which can be written as disjoint unions of elementary intervals of volume $b^{-m}$. Then we have $B_s \subseteq B \subseteq B_s'$ such that

$$A(B_s, \mathcal{P}, b^m) \leq A(B, \mathcal{P}, b^m) \leq A(B_s', \mathcal{P}, b^m)$$

and, by Lemma 5.15,

$$b^{-m} A(B_s, \mathcal{P}, b^m) = \lambda_s(B_s) \leq \lambda_s(B) \leq \lambda_s(B_s') = b^{-m} A(B_s', \mathcal{P}, b^m).$$

Thus we get

$$|b^{-m} A(B, \mathcal{P}, b^m) - \lambda_s(B)| \leq \lambda_s(B_s') - \lambda_s(B_s) = \lambda_s(B_s' \backslash B_s),$$

and the theorem will be proven once we succeed in showing that $B_s$ and $B_s'$ can always be constructed in a way that guarantees

$$\lambda_s(B_s' \backslash B_s) \leq b^{-m} \sum_{k=0}^{s-1} \binom{m}{k} (b-1)^k.$$

Of course, this is the hard part of the proof. We use induction on $s$.

- $s = 1$: Let

  $$B_1 = \left[0, \frac{a_1^{(1)}}{b} + \cdots + \frac{a_m^{(1)}}{b^m}\right) \text{ and } B_1' = \left[0, \frac{a_1^{(1)}}{b} + \cdots + \frac{a_m^{(1)}}{b^m} + \frac{1}{b^m}\right).$$

  Then

  $$B_1 = \bigcup_{k=0}^{b^{m-1}a_1^{(1)}+\cdots+a_m^{(1)}-1} \left[\frac{k}{b^m}, \frac{k+1}{b^m}\right)$$

  is a disjoint union of elementary intervals of length $b^{-m}$ and the same is true for $B_1'$. Furthermore, we have $B_1 \subseteq B \subseteq B_1'$ and $\lambda_1(B_1' \setminus B_1) = \frac{1}{b^m} = \frac{1}{b^m} \sum_{k=0}^{0} \binom{m}{k}(b-1)^k$.
- For illustration purposes we give the step $1 \to 2$ separately before showing the general induction step: Let

  $$I_1 := \left[0, \frac{a_1^{(1)}}{b}\right) \times \left[0, \frac{a_1^{(2)}}{b} + \cdots + \frac{a_{m-1}^{(2)}}{b^{m-1}}\right)$$
  $$J_1 := \left[0, \frac{a_1^{(1)}}{b}\right) \times \left[0, \frac{a_1^{(2)}}{b} + \cdots + \frac{a_{m-1}^{(2)}}{b^{m-1}} + \frac{1}{b^{m-1}}\right).$$

  Then $I_1 \subseteq B$ and

  $$I_1 = \bigcup_{k=0}^{b^{m-2}a_1^{(2)}+\cdots+a_{m-1}^{(2)}-1} \left[0, \frac{a_1^{(1)}}{b}\right) \times \left[\frac{k}{b^{m-1}}, \frac{k+1}{b^{m-1}}\right).$$

  If $a_1^{(1)} \neq 0$, then $I_1$ is a disjoint union of two-dimensional elementary intervals of area $b^{-m}$. If, however, $a_1^{(1)} = 0$, then $I_1 = \emptyset$ and is therefore trivially a disjoint union of elementary intervals of area $b^{-m}$. For the same reasons $J_1$ is a disjoint union of elementary intervals of area $b^{-m}$.

  For $k \in \{2, \ldots, m-1\}$ put

  $$I_k := \left[\frac{a_1^{(1)}}{b} + \cdots + \frac{a_{k-1}^{(1)}}{b^{k-1}}, \frac{a_1^{(1)}}{b} + \cdots + \frac{a_k^{(1)}}{b^k}\right) \times \left[0, \frac{a_1^{(2)}}{b} + \cdots + \frac{a_{m-k}^{(2)}}{b^{m-k}}\right),$$
  $$J_k := \left[\frac{a_1^{(1)}}{b} + \cdots + \frac{a_{k-1}^{(1)}}{b^{k-1}}, \frac{a_1^{(1)}}{b} + \cdots + \frac{a_k^{(1)}}{b^k}\right) \times \left[0, \frac{a_1^{(2)}}{b} + \cdots + \frac{a_{m-k}^{(2)}}{b^{m-k}} + \frac{1}{b^{m-k}}\right)$$

and

$$I_m := \left[\frac{a_1^{(1)}}{b} + \cdots + \frac{a_{m-1}^{(1)}}{b^{m-1}}, \frac{a_1^{(1)}}{b} + \cdots + \frac{a_m^{(1)}}{b^m}\right) \times [0, 0) = \emptyset,$$

$$J_m := \left[\frac{a_1^{(1)}}{b} + \cdots + \frac{a_{m-1}^{(1)}}{b^{m-1}}, \frac{a_1^{(1)}}{b} + \cdots + \frac{a_m^{(1)}}{b^m}\right) \times [0, 1).$$

Again it can be checked that $I_k, J_k$ are disjoint unions of elementary intervals of area $b^{-m}$.

Now let

$$B_2 := \bigcup_{k=1}^{m} I_k \quad \text{and} \quad B_2' := \bigcup_{k=1}^{m} J_k \cup (B_1' \setminus B_1 \times [0, 1)),$$

where $B_1$ and $B_1'$ are defined as in the case for $s = 1$ to approximate the interval $[0, \alpha^{(1)})$, such that

$$B_1' \setminus B_1 = \left[\frac{a_1^{(1)}}{b} + \cdots + \frac{a_m^{(1)}}{b^m}, \frac{a_1^{(1)}}{b} + \cdots + \frac{a_m^{(1)}}{b^m} + \frac{1}{b^m}\right)$$

and $\lambda_1(B_1' \setminus B_1) = \frac{1}{b^m}$. Then we have $B_2 \subseteq B \subseteq B_2'$ and

$$\begin{aligned}
\lambda_2(B_2' \setminus B_2) &= \sum_{k=1}^{m} \lambda_2(J_k \setminus I_k) + \lambda_1(B_1' \setminus B_1) \\
&= \sum_{k=1}^{m} \frac{a_k^{(1)}}{b^k} \frac{1}{b^{m-k}} + \frac{1}{b^m} \\
&\leq \frac{m}{b^m}(b-1) + \frac{1}{b^m} \\
&= \frac{1}{b^m} \sum_{k=0}^{1} \binom{m}{k} (b-1)^k.
\end{aligned}$$

- $s - 1 \to s$: For $k_1, \ldots, k_{s-1} \in \mathbb{N}$ with $k_1 + \cdots + k_{s-1} \leq m$ put

$$I_{k_1,\ldots,k_{s-1}} := \prod_{l=1}^{s-1} \left[\frac{a_1^{(l)}}{b} + \cdots + \frac{a_{k_l-1}^{(l)}}{b^{k_l-1}}, \frac{a_1^{(l)}}{b} + \cdots + \frac{a_{k_l-1}^{(l)}}{b^{k_l-1}} + \frac{a_{k_l}^{(l)}}{b^{k_l}}\right).$$

Furthermore, let

$$I'_{k_1,\dots,k_{s-1}} := I_{k_1,\dots,k_{s-1}} \times \left[0, \frac{a_1^{(s)}}{b} + \cdots + \frac{a_{m-k_1-\cdots-k_{s-1}}^{(s)}}{b^{m-k_1-\cdots-k_{s-1}}}\right) \text{ and}$$

$$J'_{k_1,\dots,k_{s-1}} := I_{k_1,\dots,k_{s-1}} \times \left[0, \frac{a_1^{(s)}}{b} + \cdots + \frac{a_{m-k_1-\cdots-k_{s-1}}^{(s)}}{b^{m-k_1-\cdots-k_{s-1}}} + \frac{1}{b^{m-k_1-\cdots-k_{s-1}}}\right)$$

for $k_1 + \cdots + k_{s-1} \leq m - 1$, and

$$I'_{k_1,\dots,k_{s-1}} := I_{k_1,\dots,k_{s-1}} \times [0, 0) = \emptyset \text{ and}$$
$$J'_{k_1,\dots,k_{s-1}} := I_{k_1,\dots,k_{s-1}} \times [0, 1)$$

for $k_1 + \cdots + k_{s-1} = m$.

It is readily checked that the intervals $I'_{k_1,\dots,k_{s-1}}$ are again disjoint unions of elementary intervals of volume $b^{-m}$. Furthermore,

$$I'_{k_1,\dots,k_{s-1}} \cap I'_{j_1,\dots,j_{s-1}} = \emptyset = J'_{k_1,\dots,k_{s-1}} \cap J'_{j_1,\dots,j_{s-1}}$$

for $(k_1, \dots, k_{s-1}) \neq (j_1, \dots, j_{s-1})$. Put

$$B_s := \bigcup_{\substack{k_1,\dots,k_{s-1} \in \mathbb{N} \\ k_1+\cdots+k_{s-1} \leq m}} I'_{k_1,\dots,k_{s-1}},$$

$$B'_s := \bigcup_{\substack{k_1,\dots,k_{s-1} \in \mathbb{N} \\ k_1+\cdots+k_{s-1} \leq m}} J'_{k_1,\dots,k_{s-1}} \cup (B'_{s-1} \setminus B_{s-1}) \times [0, 1),$$

where $B_{s-1}$ and $B'_{s-1}$ are the sets from the $(s-1)$-dimensional case approximating $[0, \alpha^{(1)}) \times \cdots \times [0, \alpha^{(s-1)})$ with

$$\lambda_{s-1}(B'_{s-1} \setminus B_{s-1}) \leq \frac{1}{b^m} \sum_{k=0}^{s-2} \binom{m}{k} (b-1)^k.$$

Then we have $B_s \subseteq B \subseteq B'_s$ and, like in the case $s = 2$, we obtain

$$\begin{aligned}\lambda_s(B'_s \setminus B_s) &= \sum_{\substack{k_1,\dots,k_{s-1}=1 \\ k_1+\cdots+k_{s-1} \leq m}}^{m} \lambda_s(J'_{k_1,\dots,k_{s-1}} \setminus I'_{k_1,\dots,k_{s-1}}) + \lambda_{s-1}(B'_{s-1} \setminus B_{s-1}) \\ &= \sum_{\substack{k_1,\dots,k_{s-1}=1 \\ k_1+\cdots+k_{s-1} \leq m}}^{m} \frac{1}{b^{m-k_1-\cdots-k_{s-1}}} \prod_{l=1}^{s-1} \frac{a_{k_l}^{(l)}}{b^{k_l}} + \frac{1}{b^m} \sum_{k=0}^{s-2} \binom{m}{k} (b-1)^k\end{aligned}$$

$$\leq \frac{(b-1)^{s-1}}{b^m} \sum_{\substack{k_1,\ldots,k_{s-1}=1\\ k_1+\cdots+k_{s-1}\leq m}}^{m} 1 + \frac{1}{b^m}\sum_{k=0}^{s-2}\binom{m}{k}(b-1)^k$$
$$= \frac{1}{b^m}\sum_{k=0}^{s-1}\binom{m}{k}(b-1)^k,$$

where we have used Lemma 5.19.
Thus the proof is finished. □

**Corollary 5.20.** *For the star discrepancy of a $(t, m, s)$-net $\mathcal{P}$ in base $b$ with $m \geq 2s - 2 + t$ we have*

$$D^*_{b^m}(\mathcal{P}) \leq \frac{b^t}{b^m}\left(1 + \frac{(b-1)^{s-1}}{(s-2)!}(m-t)^{s-1}\right).$$

*Proof.* For $m \geq 2s - 2 + t$ and $k \in \{1, \ldots, s-1\}$,

$$\binom{m-t}{k} \leq \binom{m-t}{s-1} = \frac{(m-t)(m-t-1)\cdots(m-t-s+2)}{(s-1)!} \leq \frac{(m-t)^{s-1}}{(s-1)!}.$$

Using this estimate in Theorem 5.18 gives the desired result. □

Since $m = (\log b^m)/\log b$, it follows that a $(0, m, s)$-net in base $b$ has star discrepancy of order of magnitude $O((\log N)^{s-1}/N)$, where $N = b^m$ with an implied constant only depending on $s$ and $b$. Furthermore, recall from Remark 5.9 that for all $s, b \in \mathbb{N}$, $b \geq 2$, there exists a $t_b(s)$ such that for all $m \geq t_b(s)$ a $(t_b(s), m, s)$-net in base $b$ exists and $t_b(s) \asymp_b s$. Let $t_b(s) \leq cs$ for $c = c(b) > 0$ and let $(\mathcal{P}_m)_{m \geq t_b(s)}$ be a sequence of $(t_b(s), m, s)$-nets in base $b$. Using Stirling's formula we find that $n! \geq \sqrt{2\pi n}(n/\mathrm{e})^n$ for $n \in \mathbb{N}$, where $\mathrm{e} = \exp(1)$. Then we have

$$\limsup_{m\to\infty} \frac{b^m D^*_{b^m}(\mathcal{P}_m)}{(\log b^m)^{s-1}} \leq \frac{b^{cs}(b-1)^{s-1}}{(s-2)!(\log b)^{s-1}} \leq \frac{b^{2c}(b-1)}{\sqrt{2\pi(s-2)}(\log b)^{s-1}}\left(\frac{\mathrm{e}b^c(b-1)}{s-2}\right)^{s-2}.$$

This expression tends to zero at a superexponential rate as $s \to \infty$.

Now we turn to discrepancy estimates for $(t, s)$-sequences in base $b$. To this end, assume that $\Delta_b(t, m, s)$ is a number such that

$$b^m D^*_{b^m}(\mathcal{P}) \leq \Delta_b(t, m, s)$$

holds for the star discrepancy of any $(t, m, s)$-net $\mathcal{P}$ in base $b$. The star discrepancy $D_N^*$ of the infinite $(t, s)$-sequence $\mathcal{S} = (\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$ in base $b$ is by definition the star discrepancy of the point set $\mathcal{P}_N = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}\}$ consisting of the first $N$ elements of $\mathcal{S}$. Let $N$ have $b$-adic representation $N = a_r b^r + \cdots + a_1 b + a_0$. Then we partition the point set $\mathcal{P}_N$ into the disjoint point sets

$$\mathcal{P}_{m,a} = \{\boldsymbol{x}_{a_r b^r + \cdots + a_{m+1} b^{m+1} + a b^m + k} \, : \, k = 0, \dots, b^m - 1\},$$

for $m = 0, \dots, r$ and $a = 1, \dots, a_m$. Hence,

$$\mathcal{P}_N = \bigcup_{m=0}^{r} \bigcup_{a=0}^{a_m - 1} \mathcal{P}_{m,a}.$$

Every $\mathcal{P}_{m,a}$ is a $(t, m, s)$-net in base $b$ and $b^m$ times its star discrepancy is at most $\Delta_b(t, m, s)$. Using the triangle inequality for the star discrepancy (see Exercise 2.14), we obtain the following result.

**Lemma 5.21.** *Let $\mathcal{S}$ be a $(t, s)$-sequence in base $b$. Let $N \in \mathbb{N}$ with $b$-adic expansion $N = a_r b^r + \cdots + a_1 b + a_0$. Then*

$$N D_N^*(\mathcal{S}) \le \sum_{m=0}^{r} a_m \Delta_b(t, m, s).$$

Theorem 5.18 shows that we can choose $\Delta_b(t, m, s) = b^t \sum_{k=0}^{s-1} \binom{m-t}{k} (b-1)^k$. Using this estimate in conjunction with Lemma 5.21 we obtain the following estimate.

**Theorem 5.22.** *For the star discrepancy of a $(t, s)$-sequence $\mathcal{S}$ in base $b$ we have*

$$D_N^*(\mathcal{S}) \le \frac{b^t (b-1)}{N} \sum_{m=0}^{r} \sum_{k=0}^{s-1} \binom{m-t}{k} (b-1)^k,$$

*where $r = \lfloor (\log N)/(\log b) \rfloor$.*

For $r \to \infty$ we have

$$\sum_{\substack{m=0 \\ m \ge 2s-2+t}}^{r} \sum_{k=0}^{s-1} \binom{m-t}{k} (b-1)^k \le \sum_{\substack{m=0 \\ m \ge 2s-2+t}}^{r} \left( 1 + \frac{m^{s-1}}{(s-2)!} (b-1)^{s-1} \right)$$

$$\leq r + 1\frac{(b-1)^{s-1}}{(s-2)!}\sum_{m=0}^{r} m^{s-1}$$

$$= \frac{(b-1)^{s-1}(s-1)}{s!} r^s + O(r^{s-1})$$

and

$$\sum_{\substack{m=0 \\ m<2s-2+t}}^{r} \sum_{k=0}^{s-1} \binom{m-t}{k} (b-1)^k = O(1).$$

Hence we conclude from Theorem 5.22 that

$$D_N^*(\mathcal{S}) \leq \frac{b^t(b-1)^s(s-1)}{s!(\log b)^s}\frac{(\log N)^s}{N} + O\left(\frac{(\log N)^{s-1}}{N}\right).$$

To compare this estimate with the Halton sequences we study the quantity $d^*$ which was introduced in (2.17). The above bound on the star discrepancy implies that

$$d^*(\mathcal{S}) = \limsup_{N\to\infty} \frac{ND_N^*(\mathcal{S})}{(\log N)^s} \leq b^t\frac{(b-1)^s(s-1)}{s!(\log b)^s} \leq b^t\frac{(b-1)^s(s-1)\mathrm{e}^s}{\sqrt{2\pi s}\, s^s(\log b)^s}.$$

It is known that for any $s, b \in \mathbb{N}$, $b \geq 2$, one can construct a $(t_b(s), s)$-sequence $\mathcal{S}_s^*$ in base $b$ with $t_b(s) \asymp_b s$. For such sequences we obtain

$$\limsup_{s\to\infty} \frac{\log d^*(\mathcal{S}_s^*)}{s\log s} \leq -1,$$

and hence $d^*(\mathcal{S}_s^*)$ tends to zero at a superexponential rate as $s \to \infty$. This result should be compared with the analogous result for Halton sequences presented in Sect. 2.4, Eq. (2.18).

## 5.3 Digital Nets and Sequences

In the last section we have shown that $(t, m, s)$-nets and $(t, s)$-sequences in base $b$ have the desired property of having a low star discrepancy (at least in an asymptotic sense). Now the question arises how to construct concrete examples of $(t, m, s)$-nets and $(t, s)$-sequences in base $b$ with good equidistribution properties. The concept of *digital* nets and sequences is a general framework for the construction of point sets and sequences, and it is the basis for virtually all concrete constructions of $(t, m, s)$-nets and $(t, s)$-sequences.

Although one can introduce digital nets and sequences in arbitrary integer bases $b \geq 2$, we restrict our treatment to prime power bases $b$ only. The reason for this restriction is that exactly for such $b$ there exists a finite field $\mathbb{F}_b$ of order $b$ and this will be helpful in the construction. We denote the elements of $\mathbb{F}_b$ by $\overline{0}, \overline{1}, \ldots, \overline{b-1}$, where $\overline{0}$ denotes the neutral element of addition in $\mathbb{F}_b$. If $b$ is a prime number, then we identify $\mathbb{F}_b$ with $\mathbb{Z}_b$, the set of residue classes modulo $b$ equipped with arithmetic operations modulo $b$, which in turn we identify with the elements of $\{0, \ldots, b-1\}$.

▶ **Remark 5.23 (Some background on finite fields)** Recall from Algebra that a finite field of order $b$ exists if and only if $b$ is a prime power. Furthermore, a finite field of prime power order $b$ is uniquely determined up to isomorphisms. The construction of a finite field uses irreducible polynomials and is based on the following result: If $K$ is a field and if $f \in K[x]$ is irreducible (i.e., it cannot be factored into the product of two or more non-trivial polynomials over $K$), then the residue class ring $K[x]/(f)$ is also a field.

Hence, to construct a finite field with $b = q^m$ elements, where $q$ is a prime number and $m \in \mathbb{N}$, one can start with the finite field $\mathbb{F}_q$ (which we identify with the set of residue classes modulo $q$ equipped with arithmetic operations modulo $q$) and an irreducible polynomial $f \in \mathbb{F}_q[x]$ of degree $m$ (whose existence is also guaranteed by a result from algebra). Then $\mathbb{F}_q[x]/(f)$ is a field that contains exactly $b = q^m$ elements which is usually denoted by $\mathbb{F}_b$.

For example, to construct $\mathbb{F}_4$, the finite field with four elements, we start with $\mathbb{F}_2 = \{0, 1\}$ and choose $f(x) = x^2 + x + 1 \in \mathbb{F}_2[x]$, which is irreducible over $\mathbb{F}_2$. Then $\mathbb{F}_2[x]/(f)$ has the four elements $0, 1, [x], [x+1]$ and we obtain the following operation tables for addition and multiplication, where we write $A = [x]$ and $B = [x+1]$ and where we use arithmetic operations modulo $(f)$, i.e., $x^2 = x + 1$:

| $+$ | $0$ | $1$ | $A$ | $B$ |
|---|---|---|---|---|
| $0$ | $0$ | $1$ | $A$ | $B$ |
| $1$ | $1$ | $0$ | $B$ | $A$ |
| $A$ | $A$ | $B$ | $0$ | $1$ |
| $B$ | $B$ | $A$ | $1$ | $0$ |

and

| $\cdot$ | $0$ | $1$ | $A$ | $B$ |
|---|---|---|---|---|
| $0$ | $0$ | $0$ | $0$ | $0$ |
| $1$ | $0$ | $1$ | $A$ | $B$ |
| $A$ | $0$ | $A$ | $B$ | $1$ |
| $B$ | $0$ | $B$ | $1$ | $A$ |

## Digital $(t, m, s)$-Nets Over $\mathbb{F}_b$

To construct a digital $(t, m, s)$-net in a prime power base $b$, we use the finite field $\mathbb{F}_b$ with $b$ elements and a bijection $\varphi : \{0, \ldots, b-1\} \to \mathbb{F}_b$ which maps the set of $b$ adic digits onto $\mathbb{F}_b$. We assume in the following that the elements of $\mathbb{F}_b$ are ordered in such a way that $\varphi(a) = \overline{a}$ for $a \in \{0, \ldots, b-1\}$. If $b$ is a prime number we identify $\mathbb{F}_b$ with $\{0, \ldots, b-1\}$ and omit the bijection $\varphi$ and the bar in this case.

Having set the stage, we now explain the digital method. Let $s, m \in \mathbb{N}$ and let $b$ be a prime power. We aim at constructing a $b^m$-element point set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{b^m-1}\}$ in $[0, 1)^s$. To generate such a point set we first choose $m \times m$ matrices $C_1, \ldots, C_s$ (one for each dimension) over $\mathbb{F}_b$, that is, with entries from $\mathbb{F}_b$.

**Digital method.** To generate the point $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,s})$ for $n \in \{0, \ldots, b^m - 1\}$ we write $n$ in its base $b$ expansion

$$n = n_0 + n_1 b + \cdots + n_{m-1} b^{m-1}$$

with digits $n_j \in \{0, \ldots, b-1\}$. Then take the $m$-dimensional column vector

$$\vec{n} = \begin{pmatrix} \varphi(n_0) \\ \varphi(n_1) \\ \vdots \\ \varphi(n_{m-1}) \end{pmatrix}$$

with entries from $\mathbb{F}_b$ and multiply it by the matrix $C_j$, $j = 1, 2, \ldots, s$:

$$C_j \vec{n} =: \begin{pmatrix} \overline{y}_{n,j,1} \\ \overline{y}_{n,j,2} \\ \vdots \\ \overline{y}_{n,j,m} \end{pmatrix}.$$

Transforming the entries of this vector with elements in $\mathbb{F}_b$ back into the set of base $b$ digits $\{0, 1, \ldots, b-1\}$ by applying the inverse map $\varphi^{-1}$ we obtain $y_{n,j,l} = \varphi^{-1}(\overline{y}_{n,j,l})$ for $l = 1, 2, \ldots, m$, which are taken as the base $b$ digits of the $j$th component of the point $\boldsymbol{x}_n$, i.e.,

$$x_{n,j} := \frac{y_{n,j,1}}{b} + \frac{y_{n,j,2}}{b^2} + \cdots + \frac{y_{n,j,m}}{b^m}.$$

Finally, we put

$$\boldsymbol{x}_n := (x_{n,1}, \ldots, x_{n,s}).$$

Every point set constructed with the digital method is a $(t, m, s)$-net in base $b$ for some quality parameter $t \in \{0, \ldots, m\}$ since it is at least an $(m, m, s)$-net in base $b$. This leads to the following definition:

**Definition 5.24 (Digital net)**

If the point set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{b^m-1}\}$ constructed by the digital method is for some $t \in \{0, \ldots, m\}$ a $(t, m, s)$-net in base $b$, then it is called a *digital* $(t, m, s)$*-net over the field* $\mathbb{F}_b$ or, for short, a *digital net* with generating matrices $C_1, \ldots, C_s$. It is called a *strict* digital $(t, m, s)$-net over $\mathbb{F}_b$ if $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{b^m-1}\}$ is a strict $(t, m, s)$-net in base $b$.

In most cases the finite field $\mathbb{Z}_b$ with $b$ prime is chosen for practical applications, and indeed $\mathbb{Z}_2$ is the most frequent choice.

## Example 5.25

We want to construct a digital $(t, 4, 2)$-net over $\mathbb{Z}_2$. First we have to choose two $4 \times 4$ matrices over $\mathbb{Z}_2$, for instance

$$C_1 = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} 0&0&0&1 \\ 0&0&1&0 \\ 0&1&0&0 \\ 1&0&0&0 \end{pmatrix}. \tag{5.4}$$

As an example we show how to construct the point $\boldsymbol{x}_{10}$. We have $n = 10 = 0+1\cdot 2+0\cdot 2^2+1\cdot 2^3$ and therefore

$$C_1\vec{n} = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix},$$

$$C_2\vec{n} = \begin{pmatrix} 0&0&0&1 \\ 0&0&1&0 \\ 0&1&0&0 \\ 1&0&0&0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Hence, $x_{10,1} = \frac{1}{4} + \frac{1}{16} = \frac{5}{16}$ and $x_{10,2} = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$, and thus $\boldsymbol{x}_{10} = \left(\frac{5}{16}, \frac{5}{8}\right)$.

We will show now how the (strict) quality parameter $t$ is connected with the choice of the generating matrices $C_1, \ldots, C_s$. This can be answered with the help of the following quantity $\rho$ which, in some sense, "measures" the "linear independence of the $s$ matrices $C_1, \ldots, C_s$".

### Definition 5.26

Let $b$ be a prime power and let $C_1, \ldots, C_s$ be $m \times m$ matrices over $\mathbb{F}_b$. Let $\rho = \rho(C_1, \ldots, C_s)$ be the largest integer such that for any choice of $d_1, \ldots, d_s \in \mathbb{N}_0$ with

$$d_1 + \cdots + d_s = \rho$$

the following holds:
The first $d_1$ row vectors of $C_1$ together with
the first $d_2$ row vectors of $C_2$ together with
$\vdots$
the first $d_s$ row vectors of $C_s$,
(these are together $\rho$ vectors in $\mathbb{F}_b^m$) are linearly independent over the finite field $\mathbb{F}_b$. We call $\rho$ the *linear independence parameter* of the matrices $C_1, \ldots, C_s$.

## Example 5.27

Consider $C_1$ and $C_2$ over $\mathbb{Z}_2$ from Example 5.25. Clearly, for these $4 \times 4$ matrices $\rho(C_1, C_2)$ is at most 4. It is easy to see that in the present case $\rho(C_1, C_2)$ is indeed 4, since for any choice of $d_1, d_2 \in \mathbb{N}_0$ with $d_1 + d_2 = 4$, the first $d_1$ rows of $C_1$ together with the first $d_2$ rows of $C_2$ provides the four canonical row-vectors $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$, which are linearly independent over $\mathbb{Z}_2$. Hence, $\rho(C_1, C_2) = 4$.

Now we can determine the strict quality parameter $t$ of a digital net generated by matrices $C_1, \dots, C_s$ over $\mathbb{Z}_b$.

**Theorem 5.28 (Niederreiter).** *Let $s, m \in \mathbb{N}$ and let $b$ be a prime power. The point set constructed by the digital method with the $m \times m$ matrices $C_1, \dots, C_s$ over a finite field $\mathbb{F}_b$ is a strict $(m - \rho, m, s)$-net in base $b$, where $\rho = \rho(C_1, \dots, C_s)$ is the linear independence parameter introduced in Definition 5.26.*

*Proof.* First we have to show that every $s$-dimensional elementary $b$-adic interval of volume $b^{-\rho}$ contains exactly $b^{m-\rho}$ of the generated points. Let

$$J = \prod_{j=1}^{s} \left[ \frac{a_j}{b^{d_j}}, \frac{a_j + 1}{b^{d_j}} \right)$$

with $d_1, \dots, d_s \in \mathbb{N}_0$ such that $d_1 + \dots + d_s = \rho$, and $a_j \in \{0, \dots, b^{d_j} - 1\}$ for $j = 1, 2, \dots, s$. We have to count the number of indices $n \in \{0, 1, \dots, b^m - 1\}$ for which $\boldsymbol{x}_n$ belongs to $J$. We have $\boldsymbol{x}_n \in J$ if and only if

$$x_{n,j} \in \left[ \frac{a_j}{b^{d_j}}, \frac{a_j + 1}{b^{d_j}} \right) \qquad (5.5)$$

for all $j = 1, 2, \dots, s$. Let $a_j = e_{d_j}^{(j)} + \dots + e_1^{(j)} b^{d_j - 1}$ and hence

$$\frac{a_j}{b^{d_j}} = \frac{e_1^{(j)}}{b} + \dots + \frac{e_{d_j}^{(j)}}{b^{d_j}}.$$

Then (5.5) is equivalent to

$$\frac{e_1^{(j)}}{b} + \dots + \frac{e_{d_j}^{(j)}}{b^{d_j}} \le x_{n,j} < \frac{e_1^{(j)}}{b} + \dots + \frac{e_{d_j}^{(j)}}{b^{d_j}} + \frac{1}{b^{d_j}},$$

which, in turn, is equivalent to

$$x_{n,j} = \frac{e_1^{(j)}}{b} + \dots + \frac{e_{d_j}^{(j)}}{b^{d_j}} + \cdots.$$

Hence the condition (5.5) uniquely determines the first $d_j$ base $b$ digits of $x_{n,j}$.

According to the digital construction scheme, the $i$th base $b$ digit of $x_{n,j}$ is given by $\varphi^{-1}$ applied to the product $\vec{c}_i^{\,(j)}\vec{n}$ of the $i$th row vector $\vec{c}_i^{\,(j)}$ of $C_j$ with the column vector $\vec{n}$, where $\varphi$ is the bijection used in the construction. Hence $\boldsymbol{x}_n \in J$ if and only if the following system of equations over $\mathbb{F}_b$ is satisfied:

$$\begin{aligned}
\vec{c}_1^{\,(1)}\vec{n} &= \varphi(e_1^{(1)}),\\
&\ \vdots\\
\vec{c}_{d_1}^{\,(1)}\vec{n} &= \varphi(e_{d_1}^{(1)}),\\
\vec{c}_1^{\,(2)}\vec{n} &= \varphi(e_1^{(2)}),\\
&\ \vdots\\
\vec{c}_{d_2}^{\,(2)}\vec{n} &= \varphi(e_{d_2}^{(2)}),\\
&\ \vdots\\
\vec{c}_1^{\,(s)}\vec{n} &= \varphi(e_1^{(s)}),\\
&\ \vdots\\
\vec{c}_{d_s}^{\,(s)}\vec{n} &= \varphi(e_{d_s}^{(s)}).
\end{aligned} \tag{5.6}$$

This system consists of exactly $d_1 + d_2 + \cdots + d_s = \rho$ equations. Let $A$ be the $\rho \times m$ matrix over $\mathbb{F}_b$ with row vectors

$$\vec{c}_1^{\,(1)}, \ldots, \vec{c}_{d_1}^{\,(1)}, \vec{c}_1^{\,(2)}, \ldots, \vec{c}_{d_2}^{\,(2)}, \ldots, \vec{c}_1^{\,(s)}, \ldots, \vec{c}_{d_s}^{\,(s)}, \tag{5.7}$$

and let

$$\vec{f}^{\top} = (\varphi(e_1^{(1)}), \ldots, \varphi(e_{d_1}^{(1)}), \varphi(e_1^{(2)}), \ldots, \varphi(e_{d_2}^{(2)}), \ldots, \varphi(e_1^{(s)}), \ldots, \varphi(e_{d_s}^{(s)})) \in \mathbb{F}_b^{\rho}.$$

Then the linear system (5.6) can be rewritten into the system

$$A\vec{n} = \vec{f}. \tag{5.8}$$

Since the system of row vectors (5.7) by the definition of $\rho$ is linearly independent over $\mathbb{F}_b$, the $\rho \times m$ matrix $A$ has rank $\rho$. This implies that (5.8) is solvable and that the nullspace of $A$ has dimension $m-\rho$. Hence (5.8) has exactly $b^{m-\rho}$ solutions and thus $J$ contains $b^{m-\rho}$ elements from the digital net. This means that the matrices $C_1, \ldots, C_s$ generate a digital $(m-\rho, m, s)$-net over $\mathbb{F}_b$.

It remains to show the strictness of the quality parameter. If $\rho = m$, then there is nothing to prove. If $\rho \leq m-1$, then there are $d_1, \ldots, d_s \in \mathbb{N}_0$ with $d_1 + \cdots + d_s = \rho + 1$, and such that the vectors

$$\vec{c}_1^{\,(1)}, \ldots, \vec{c}_{d_1}^{\,(1)}, \ldots, \vec{c}_1^{\,(s)}, \ldots, \vec{c}_{d_s}^{\,(s)}$$

are linearly dependent over $\mathbb{F}_b$. But then the system (5.6) with $\varphi(e_i^{(j)}) = 0$ for $i = 1, \ldots, d_j$ and $j = 1, \ldots, s$, also has $b^{m-\rho}$ solutions $\vec{n}$ (although it consists of $\rho + 1$ equations in $m$ variables). This means that the corresponding elementary interval

$$\tilde{J} := \prod_{j=1}^{s} \left[ \frac{a_j}{b^{d_j}}, \frac{a_j + 1}{b^{d_j}} \right)$$

with $a_j = e_{d_j}^{(j)} + \cdots + e_1^{(j)} b^{d_j - 1}$ for $j = 1, 2, \ldots, s$ and of volume $b^{-(\rho+1)}$ contains $b^{m-\rho}$ points of the digital net. Consequently,

$$A(\tilde{J}, b^m) - b^m \lambda_s(\tilde{J}) = b^{m-\rho} - b^{m-\rho-1} = (b-1)b^{m-\rho-1} \neq 0,$$

and so the matrices $C_1, \ldots, C_s$ generate a strict digital $(m-\rho, m, s)$-net over $\mathbb{F}_b$. □

## Digital $(t, s)$-Sequences Over $\mathbb{F}_b$

To construct a $(t, s)$-sequence in a prime power base $b$ by the digital method, we again use a finite field $\mathbb{F}_b$ and a bijection $\varphi : \{0, \ldots, b-1\} \to \mathbb{F}_b$ for which we demand that $\varphi(0) = \overline{0}$.

First we have to choose $\mathbb{N} \times \mathbb{N}$ matrices $C_1, \ldots, C_s$ (one for each component) over $\mathbb{F}_b$. That is, matrices of the form

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{13} & \ldots \\ c_{21} & c_{22} & c_{23} & \ldots \\ c_{31} & c_{32} & c_{33} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \in \mathbb{F}_b^{\mathbb{N} \times \mathbb{N}}.$$

**Digital method.** To generate one of the points $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,s})$ for $n \in \mathbb{N}_0$ of the $(t, s)$-sequence, we again write $n$ in its base $b$ expansion $n = \sum_{i=0}^{\infty} a_i b^i$ with $a_i \in \{0, \ldots, b-1\}$ and $a_i = 0$ for all $i$ large enough. Then take the column vector

$$\vec{n} = \begin{pmatrix} \varphi(a_0) \\ \varphi(a_1) \\ \varphi(a_2) \\ \vdots \end{pmatrix}$$

with entries in $\mathbb{F}_b$ and multiply it by $C_j$, $j = 1, 2, \ldots, s$:

$$C_j \vec{n} = \begin{pmatrix} \overline{y}_{n,j,1} \\ \overline{y}_{n,j,2} \\ \vdots \end{pmatrix}.$$

Note that for the multiplication only finitely many of the entries of $\vec{n}$ are different from zero, as we assumed that $\varphi(0) = \overline{0}$. Transforming the entries of $C_j \vec{n}$ back into the set of base $b$ digits $\{0, 1, \ldots, b-1\}$ by applying the inverse map $\varphi^{-1}$ we obtain

$$x_{n,j} = \frac{\varphi^{-1}(\overline{y}_{n,j,1})}{b} + \frac{\varphi^{-1}(\overline{y}_{n,j,2})}{b^2} + \cdots .$$

Finally, we put

$$\boldsymbol{x}_n := (x_{n,1}, \ldots, x_{n,s}).$$

**Definition 5.29**

We call the sequence $(\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$ constructed in this way a *digital sequence over* $\mathbb{F}_b$ *with generating matrices* $C_1, \ldots, C_s$, or short, a *digital sequence*.

► **Remark 5.30** Depending on the matrices $C_1, \ldots, C_s$, it may happen that the vector $C_j \vec{n} =: \vec{y}_{n,j}$ contains infinitely many entries different from zero. For practical purposes this requires an adaptation of the point generation. Usually the vector $\vec{y}_{n,j}$ is truncated at a suitable place.

Furthermore, another theoretical problem may arise. Namely, it should be avoided that the vector $\vec{y}_{n,j}$ contains only finitely many elements different from $\varphi(b-1)$. Because of the nonuniqueness of representation of the $b$-adic real numbers represented by such "digit vectors", the net structure of the sequence under consideration would be destroyed. This is the reason for the following additional "finiteness" condition on the matrices $C_1, \ldots, C_s$. Let

$$C_j = (c_{i,r}^{(j)})_{i,r \in \mathbb{N}} \in \mathbb{F}_b^{\mathbb{N} \times \mathbb{N}}$$

for $j = 1, \ldots, s$.

**F:** We demand that for all $j$ and $r$ we have $c_{i,r}^{(j)} = 0$ for all $i$ large enough.

To eliminate the problems with ambiguous digit expansions without insisting on the finiteness condition **F** one can use a slightly modified definition of digital sequences which is based on a certain truncation operator. Such sequences are then occasionally called *digital sequences in the broad sense*.

We will show in the following that under certain conditions on the generating matrices a digital sequence is a $(t, s)$-sequence in base $b$ for some $t \in \mathbb{N}_0$ which we then call a *digital* $(t, s)$*-sequence over* $\mathbb{F}_b$. Before we do so we again have to introduce the quantities $\rho_m$, which in some sense “measure” the “linear independence” of the $s$ infinite matrices $C_1, \ldots, C_s$.

**Definition 5.31**

Let $C_1, \ldots, C_s$ be $\mathbb{N} \times \mathbb{N}$ matrices over the finite field $\mathbb{F}_b$. For any $j = 1, 2, \ldots, s$ and $m \in \mathbb{N}$ we denote by $C_j^{(m)}$ the left upper $m \times m$ sub-matrix of $C_j$. Then

$$\rho_m = \rho_m(C_1, \ldots, C_s) := \rho(C_1^{(m)}, \ldots, C_s^{(m)}),$$

where $\rho$ is the linear independence parameter defined for $s$-tuples of $m \times m$ matrices over $\mathbb{F}_b$ in Definition 5.26.

Now we can determine the quality parameter $t$ of a digital sequence.

**Theorem 5.32 (Niederreiter).** *Let $C_1, \ldots, C_s$ be $\mathbb{N} \times \mathbb{N}$ matrices over $\mathbb{F}_b$. If there exists a $t \in \mathbb{N}_0$ such that for each $m > t$ we have $\rho_m(C_1, \ldots, C_s) \geq m - t$, then the digital sequence generated by $C_1, \ldots, C_s$ is a digital $(t, s)$-sequence over $\mathbb{F}_b$.*

*Proof.* Let $(\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$ denote the digital sequence generated by $C_1, \ldots, C_s$. By the definition of a $(t, s)$-sequence we have to show that for any $m \in \mathbb{N}$ and any $k \in \mathbb{N}_0$ the point set

$$\mathcal{P}_{k,m} = \{\boldsymbol{x}_{kb^m}, \ldots, \boldsymbol{x}_{kb^m + b^m - 1}\}$$

is a $(t, m, s)$-net in base $b$. Indeed, for given $k$ and $m$, and any $l \in \{0, \ldots, b^m - 1\}$ let $k = \kappa_{r+1} b^r + \cdots + \kappa_1$ and $l = \lambda_{m-1} b^{m-1} + \cdots + \lambda_0$ be the base $b$ representations of $k$ and $l$, respectively. For $n = kb^m + l$ we have

$$\vec{n} = (\overline{\lambda}_0, \ldots, \overline{\lambda}_{m-1}, \overline{\kappa}_1, \ldots, \overline{\kappa}_{r+1}, \ldots)^\top$$

and with the following representation of the matrices $C_j$,

$$C_j = \left(\begin{array}{c|c} C_j^{(m)} & F_j^{(m)} \\ \hline \multicolumn{2}{c}{H_j^{(m)}} \end{array}\right) \in \mathbb{F}_b^{\mathbb{N} \times \mathbb{N}},$$

where $F_j^{(m)} \in \mathbb{F}_b^{m\times\mathbb{N}}$ and $H_j^{(m)} \in \mathbb{F}_b^{\mathbb{N}\times\mathbb{N}}$, we have

$$C_j\vec{n} = \begin{pmatrix} C_j^{(m)}\vec{l} \\ \overline{0} \\ \overline{0} \\ \vdots \end{pmatrix} + \begin{pmatrix} F_j^{(m)}\vec{k} \\ \overline{0} \\ \overline{0} \\ \vdots \end{pmatrix} + \begin{pmatrix} \overline{0} \\ \vdots \\ \overline{0} \\ H_j^{(m)}\vec{n} \end{pmatrix},$$

where $\vec{l} = (\overline{\lambda}_0, \dots, \overline{\lambda}_{m-1})^\top$ and $\vec{k} = (\overline{\kappa}_1, \dots, \overline{\kappa}_{r+1}, \overline{0}, \overline{0}, \dots)^\top$.

In order to show that $\mathcal{P}_{k,m}$ is a $(t, m, s)$-net in base $b$ we need to show that every elementary interval of the form

$$J = \prod_{j=1}^{s} \left[\frac{a_j}{b^{d_j}}, \frac{a_j + 1}{b^{d_j}}\right)$$

with $d_1, \dots, d_s \in \mathbb{N}_0$ such that $d_1 + \dots + d_s = m - t$, and $a_j \in \{0, \dots, b^{d_j} - 1\}$ for $j = 1, 2, \dots, s$ contains exactly $b^t$ elements from $\mathcal{P}_{k,m}$. So we count the number of indices $n \in \{kb^m, kb^m + 1, \dots, kb^m + b^m - 1\}$ for which $\boldsymbol{x}_n \in J$, that is, the number of $l \in \{0, 1, \dots, b^m - 1\}$ for which $\boldsymbol{x}_{kb^m+l} \in J$. Let

$$\frac{a_j}{b^{d_j}} = \frac{e_1^{(j)}}{b} + \dots + \frac{e_{d_j}^{(j)}}{b^{d_j}}.$$

Then we have, as in the proof of Theorem 5.28, that $\boldsymbol{x}_n \in J$ if and only if

$$x_{n,j} = \frac{e_1^{(j)}}{b} + \dots + \frac{e_{d_j}^{(j)}}{b^{d_j}} + \cdots.$$

With the above considerations, this is equivalent to

$$\begin{aligned} \vec{c}_1^{\,(j)}\vec{l} &= \varphi(e_1^{(j)}) - \vec{f}_1^{\,(j)}\vec{k} \\ &\vdots \\ \vec{c}_{d_j}^{\,(j)}\vec{l} &= \varphi(e_{d_j}^{(j)}) - \vec{f}_{d_j}^{\,(j)}\vec{k} \end{aligned}$$

where $\vec{c}_i^{\,(j)}$ denotes the $i$th row vector of the matrix $C_j^{(m)}$ and $\vec{f}_i^{\,(j)}$ the $i$th row vector of the matrix $F_j^{(m)}$. For $j = 1, 2, \dots, s$ this gives a linear system of $d_1 + \dots + d_s = m - t$ linear equations in the $m$ variables $\overline{\lambda}_0, \dots, \overline{\lambda}_{m-1} \in \mathbb{F}_b$.

Since the row vectors $\vec{c}_i^{\,(j)}$ for $i = 1, \dots, d_j$ and $j = 1, 2, \dots, s$ are, by the definition of $\rho_m$, linearly independent over $\mathbb{F}_b$, the linear system has exactly $b^{m-(m-t)} = b^t$ solution and hence $J$ contains $b^t$ elements from $\mathcal{P}_{k,m}$. □

### Example 5.33

The van der Corput sequence can be generated by the digital construction scheme using the matrix

$$C_1 = \begin{pmatrix} \overline{1} & \overline{0} & \overline{0} & \overline{0} & \dots \\ \overline{0} & \overline{1} & \overline{0} & \overline{0} & \dots \\ \overline{0} & \overline{0} & \overline{1} & \overline{0} & \dots \\ \overline{0} & \overline{0} & \overline{0} & \overline{1} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

over $\mathbb{F}_b$. Since $\rho_m(C_1) = m$ for every $m \in \mathbb{N}$, it follows from Theorem 5.32 that the van der Corput sequence in prime power base $b$ is a digital $(0, 1)$-sequence over $\mathbb{F}_b$.

Interestingly, the construction of a second generating matrix to obtain a, say digital $(0, 2)$-sequence over $\mathbb{F}_b$, is already nontrivial. This question is addressed in the following section.

## 5.4 Special Constructions of Digital Nets and Sequences

Although I.M. Sobol' was the first to construct digital $(t, s)$-sequences in base 2 in 1967 and H. Faure introduced constructions of digital $(0, s)$-sequences in prime base $b$ with $b \geq s$ in 1982, it now seems most convenient to introduce digital $(t, s)$-sequences using H. Niederreiter's unifying approach based on polynomial arithmetic over finite fields. By considering generalised Niederreiter sequences, the constructions by Sobol' and Faure appear as special cases.

For a prime power $b$, let $\mathbb{F}_b((x^{-1}))$ be the field of *formal Laurent series* over $\mathbb{F}_b$, that is, series of the form

$$L(x) = \sum_{l=w}^{\infty} t_l x^{-l},$$

where $w$ is an arbitrary integer and all $t_l \in \mathbb{F}_b$. Note that $\mathbb{F}_b((x^{-1}))$ contains the field of rational functions over $\mathbb{F}_b$ as a subfield. Furthermore, let $\mathbb{F}_b[x]$ be the set of all polynomials over $\mathbb{F}_b$. The *discrete exponential valuation* $\nu$ on $\mathbb{F}_b((x^{-1}))$ is defined by $\nu(L) = -w$ if $L \neq 0$, where $w$ is the least index with $t_w \neq 0$. For $L = 0$ we set $\nu(0) = -\infty$. Observe that $\nu(p/q) = \deg(p) - \deg(q)$ for all nonzero polynomials $p, q \in \mathbb{F}_b[x]$ (cf. Exercise 5.15).

### Classical Niederreiter Sequence

Let $s \in \mathbb{N}$, $b$ a prime power and let $p_1, \dots, p_s \in \mathbb{F}_b[x]$ be distinct monic irreducible polynomials over $\mathbb{F}_b$. Recall that a polynomial $p \in \mathbb{F}_b[x]$ is called irreducible if it cannot be factored into the product of two or more non-trivial polynomials and it is called monic if its leading coefficient is equal to the identity $\overline{1}$ of $\mathbb{F}_b$. Let $e_j = \deg(p_j)$ for $j = 1, 2, \ldots, s$. For $j \in \{1, \ldots, s\}$, $i \in \mathbb{N}$ and $k \in \{0, \ldots, e_j - 1\}$, consider the expansions

$$\frac{x^k}{p_j(x)^i} = \sum_{r=0}^{\infty} a^{(j)}(i, k, r) x^{-r} \tag{5.9}$$

over the field $\mathbb{F}_b((x^{-1}))$. Then we define the matrix $C_j = (c_{i,r}^{(j)})_{i,r \in \mathbb{N}}$ by

$$c_{i,r}^{(j)} = a^{(j)}(Q + 1, k, r) \in \mathbb{F}_b \quad \text{for } j \in \{1, \ldots, s\},\ i, r \in \mathbb{N}, \tag{5.10}$$

where the parameters $Q$ and $k$ are uniquely determined by $i$ and $j$ via the Euclidean division $i - 1 = Qe_j + k$ with $Q = Q(j, i) \in \mathbb{N}_0$ and remainder $k = k(j, i) \in \{0, \ldots, e_j - 1\}$. Hence $Q(j, i) = \lfloor (i - 1)/e_j \rfloor$.

Since the exponential evaluation $\nu$ applied to the left-hand side of (5.9) gives $\nu(x^k/p_j(x)^i) = k - ie_j$ and this tends to $-\infty$ for growing $i$, it follows that $c_{i,r}^{(j)} = 0$ for $i$ large enough. Thus condition **F** is satisfied.

**Definition 5.34**

A digital sequence over $\mathbb{F}_b$ generated by the $\mathbb{N} \times \mathbb{N}$ matrices $C_j = (c_{i,r}^{(j)})_{i,r \in \mathbb{N}}$ for $j = 1, 2, \ldots, s$, where the $c_{i,r}^{(j)}$ are given by (5.10), is called a *Niederreiter sequence*.

**Theorem 5.35 (Niederreiter).** *The Niederreiter sequence with generating matrices defined as above, is a digital $(t, s)$-sequence over $\mathbb{F}_b$ with*

$$t = \sum_{j=1}^{s} (e_j - 1).$$

*Proof.* According to Theorem 5.32, we need to show that for all integers $m > \sum_{j=1}^{s} (e_j - 1)$ and all $d_1, \ldots, d_s \in \mathbb{N}_0$ with $1 \leq \sum_{j=1}^{s} d_j \leq m - \sum_{j=1}^{s} (e_j - 1)$, the vectors

$$\vec{c}_i^{\,(j)} = (c_{i,1}^{(j)}, \ldots, c_{i,m}^{(j)}) \in \mathbb{F}_b^m, \quad i = 1, \ldots, d_j, \quad j = 1, 2, \ldots, s, \tag{5.11}$$

are linearly independent over $\mathbb{F}_b$. Suppose that

$$\sum_{j=1}^{s} \sum_{i=1}^{d_j} f_i^{(j)} \vec{c}_i^{\,(j)} = \vec{0} \in \mathbb{F}_b^m$$

for some $f_i^{(j)} \in \mathbb{F}_b$, where, without loss of generality, we may assume that $d_j \geq 1$ for all $j = 1, 2, \ldots, s$. By comparing components, we obtain

$$\sum_{j=1}^{s} \sum_{i=1}^{d_j} f_i^{(j)} c_{i,r}^{(j)} = 0 \quad \text{for } r = 1, 2, \ldots, m. \tag{5.12}$$

Consider the rational function

$$L := \sum_{j=1}^{s} \sum_{i=1}^{d_j} f_i^{(j)} \frac{x^{k(j,i)}}{p_j(x)^{Q(j,i)+1}} = \sum_{r=1}^{\infty} \left( \sum_{j=1}^{s} \sum_{i=1}^{d_j} f_i^{(j)} c_{i,r}^{(j)} \right) x^{-r},$$

where we used (5.9) and (5.10). From (5.12) we obtain that $\nu(L) < -m$.

Recall that $Q(j, i) = \lfloor (i-1)/e_j \rfloor$. Hence, if we put $Q_j = \lfloor (d_j - 1)/e_j \rfloor$ for $j = 1, 2, \ldots, s$, then a common denominator of $L$ is $g(x) = \prod_{j=1}^{s} p_j(x)^{Q_j+1}$, which implies that $Lg$ is a polynomial. On the other hand,

$$\nu(Lg) < -m + \deg(g) = -m + \sum_{j=1}^{s} (Q_j + 1)e_j \leq -m + \sum_{j=1}^{s} (d_j - 1 + e_j) \leq 0.$$

Thus $Lg = 0$, hence $L = 0$, and therefore

$$\sum_{j=1}^{s} \sum_{i=1}^{d_j} f_i^{(j)} \frac{x^{k(j,i)}}{p_j(x)^{Q(j,i)+1}} = 0.$$

The left-hand side is a partial fraction decomposition of a rational function, and because of its uniqueness all $f_i^{(j)} = 0$. This means that the vectors given in (5.11) are linearly independent over $\mathbb{F}_b$ and the proof is finished. □

► **Remark 5.36** It was shown by J. Dick and H. Niederreiter in 2008 that the quality parameter $t = \sum_{j=1}^{s} (e_j - 1)$ is even exact, i.e., the Niederreiter sequence is a strict digital $(\sum_{j=1}^{s} (e_j - 1), s)$-sequence over $\mathbb{F}_b$.

Now we consider a special choice of polynomials $p_1, \ldots, p_s$. For fixed $s$ and $b$, list all monic irreducible polynomials over $\mathbb{F}_b$ in a sequence according to nondecreasing degrees, and let $p_1, \ldots, p_s$ be the first $s$ terms of this sequence. Then it has been shown by H. Niederreiter that for the strict quality parameter $t$ of the corresponding Niederreiter sequence we have

$$t \leq \begin{cases} 0 & \text{for } s \leq b, \\ s(\log_b s + \log_b \log_b s + 1) & \text{for } s > b. \end{cases} \tag{5.13}$$

This shows that for prime power bases $b$ the necessary condition from Theorem 5.16 for the existence of $(0, s)$-sequences in base $b$ is also sufficient.

**Corollary 5.37.** *Let $s \in \mathbb{N}$ and let $b$ be a prime power. Then a $(0, s)$-sequence in base $b$ exists if and only if $s \leq b$.*

Furthermore, also the necessary condition from Theorem 5.7 is sufficient for prime power bases.

**Corollary 5.38.** *Let $s \in \mathbb{N}$ and let $b$ be a prime power. Then for $m \geq 2$ a $(0, m, s)$-net in base $b$ exists if and only if $s \leq b + 1$.*

*Proof.* Assume that $s \leq b + 1$. Then Corollary 5.37 implies the existence of a $(0, s - 1)$-sequence in base $b$. From this we obtain with Lemma 5.17 a $(0, m, s)$-net in base $b$. The other direction is just Theorem 5.7. □

## Generalised Niederreiter Sequence

In 1993 S. Tezuka proposed a generalisation of Niederreiter's sequence. This sequence differs from the Niederreiter sequence introduced above, in that $x^k$ in (5.9) are replaced by polynomials $y_{j,i,k}(x)$, where $j \in \{1, \ldots, s\}$, $i \in \mathbb{N}$, and $k \in \{0, \ldots, e_j - 1\}$. In order for Theorem 5.35 to apply to these sequences, for each $i \in \mathbb{N}$ and $j = 1, 2, \ldots, s$ the set of polynomials $\{y_{j,i,k}(x) : 0 \leq k < e_j\}$ needs to be linearly independent modulo $p_j(x)$ over $\mathbb{F}_b$. The generalised Niederreiter sequence is then defined by the expansion

$$\frac{y_{j,i,k}(x)}{p_j(x)^i} = \sum_{r=1}^{\infty} a^{(j)}(i, k, r) x^{-r}$$

over the field $\mathbb{F}_b((x^{-1}))$. We then define the matrix $C_j = (c_{i,r}^{(j)})_{i,r \in \mathbb{N}}$ by

$$c_{i,r}^{(j)} = a^{(j)}(Q + 1, k, r) \in \mathbb{F}_b \quad \text{for } j \in \{1, \ldots, s\},\ i, r \in \mathbb{N}, \tag{5.14}$$

where $i - 1 = Qe_j + k$, with integers $Q = Q(j, i)$ and $k = k(j, i)$ satisfying $0 \leq k < e_j$.

**Definition 5.39**

A digital sequence over $\mathbb{F}_b$ generated by the matrices $C_j = (c_{i,r}^{(j)})_{i,r \in \mathbb{N}}$, $j = 1, 2, \ldots, s$, where the $c_{i,r}^{(j)}$ are given by (5.14), is called a *generalised Niederreiter sequence*.

The proof of Theorem 5.35 still applies, hence a generalised Niederreiter sequence is a digital $(t, s)$-sequence over $\mathbb{F}_b$ with $t = \sum_{j=1}^{s}(e_j - 1)$. Under suitable conditions this quality parameter can be shown to be strict.

## Sobol' Sequence

In 1967 I.M. Sobol' was the first to introduce a construction of $(t, s)$-sequences. This sequence, nowadays referred to as *Sobol' sequence*, is the generalised Niederreiter sequence where $b = 2$, $p_1(x) = x$, and, for $j = 2, \ldots, s$, $p_j(x)$ is the $(j - 1)$th primitive polynomial in a sequence of all primitive polynomials over $\mathbb{F}_2$ arranged according to nondecreasing degrees. Further, there are polynomials $g_{j,0}, \ldots, g_{j,e_j-1}$ with $\deg(g_{j,h}) = e_j - h + 1$ such that $y_{j,i,k} = g_{j,k}$ for all $i \in \mathbb{N}$, $k \in \{0, \ldots, e_j - 1\}$, and $j \in \{1, \ldots, s\}$.

► **Remark 5.40** Recall that a polynomial $p \in \mathbb{F}_b[x]$ of degree $m \in \mathbb{N}$ is called a *primitive polynomial* over $\mathbb{F}_b$ if it is the minimal polynomial over $\mathbb{F}_b$ of a primitive element of $\mathbb{F}_{b^m}$. In other words, $p \in \mathbb{F}_b[x]$ of degree $m \in \mathbb{N}$ is primitive if it is monic and irreducible over $\mathbb{F}_b$ and it has a root $\alpha \in \mathbb{F}_{b^m}$ that generates the multiplicative group $\mathbb{F}_{b^m}^*$ of $\mathbb{F}_{b^m}$.

A Sobol' sequence can also be generated in the following way: Let $p_1, \ldots, p_s \in \mathbb{F}_2[x]$ be primitive polynomials ordered according to their degree, and let

$$p_j(x) = x^{e_j} + a_{1,j}x^{e_j-1} + a_{2,j}x^{e_j-2} + \cdots + a_{e_j-1}x + 1 \quad \text{for } j = 1, 2, \ldots, s.$$

Choose odd natural numbers $1 \le m_{1,j}, \ldots, m_{e_j,j}$ such that $m_{k,j} < 2^k$ for $1 \le k \le e_j$, and for all $k > e_j$ define $m_{k,j}$ recursively by

$$m_{k,j} = 2a_{1,j}m_{k-1,j} \oplus \cdots \oplus 2^{e_j-1}a_{e_j-1}m_{k-e_j+1,j} \oplus 2^{e_j}m_{k-e_j,j} \oplus m_{k-e_j,j},$$

where $\oplus$ is the bit-by-bit exclusive-or operator. The numbers

$$v_{k,j} := \frac{m_{k,j}}{2^k}$$

are called *direction numbers*. Then for $n \in \mathbb{N}_0$ with base 2 expansion $n = n_0 + 2n_1 + \cdots + 2^{r-1}n_{r-1}$ we define

$$x_{n,j} = n_0v_{1,j} \oplus n_1v_{2,j} \oplus \cdots \oplus n_{r-1}v_{r,j} \quad \text{and} \quad \boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,s}).$$

The Sobol' sequence is then the sequence of points $(\boldsymbol{x}_n)_{n \in \mathbb{N}_0}$.

## Faure Sequence

In 1982 H. Faure introduced a construction of $(0, s)$-sequences over prime fields $\mathbb{F}_b$ with $s \leq b$. This shows once more that for prime bases $b$ the necessary condition from Theorem 5.16 for the existence of $(0, s)$-sequences in base $b$ is also sufficient. These sequences, nowadays referred to as *Faure sequences*, correspond to the case where the base $b$ is a prime number such that $b \geq s$, $p_j(x) = x - j + 1$ for $j = 1, 2, \ldots, s$ and all $y_{i,j,k}(x) = 1$.

The generating matrices of Faure sequences can also be written down explicitly in terms of the *Pascal matrix*, which is given by

$$P = \begin{pmatrix} \binom{0}{0} & \binom{0}{1} & \binom{0}{2} & \cdots \\ \binom{1}{0} & \binom{1}{1} & \binom{1}{2} & \cdots \\ \binom{2}{0} & \binom{2}{1} & \binom{2}{2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where we set $\binom{k}{l} = 0$ for $l > k$. The generating matrices $C_1, \ldots, C_s$ of the Faure sequence are now given by

$$C_j = (P^\top)^{j-1} \quad (\text{mod } b) \quad \text{for } j = 1, 2, \ldots, s. \tag{5.15}$$

This yields $C_j = (c_{i,r}^{(j)})_{i,r \in \mathbb{N}}$, where

$$c_{i,r}^{(j)} = \begin{cases} 0 & \text{if } 1 \leq r < i, \\ \binom{r-1}{i-1}(j-1)^{r-i} & \text{if } l \geq k, \end{cases}$$

where $\binom{r-1}{i-1}(j-1)^{r-i}$ is evaluated modulo $b$ and where $0^0 := 1$ by convention.

### Example 5.41

In the case $s = b = 2$ we obtain the matrices

$$C_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots \\ 0 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \in \mathbb{Z}_2^{\mathbb{N} \times \mathbb{N}}$$

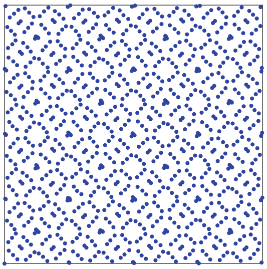

**Fig. 5.3** The first 1024 elements of the Faure sequence for $s = b = 2$

and

$$C_2 = \begin{pmatrix} \binom{0}{0} & \binom{1}{0} & \binom{2}{0} & \binom{3}{0} & \cdots \\ 0 & \binom{1}{1} & \binom{2}{1} & \binom{3}{1} & \cdots \\ 0 & 0 & \binom{2}{2} & \binom{3}{2} & \cdots \\ 0 & 0 & 0 & \binom{3}{3} & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \ldots \\ 0 & 1 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 1 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \in \mathbb{Z}_2^{\mathbb{N}\times\mathbb{N}}.$$

The first 1024 points of the resulting Faure sequence are shown in Fig. 5.3.

## 5.5 Further Reading and Exercises

### Further Reading

The theory of (digital) $(t, m, s)$-nets and $(t, s)$-sequences, respectively, was developed by Niederreiter in 1980s. The original paper is [68]. The early development is very well presented in Niederreiter's book [69]. A more current reference is the book by Dick and Pillichshammer [18] which, beside a detailed introduction, also describes the development since the early 1990s. An introduction into the theory of finite fields can be found in the book by Lidl and Niederreiter [64].

For the proof of the result in Remark 5.9, item 6, we refer to [70] for the upper bound on $t_b(s)$ and to [71] for the lower bound. The currently best asymptotic discrepancy bound for $(t, m, s)$-nets and $(t, s)$-sequences was shown by Faure and Kritzer [28]. For the definition of digital sequences in the broad sense we refer to the book [72] by Niederreiter and Xing. First examples of digital sequences were already introduced, some years before Niederreiter's unifying approach, by Sobol' in [89], called binary $LP_\tau$-sequences, and by Faure [27]. The currently best constructions for digital nets and sequences with respect to the quality parameter are based on methods from Algebraic Geometry. These constructions were developed in a series of papers by Niederreiter and Xing. For an introduction into this subject and an overview we refer to the book by Niederreiter and Xing [72, Chapter 8]. Section 5.4 is based on parts of [18, Chapter 8], to which we refer for further constructions of digital sequences such as, e.g., the ones by Niederreiter and Xing. A discussion of the quality parameter $t$ of the digital sequences in Sect. 5.4 can be found in the paper [17] by Dick and Niederreiter and a proof for the bound (5.13) on the quality parameter of Niederreiter sequences can be found in [69, Theorem 4.54].

An online database for optimal parameters of $(t, m, s)$-nets and $(t, s)$-sequences called "MinT" is available under: http://mint.sbg.ac.at/index.php

## Exercises

5.1 Construct "by hand" a $(0, 2, 2)$-net in base 3.

5.2 Let $b \in \mathbb{N}, b \geq 2$. Show that for any $m, s \in \mathbb{N}, s \geq 2$ and $m \geq 2$ there exists an $(m-1, m, s)$-net in base $b$.

5.3 For $j = 1, \ldots, r$ let $\mathcal{P}_j$ be a $(t_j, m_j, s)$-net in base $b$ and let $b^{m_1} + \cdots + b^{m_r} = b^m$. Show that the point set $\mathcal{P}$ consisting of the elements of $\mathcal{P}_1, \ldots, \mathcal{P}_r$ is a $(t, m, s)$-net in base $b$ with quality parameter

$$t = m - \min_{j=1,2,\ldots,r} (m_j - t_j).$$

5.4 Let $b, k \geq 2$ be integers. Show that every $s$-dimensional elementary interval in base $b^k$ is also an $s$-dimensional elementary interval in base $b$.

5.5 Show that every $(t, \mu k, s)$-net in base $b$ is also a $(\lceil t/k \rceil, \mu, s)$-net in base $b^k$. *Hint:* Use Exercise 5.4.

5.6 Deduce the assertion of Theorem 5.18 for the special case $s = 1$ and $t = 0$ from Proposition 2.31.

5.7 Construct the finite field $\mathbb{F}_{25}$ with 25 elements and determine the operation tables for addition and multiplication.

5.8 Show that the $4 \times 4$ matrices

$$C_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

$$C_3 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad C_4 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

over $\mathbb{F}_2$ generate a digital $(1, 4, 4)$-net over $\mathbb{F}_2$.

5.9 Let $b$ be a prime power and let the $m \times m$ matrices $C_1, \ldots, C_s$ over $\mathbb{F}_b$ be the generator matrices of a (strict) digital $(t, m, s)$-net over $\mathbb{F}_b$. Let $Z$ be a non-singular $m \times m$ matrix over $\mathbb{F}_b$. Show that the matrices $C_1', \ldots, C_s'$, where $C_j' := C_j Z$ for $j = 1, 2, \ldots, s$, generate a (strict) digital $(t, m, s)$-net over $\mathbb{F}_b$, namely the same net but with a different ordering of points.

5.10 Let $b$ be a prime power and $\varphi : \{0, \ldots, b-1\} \to \mathbb{F}_b$ be a bijection with $\varphi(0) = \overline{0}$. For $x = \sum_{i=1}^{\infty} \frac{\xi_i}{b^i} \in [0, 1)$ and $\sigma = \sum_{i=1}^{\infty} \frac{\varsigma_i}{b^i} \in [0, 1)$, where $\xi_i, \varsigma_i \in \{0, \ldots, b-1\}$, we define the ($b$-adic) *digital shifted point* $y$ by $y = x \oplus_{b,\varphi} \sigma := \sum_{i=1}^{\infty} \frac{\eta_i}{b^i}$, where $\eta_i = \varphi^{-1}(\varphi(\xi_i) + \varphi(\varsigma_i))$, and where the "+" is addition in $\mathbb{F}_b$.

For dimensions $s > 1$ let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_s) \in [0, 1)^s$. For $\boldsymbol{x} = (x_1, \ldots, x_s) \in [0, 1)^s$ we define the ($b$-adic) *digital shifted point* $\boldsymbol{y}$ by $\boldsymbol{y} = \boldsymbol{x} \oplus_{b,\varphi} \boldsymbol{\sigma} = (x_1 \oplus_{b,\varphi} \sigma_1, \ldots, x_s \oplus_{b,\varphi} \sigma_s)$.

Let $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{b^m-1}\}$ be a (strict) $(t, m, s)$-net in base $b$ and let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_s) \in [0, 1)^s$. Show that the digitally shifted point set formed by the points $\boldsymbol{y}_n = \boldsymbol{x}_n \oplus_{b,\varphi} \boldsymbol{\sigma}$ for $n = 0, \ldots, b^m - 1$, is again a (strict) $(t, m, s)$-net in base $b$ with probability one with respect to the Lebesgue measure of $\boldsymbol{\sigma}$'s. (If the $\sigma_i$'s have only finitely many $b$-adic digits different from zero, then the assertion is always true.) *Hint:* This is [18, Lemma 4.67].

5.11 Assume that $\mathcal{P} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{b^m-1}\}$ is a digital net over $\mathbb{F}_b$ and let $\oplus_{b,\varphi}$ be the addition from Exercise 5.10. Show that $(\mathcal{P}, \oplus_{b,\varphi})$ is a group. If the points of $\mathcal{P}$ are pairwise different, then this group is isomorphic to $\mathbb{F}_b^m$.

5.12 Remember the definition of Walsh functions ${}_b\mathrm{wal}_k(x)$ from Exercise 3.5. Let $b$ be a prime number and let $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{b^m-1}\}$ be a digital net over $\mathbb{F}_b$ (which we identify with $\mathbb{Z}_b$) with generating matrices $C_1, \dots, C_s \in \mathbb{F}_b^{m \times m}$. Show that for all $k_1, \dots, k_s \in \{0, 1, \dots, b^m - 1\}$ we have

$$\sum_{n=0}^{b^m-1} {}_b\mathrm{wal}_{k_1}(x_{n,1}) \cdots {}_b\mathrm{wal}_{k_s}(x_{n,s}) = \begin{cases} b^m & \text{if } C_1^\top \vec{k}_1 + \cdots + C_s^\top \vec{k}_s = \vec{0}, \\ 0 & \text{otherwise,} \end{cases}$$

where $x_{n,j}$ denotes the $j$th component of $\boldsymbol{x}_n$ and where for $k = \kappa_0 + \kappa_1 b + \cdots + \kappa_{m-1} b^{m-1}$ we denote $\vec{k} = (\kappa_0, \kappa_1, \dots, \kappa_{m-1})^\top$. *Hint:* This is [18, Lemma 4.75] where one can find a proof. *Remark:* This relation is a very crucial tool in the analysis of digital nets and their application for numerical integration. Further information in this direction can be found in [18].

5.13 Let $b$ be a prime power. Show that every upper triangular $\mathbb{N} \times \mathbb{N}$ matrix over $\mathbb{F}_b$ with non-zero diagonal entries generates a digital $(0, 1)$-sequence over $\mathbb{F}_b$.

5.14 Let $b$ be a prime power and let the $\mathbb{N} \times \mathbb{N}$ matrices $C_1, \dots, C_s$ generate a digital $(t, s)$-sequence over the finite field $\mathbb{F}_b$. For any $m \in \mathbb{N}$ consider the left upper $m \times m$ sub-matrices $C_1^{(m)}, \dots, C_s^{(m)}$. Take

$$C_{s+1}^{(m)} := E_m' = \begin{pmatrix} \overline{0} & \overline{0} & \dots & \overline{0} & \overline{1} \\ \overline{0} & & \iddots & \overline{1} & \overline{0} \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ \overline{0} & \overline{1} & \iddots & & \overline{0} \\ \overline{1} & \overline{0} & \dots & \overline{0} & \overline{0} \end{pmatrix} \in \mathbb{F}_b^{m \times m}.$$

Show that $C_1^{(m)}, \dots, C_s^{(m)}, C_{s+1}^{(m)}$ generate a digital $(t, m, s+1)$-net over $\mathbb{F}_b$. *Remark:* Note that this is a "digital version" of Lemma 5.17. Note also the increase of the dimension from $s$ to $s + 1$.

5.15 Prove that for the exponential discrete evaluation we have $\nu(p/q) = \deg(p) - \deg(q)$ for all nonzero polynomials $p, q \in \mathbb{F}_b[x]$. In particular, $\nu(p) = \deg(p)$ for all nonzero $p \in \mathbb{F}_b[x]$

5.16 Consider the first $2^4$ points of a Niederreiter sequence in base 2 and dimension 3. Find the smallest value of $t$ such that these points form a digital $(t, m, s)$-net.

5.17 Show that the Sobol' sequence defined via the recurrence relation is a special case of a generalised Niederreiter sequence, as stated in the section on Sobol' sequences.

5.18 Show that the generating matrices of a Faure sequence defined via the polynomials can be written in terms of Pascal matrices as stated in (5.15).

# A Brief Discussion of the Discrepancy Bounds 6

## 6.1 The Curse of Dimensionality

In many applications the dimension $s$ can be rather large. In this case, the asymptotically almost optimal bounds on the discrepancy which we obtained, e.g., for the Hammersley point set or for $(t, m, s)$-nets soon become useless for a modest number $N$ of points. For example, assume that for every $s, N \in \mathbb{N}$ we have a point set $\mathcal{P}_{s,N}$ in the $s$-dimensional unit cube of cardinality $N$ with star discrepancy of at most

$$D_N^*(\mathcal{P}_{s,N}) \leq c_s \frac{(\log N)^{s-1}}{N}, \tag{6.1}$$

with some $c_s > 0$ that is independent of $N$. Hence for any $\delta > 0$ the star discrepancy behaves asymptotically like $N^{-1+\delta}$, which is the optimal rate of convergence since for dimension $s = 1$ we already have $D_N^*(\mathcal{P}_{1,N}) \geq 1/(2N)$. However, the function $N \mapsto (\log N)^{s-1}/N$ does not start to decrease to zero until $N \geq \exp(s-1)$. For $N \leq \exp(s-1)$ this function is increasing, which means that for cardinality $N$ in this range our discrepancy bounds are useless. In fact, even for moderately large dimension $s$, the value of $\exp(s-1)$ is huge, such that point sets with cardinality $N \geq \exp(s-1)$ cannot be used for practical applications. Therefore, the bound (6.1) is useful only if $N$ is large compared to the dimension $s$.

For practical applications one is interested in the discrepancy of point sets with cardinality $N$ not too large (compared to $s$). To analyse this problem systematically one considers the following quantity.

**Definition 6.1**

For $s, N \in \mathbb{N}$ define the *$N$th minimal star discrepancy* by

$$\mathrm{disc}_\infty(N, s) := \inf_{\substack{\mathcal{P} \subseteq [0,1)^s \\ \#\mathcal{P} = N}} D_N^*(\mathcal{P}).$$

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_6,

For example, one may ask if

$$\lim_{s\to\infty} \mathrm{disc}_\infty(2^s, s) = 0\,? \tag{6.2}$$

In other words, do there exist point sets in dimension $s$ of cardinality $2^s$ (which is already dramatically large for dimensions $s \geq 30$) for which the star discrepancy approaches zero when $s$ tends to infinity? For example, the discrepancy bound from Corollary 5.20 for $(t, m, s)$-nets $\mathcal{P}_s$ in base 2 gives for $m = s$ and $t = t(s) \asymp s$ (which is optimal)

$$D^*_{2^s}(\mathcal{P}_s) \leq \frac{2^{t(s)}}{2^s}\left(1 + \frac{s^{s-1}}{(s-2)!}\right) \sim 2^{t(s)}\left(\frac{\mathrm{e}}{2}\right)^s \sqrt{\frac{2s}{\pi}} \to \infty \quad \text{as} \quad s \to \infty.$$

So the asymptotically excellent discrepancy bounds for $(0, m, s)$-nets cannot help to answer the question (6.2).

As another attempt, let $s, m \in \mathbb{N}$ where $m \geq 2$ and consider the regular lattice $\Gamma_{m,s}$. According to Theorem 2.19, its star discrepancy is given by

$$D^*_N(\Gamma_{m,s}) = 1 - \left(1 - \frac{1}{2m}\right)^s,$$

where $N = m^s$. Let $\varepsilon \in (0, 1)$. Then to achieve a star discrepancy of at most $\varepsilon$ we would require a regular lattice consisting of

$$N \approx \left(\frac{s}{2|\log(1-\varepsilon)|}\right)^s$$

points (i.e., choose $m \approx \frac{s}{2|\log(1-\varepsilon)|}$). This number grows super-exponentially in $s$. For example, for $\varepsilon = 1/3$ one requires $N \approx (1.23s)^s$. Hence, for a star discrepancy of at most $\varepsilon$ we require a cardinality which is exponential in $s$. Such an exponential dependence is called the *curse of dimensionality*, a notion coined by R. Bellman in 1957.

The above discussion shows that although we have excellent asymptotic results for the star discrepancy of many sequences, these results do not help to answer question (6.2). This means that, according to the classical theory, QMC methods could not be expected to work for very high dimensions. Nevertheless, and surprisingly, we know from practical applications that QMC rules often do very well and even work much better than we have any right to expect. As already mentioned in the preface, I.H. Sloan[1] spoke in this context of "The unreasonable effectiveness of QMC". A spectacular and surprising example in this direction was reported by S.H. Paskov and J.F. Traub in 1995. They valuated a financial derivative, a 30-year

[1] Talk at the MCQMC conference in Warsaw, August 15, 2010.

Collateralized Mortgage Obligation, which required the computation of ten 360-dimensional integrals. They tested two QMC rules based on Halton and Sobol' sequences, and compared them with the MC rule. Their experiments showed that both QMC rules outperform the MC rule and that the convergence of the QMC rules is much smoother than that of the MC rule. Following the success of those experiments it is nowadays an important stream of research to explain *why* QMC rules do so well for many high-dimensional applications. Motivated by the idea that perhaps the reason for the success of QMC is that some coordinate directions are more important than others, I.H. Sloan and H. Woźniakowski began to study weighted function spaces. The weights model the behavior of different coordinate directions and under suitable conditions one can break the curse of dimensionality. As an example we already discussed QMC integration in weighted Korobov spaces in Sect. 4.5.

Furthermore, we nowadays also know that the situation for the star discrepancy is not so bad as it seems from the discussion above. Actually the question (6.2) can be answered in the affirmative, and even more is possible. This will be explained in the following section.

## 6.2 Tractability of the Star Discrepancy

In 2001 S. Heinrich, E. Novak, G.W. Wasilkowski, and H. Woźniakowski established the following very surprising result.

**Theorem 6.2 (Heinrich, Novak, Wasilkowski, and Woźniakowski).** *There exists a constant $c > 0$ with the property that*

$$\mathrm{disc}_\infty(N,s) \le c\sqrt{\frac{s}{N}} \quad \textit{for all} \quad N,s \in \mathbb{N}. \tag{6.3}$$

The proof of (6.3) uses deep results from probability theory and is beyond the scope of this book. Later, Ch. Aistleitner showed by a simplified argument that in (6.3) one can choose $c = 10$. Both proofs are unfortunately non-constructive, and until now we do not know point sets for which the bound (6.3) holds. However, using a probabilistic argument, it can be shown that many $N$-element point sets in $[0,1)^s$ satisfy the bound modulo a multiplicative factor greater than one.

**Definition 6.3**

For $s \in \mathbb{N}$ and $\varepsilon > 0$ the *inverse of the star discrepancy* is defined as

$$N_\infty(s,\varepsilon) = \min\{N \in \mathbb{N} : \mathrm{disc}_\infty(N,s) \le \varepsilon\},$$

i.e., the minimal cardinality $N$ that a point set in $[0,1)^s$ must have so that we can achieve a star discrepancy not larger than $\varepsilon$.

It follows from (6.3) that

$$N_\infty(s,\varepsilon) \le Cs\varepsilon^{-2} \tag{6.4}$$

for some positive constant $C$ (for which we may choose 100 according to the result of Aistleitner). Hence, $N_\infty(s,\varepsilon)$ depends only polynomially on $s$ and $\varepsilon^{-1}$. In Information-Based Complexity (IBC) theory such a behaviour is called *polynomial tractability* (cf. Definition 4.23).

Furthermore, it is known that the dependence on the dimension $s$ of the inverse of the star discrepancy in (6.4) cannot be improved. For example, it was shown by A. Hinrichs in 2004 that there exist constants $c, \varepsilon_0 > 0$ such that

$$N_\infty(s,\varepsilon) \ge cs\varepsilon^{-1} \quad \text{for all } \varepsilon \in (0,\varepsilon_0),\ s \in \mathbb{N}$$

and $\mathrm{disc}_\infty(N,s) \ge \min(\varepsilon_0, cs/N)$. The exact dependence of $N_\infty(s,\varepsilon)$ on $\varepsilon^{-1}$ is still an open question which seems to be very difficult.

We will prove a result which also implies polynomial tractability of the star discrepancy, although it is slightly weaker than the one presented in Theorem 6.2. This result was also first shown by Heinrich, Novak, Wasilkowski, and Woźniakowski in 2001.

**Theorem 6.4 (Heinrich, Novak, Wasilkowski, and Woźniakowski).** *We have*

$$\mathrm{disc}_\infty(N,s) \le \frac{2\sqrt{2}}{\sqrt{N}} \left( s \log \left\lceil \frac{s\sqrt{N}}{2(\log 2)^{1/2}} \right\rceil + \log 2 \right)^{1/2} \quad \textit{for all } N, s \in \mathbb{N}, \tag{6.5}$$

*and*

$$N_\infty(s,\varepsilon) \le \lceil 8\varepsilon^{-2}(s \log(\lceil 2s/\varepsilon \rceil + 1) + \log 2) \rceil \quad \textit{for all } s \in \mathbb{N} \textit{ and } \varepsilon > 0. \tag{6.6}$$

Again this is a pure existence result. The idea of the proof of Theorem 6.4 is to show that the probability of randomly picking an $N$-element point set in $[0,1)^s$ with star discrepancy at most the right-hand side of (6.5) is positive. To this end we need an auxiliary result from probability theory which we state here without proof.

**Lemma 6.5 (Höffding's inequality).** *Let $X_1, \ldots, X_N$ be independent random variables satisfying $u_i \le X_i \le v_i$ for some $u_i < v_i$, and with expectation $\mathbb{E}[X_i] = 0$ for all $i = 1, \ldots, N$. Then*

$$\mathbb{P}\left[|X_1 + \cdots + X_N| \geq N\eta\right] \leq 2\exp\left(-2\eta^2 N^2 / \sum_{i=1}^{N}(v_i - u_i)^2\right),$$

*where* $\mathbb{P}[\cdot]$ *is the probability on the probability space supporting* $X_1, \ldots, X_N$.

*Proof of Theorem 6.4*. For given $\delta > 0$ we define $m = \lceil s/\delta \rceil$. Then Remark 2.17 implies that for every $N$-element point set $\mathcal{P} = \{\boldsymbol{t}_0, \ldots, \boldsymbol{t}_{N-1}\}$ in $[0,1)^s$ we have

$$D_N^*(\mathcal{P}) \leq \max_{\boldsymbol{x} \in \mathcal{G}_{m,s}} \left| x_1 \cdots x_s - \frac{1}{N} \sum_{i=0}^{N-1} 1_{[\boldsymbol{0},\boldsymbol{x})}(\boldsymbol{t}_i) \right| + \delta, \tag{6.7}$$

where $\mathcal{G}_{m,s} = \left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m}\right\}^s$. Note that $\#\mathcal{G}_{m,s} = (m+1)^s$.

Let $\tau_0, \ldots, \tau_{N-1}$ be independent and uniformly distributed random variables on $[0,1)^s$. For $\boldsymbol{x} \in [0,1]^s$ let

$$\zeta_{\boldsymbol{x}}^{(i)} = x_1 \cdots x_s - 1_{[\boldsymbol{0},\boldsymbol{x})}(\tau_i) \quad \text{for all} \quad i = 0, \ldots, N-1.$$

Then $\mathbb{E}[\zeta_{\boldsymbol{x}}^{(i)}] = 0$ and $|\zeta_{\boldsymbol{x}}^{(i)}| \leq 1$ for all $i = 0, \ldots, N-1$. Thus we can apply Lemma 6.5 to the sum of the $\zeta_{\boldsymbol{x}}^{(i)}$'s, which yields

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=0}^{N-1} \zeta_{\boldsymbol{x}}^{(i)}\right| \geq \delta\right] \leq 2\exp\left(-\frac{\delta^2 N}{2}\right) \quad \text{for} \quad \boldsymbol{x} \in [0,1]^s.$$

Using the estimate (6.7) it follows that

$$\begin{aligned}\mathbb{P}\left[D_N^*(\{\tau_0, \ldots, \tau_{N-1}\}) \leq 2\delta\right] &\geq \mathbb{P}\left[\max_{\boldsymbol{x} \in \mathcal{G}_{m,s}} \left| x_1 \cdots x_s - \frac{1}{N}\sum_{i=0}^{N-1} 1_{[\boldsymbol{0},\boldsymbol{x})}(\tau_i)\right| \leq \delta\right] \\ &\geq 1 - 2(m+1)^s \exp(-\delta^2 N/2).\end{aligned}$$

The last term is strictly positive if

$$\log 2 + s \log\left(\left\lceil \frac{s}{\delta} \right\rceil + 1\right) - \delta^2 \frac{N}{2} < 0.$$

This inequality is valid for all $\delta > \delta_0 = \delta_0(N, s)$, where

$$\delta_0^2 = \frac{2}{N}\left(s \log\left(\left\lceil \frac{s}{\delta_0} \right\rceil + 1\right) + \log 2\right). \tag{6.8}$$

This implies $\frac{1}{\delta_0} \le \left(\frac{N}{4\log 2}\right)^{1/2}$. Re-inserting into (6.8) leads to

$$\delta_0^2 \le \frac{2}{N}\left(s \log\left\lceil \frac{s\sqrt{N}}{2(\log 2)^{1/2}} \right\rceil + \log 2\right). \tag{6.9}$$

Thus we have shown that for all $\delta > \delta_0$ there exist $\tau_0, \ldots, \tau_{N-1} \in [0,1)^s$ such that $D_N^*(\{\tau_0, \ldots, \tau_{N-1}\}) \le 2\delta_0$. This shows $\mathrm{disc}_\infty(N,s) \le 2\delta_0$, which implies, together with (6.9), the bound (6.5). The proof of (6.6) is left as exercise. □

## 6.3 Further Reading and Exercises

### Further Reading

Nice introductions to the topic are provided by the overview article by Novak [73] and the book by Traub and Werschulz [92]. A report on the 360-dimensional experiment of Paskov and Traub can be found in [81], see also [92, Section 4]. A warmly recommended introduction to the study of weighted spaces in the context of QMC is the paper by Sloan and Woźniakowski [87]. The original proof of Theorem 6.4 can be found in [34]. A simplified version according to Aistleitner with an explicit constant can be found in [1]. The mentioned lower bound on the inverse of the star discrepancy was shown by Hinrichs [38, Theorem 1]. The current state of the art of tractability theory is summarized in the three volumes of Novak und Woźniakowski [74–76]. Concerning the results presented in this chapter we particularly refer to [75, Chapter 9]. A proof of Höffding's inequality can be found in [44].

### Exercises

6.1 Show that $\lim_{s\to\infty} D_{2^s}^*(\Gamma_{2,s}) = 1$.

6.2 Revisit the proof of Theorem 6.4 and show that the probability of randomly picking an $N$-element point set $\mathcal{P}_{N,s}$ in $[0,1)^s$ with star discrepancy

$$D_N^*(\mathcal{P}_{N,s}) \le \frac{2\sqrt{2}}{\sqrt{N}}\left(s \log\left\lceil \frac{s\sqrt{N}}{2(\log 2)^{1/2}} \right\rceil + \log 2 + \eta\right)^{1/2}$$

for arbitrary $\eta > 0$ is larger than $1 - \exp(-\eta)$.

6.3 Prove (6.6). *Hint:* The solution can be found in [18, Theorem 3.54].

6.4 For $s, N \in \mathbb{N}$ define the *$N$th minimal $L_2$ discrepancy* by

$$\mathrm{disc}_2(N,s) := \inf_{\substack{\mathcal{P}\subseteq[0,1)^s\\ \#\mathcal{P}=N}} L_{2,N}(\mathcal{P}).$$

Show that $\mathrm{disc}_2(0,s) = 3^{-s/2}$ and $\mathrm{disc}_2(N,s) \le (2^{-s} - 3^{-s})^{1/2}/N^{1/2}$.
For $\varepsilon > 0$ the *inverse of the $L_2$ discrepancy* is defined as

$$N_2(s,\varepsilon) = \min\{N \in \mathbb{N} : \mathrm{disc}_2(N,s) \le \varepsilon\, \mathrm{disc}_2(0,s)\}.$$

Show that $N_2(s,\varepsilon) \le (3/2)^s \varepsilon^{-2}$. *Remark:* It is known that $N_2(s,\varepsilon) \ge (9/8)^s(1-\varepsilon^2)$. Hence the $L_2$ discrepancy suffers from the curse of dimensionality.

# Basics of Financial Mathematics

# 7

In this chapter we will give some background on mathematical finance, or, to be precise, on the mathematical theory that lies behind derivative pricing. Since the 1980s, financial mathematics has become a huge field that uses methods from many other branches of mathematics, most notably from probability theory. The reliance on probability theory provides us with a wealth of applications for simulation techniques.

We want to give a quick entry into topics of Monte Carlo and quasi-Monte Carlo pricing of financial derivatives. Obviously, a rigorous treatment of the probabilistic theory is well beyond the scope of this introductory text on QMC methods.

## 7.1 Bonds, Stocks and Derivatives

Since we want to discuss valuation of financial derivatives, we have to define at least the most basic ones as well as some of the primary instruments from which they "derive". The explanations given below are more general than would be necessary for a quick introduction to derivative pricing. But we do not want to hide from the reader the fact that most mathematical models in finance are rather coarse, which is done best by showing what *does not* enter the models.

- A *bond* is a financial instrument that pays its owner a fixed amount of money at a pre-specified date in the future. The writer of the bond is usually a big company or a government. The owner effectively becomes a creditor to the writer. If the quality of the debtor is very high, as is the case for many government bonds, the bond can be modeled as a deterministic payment. The bond usually sells at a lower price than its payoff and thus pays *interest*.

  Bond payments may be subject to credit risk, i.e., for some debtors the probability that the debtor will not be able to make the payment in full at the required date cannot be ignored. Fruitful theories have been put forward for valuation of credit risky bonds, and those theories provide fields of applications for simulation methods, too.

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_7,

However, we will not consider that topic and for us a bond will always be non-defaultable. Nevertheless, in practice even then the price of the bond may and will vary over time, due to a change in inflation and other parameters of the general economy.

- A *share* is a financial instrument that warrants its holder ownership of a fraction of a corporation. In particular, the shareholder participates in the business revenue by the means of dividend payments.

  However, dividend payments are not the only possible source of income through a share. At least equally important is the possible gain due to a price change: an investor may buy a share and sell it shortly afterwards if the price has risen. The possible gain is usually one or two orders of magnitude higher than that through interest payments. On the downside, the price change may as well result in a loss. If the shares of a company are traded at a stock exchange, then buying and selling them is particularly convenient and high frequency traders may buy and sell large contingents of shares several times per second.

  The value of a share depends on many things, such as the preferences of the individual agent, the assets of the company, the state of the economy, the future dividend payments, and future interest rates and inflation.

  The so-called *efficient market hypothesis* assumes that the value of the share at a given time is just the market price at that very time. Under this hypothesis it does not make sense to compute the objective value of a share in a mathematical model and compare it to the market price. The only way that a computed value of a share can differ from its market price is that our preferences and/or expectations differ from those of the majority of the market, thus giving a subjective price.
- A *contingent claim* is a financial instrument whose value at a future date can be completely described in terms of the prices of other financial instruments, its so-called underlyings. A typical example is an option on a share. A *European call option* on a share with *maturity* $T$ is a contract which gives its holder the right (but not the obligation) to buy one share from the option writer at some fixed time $T$ in the future at the previously agreed price $K$. In exchange, the option writer is paid a *premium* at the time of writing the option. One also says that the holder has the long position of the option, while the writer has the short position.

  Denote the price of the share at the future date $T$ by $S_T$. Since the option holder may sell the share instantly at the stock exchange, the value of the option at time $T$ is $S_T - K$ if $S_T > K$, and 0 if $S_T \leq K$.

  The left-hand side of Fig. 7.1 shows the payoff of a European call option dependent on the price of the share at maturity. An important feature is the kink at the strike price $K$. It introduces a non-differentiability that makes estimates of the QMC error substantially more difficult.

  Note that the buyer of the option swaps a fixed amount of money – the price of the option at time 0, i.e., the premium – for an uncertain payment in the future. This is a feature that an option shares with an insurance contract, and indeed an option can be used to insure against "unfavorable" events. For example, an investor who intends to buy a share in the future can buy an option now and is therefore insured against any rise of the share price.

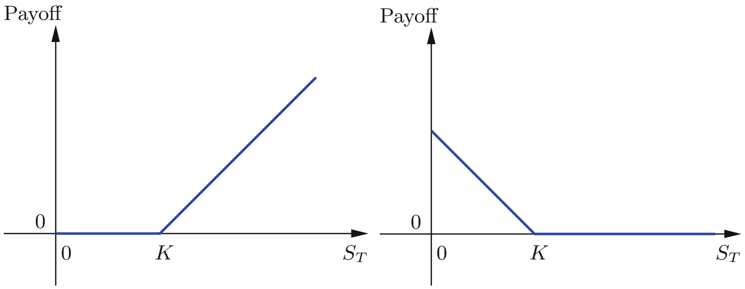

**Fig. 7.1** Payoff of a European call and put option

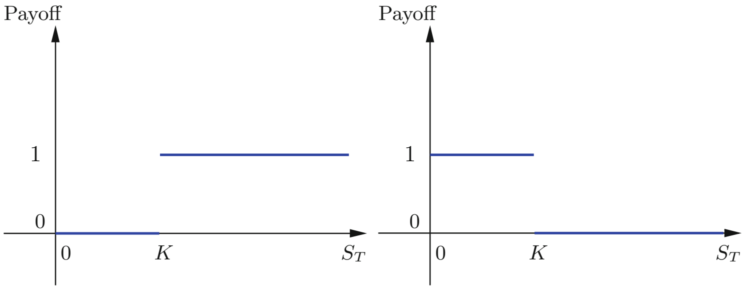

**Fig. 7.2** Payoff of a digital cash-or-nothing call and put option

On the right-hand side of Fig. 7.1 we plot the payoff of a European put option. This is an option which gives its holder the right (but not the obligation) to sell one share to the option writer at some fixed time $T$ in the future, at the previously agreed price $K$. If the share price satisfies $S_T \geq K$ at time $T$, then the option is worthless. But if $S_T < K$, then the option holder may buy the share at the stock exchange at price $S_T$ and sell it immediately to the writer at price $K$, thus realizing a gain of $K - S_T$.

In Fig. 7.2 we show the payoff of another contingent claim, a so-called *digital cash-or-nothing call option*. This option pays a fixed amount of cash at expiry if at that time the price $S_T$ of the underlying is above the strike $K$. We also plot the payoff of the corresponding put option. The digital option serves as an example of a contingent claim with discontinuous payoff.

Since the value of an option is strongly tied to that of the underlying, and in simple models is completely determined by the parameters of the model, an objective value of the option can be computed in these models using arbitrage arguments. Even in more general models a – not necessarily unique – value can be computed with the property that trading the option at that price does not introduce the possibility of arbitrage into the market, that is, the possibility of making risk-less profit above the interest rate.

## 7.2 Arbitrage and the No-Arbitrage Principle

Suppose you are given an option on a stock. You know the specifications of the option and therefore you know the uncertain payoff at its maturity $T$, given the uncertain value of the stock at that particular time. That is, the payoff is of the form $f(S_T)$, where $S_T$ is the share price at time $T$. One is tempted to model $S_T$ as a random variable, then use statistical methods to estimate its parameters from historical stock prices, and a fortiori estimate the value of the option as its discounted expected value. Indeed this is exactly how one usually proceeds in determining the fair stake in a game of chance. We will show in this section that this reasonable program will in general yield a price that is unreasonable from a more basic perspective, in that it allows for risk-less profit.

While general arbitrage theory is well beyond the scope of this book, the underlying principle can be illustrated rather quickly. Assume the following simple market model where we have only two times, 0 and 1, and three instruments, a bond, a share, and a European call option with strike $K = 1$ and maturity $T = 1$. Let $B = (B_t)_{t\in\{0,1\}}, S = (S_t)_{t\in\{0,1\}}, C = (C_t)_{t\in\{0,1\}}$ denote the price processes of the bond, share, option, respectively, and assume the following parameters: $B_0 > 0$, $B_1 = B_0(1 + r)$, $r \geq 0$, $S_0 > 0$, $S_1 = S_0 u$ with probability $p$, and $S_1 = S_0 d$ with probability $1 - p$, where $0 < d < 1 + r < u$. The value of the option at time 1 is $\max(S_1 - K, 0)$, thus $C_1 = \max(S_0 u - K, 0)$ with probability $p$, and $C_1 = \max(S_0 d - K, 0)$ with probability $1 - p$. Suppose we know, for example from statistical studies, the value of $p$.

Following the program proposed above, we would get as the price of the option at time 0 the value

$$\hat{C}_0 = \frac{B_0}{B_1}\mathbb{E}[\max(S_1 - K, 0)] = \frac{1}{1+r}\left(p\max(S_0 u - K, 0) + (1-p)\max(S_0 d - K, 0)\right).$$

However, this formula can give an unreasonable price. Suppose $r = 0$, $u = 2$, $d = \frac{1}{2}$, $S_0 = K = 1$ and $p = \frac{1}{2}$, for which the above formula gives $\hat{C}_0 = \frac{1}{2}$.

Then as a trader we could do the following:

- At time 0, write 4 options and sell them for € 2, borrow one additional Euro to buy three shares. Note that the net investment is zero.
- Now wait until time 1. Then
  - If the share price has gone up, the shares are worth 6. We sell them to get € 6 in Cash. Since the share price $S_T$ (which is 2) is bigger than the strike $K$ (which is 1), the options will be executed, costing us € 4 and we have to pay € 1 back. Thus our strategy leaves us with a net profit of € 1.
  - If the share price has gone down, the options have become worthless. We sell the shares, giving us € $\frac{3}{2}$ and thereby, after paying € 1 back, leaving us with a profit of € $\frac{1}{2}$.

Thus, whatever happens, we are left with a strictly positive profit without having taken any risk and even without initial investment of capital. Such a situation is called an *arbitrage opportunity* and it is usually assumed that such opportunities do not exist in a viable market for the following reason: if you can make a riskless profit of at least € $\frac{1}{2}$, without initial investment, then you can as well make a riskless profit of a million Euros by simply scaling up the above strategy, that is by writing eight million options and so on. In fact, in our model the riskless profit is not limited at all. Obviously, a market with such a feature is useless.[1]

On the other hand, it can easily be shown that there is one and only one option price in this model that does not allow for arbitrage, namely

$$C_0 = \frac{1}{1+r}\big(\tilde{p}\max(S_0u - K, 0) + (1-\tilde{p})\max(S_0d - K, 0)\big),$$

where $\tilde{p} = \frac{1+r-d}{u-d}$. The distinctive feature of $\tilde{p}$ is that $\frac{1}{1+r}(\tilde{p}S_0u + (1-\tilde{p})S_0d) = S_0$, that is, the discounted stock price process is a *martingale* with respect to this new probability. We do not give a precise definition for this. Intuitively, a martingale is a process $X$ such that the conditional expectation of $X_{t+s}$ given the evolution of $X$ up to time $s$, is $X_t$. Thus a martingale is a model for the gain process of a player in a fair game.

To summarize, the arbitrage-free price of an option can be written as the discounted expected value of its payoff, but the expectation has to be taken under a suitable probability measure equivalent – but in general not equal – to the original one.

► **Remark 7.1** No-arbitrage arguments usually make rather strong assumptions:

- Bonds can be bought and sold at the same rate and for every positive maturity. In particular, an investor can lend and borrow money at the same rate.
- There is no restriction on short-selling. That is, an investor may borrow a share and sell it in order to gain from falling prices. At a later time, she or he has to buy the share back at the market price to be able to return it.
- There are no transaction costs.
- There are no liquidity effects: an investor can buy or sell unlimited numbers of securities at the market price without changing the price. In particular, the bid and ask prices coincide.
- All investors have the same information about the market.
- In continuous-time models, an investor may trade continuously at every point in time.

[1] Note, however, that we have tacitly assumed that the strategy scales indefinitely and that, in particular, our buying of shares does not change the market price.

## 7.3 The Black-Scholes Model

The simple model in the preceding section can be extended to an $n$-step setup. Again, the unique arbitrage-free price for a contingent claim is given as a discounted expected value, under some equivalent probability measure, of the claim's payoff at expiry. It is now tempting to let $n$ go to infinity to obtain a continuous-time model. Indeed, this can be done in rigorous fashion so that we arrive at a model of the form

$$\begin{aligned} B_t &= B_0 \exp(rt), \\ S_t &= S_0 \exp(\mu t + \sigma W_t)\,, \end{aligned} \tag{7.1}$$

$t \in [0, T]$, where $W$ is a Brownian motion (sometimes also called "Wiener process", after the mathematician who first described the modern mathematical model for this process), that is, a continuous-time stochastic process with specific properties. The exact mathematical definition of Brownian motion will be given in section "Brownian Motion: Definition and Properties" of Chap. 8, and we invite the reader who is not already familiar with this concept to skip ahead and have a look at Definition 8.15. The model described by (7.1) is often called the *Black-Scholes model.*

As in the one-step model, there exists a probability measure $\tilde{\mathbb{P}}$, equivalent to the original measure $\mathbb{P}$, such that $t \mapsto B_t^{-1} S_t$ becomes a martingale. Under this new probability measure

$$S_t = S_0 \exp\left(\left(r - \frac{\sigma^2}{2}\right) t + \sigma \tilde{W}_t\right)\,,$$

where $\tilde{W}_t = W_t - \frac{1}{\sigma}(r - \mu - \frac{\sigma^2}{2})t$ is a Brownian motion under $\tilde{\mathbb{P}}$. The measure $\tilde{\mathbb{P}}$ can in fact be explicitly given in terms of the Radon-Nikodým derivative, which is given by $\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \exp(\nu W_T - \frac{\nu^2}{2}T)$, with $\nu = \frac{1}{\sigma}(r - \mu - \frac{\sigma^2}{2})$. That means that the expected value $\tilde{\mathbb{E}}[X]$ of any bounded random variable measurable with respect to $\{W_s \;:\; 0 \le s \le T\}$ can be computed by the formula

$$\tilde{\mathbb{E}}[X] = \mathbb{E}\left[\exp\left(\nu W_T - \frac{\nu^2}{2}T\right) X\right],$$

where $\tilde{\mathbb{E}}$ denotes expectation with respect to $\tilde{\mathbb{P}}$.

This new probability measure is now used to price derivatives in this model: if $C$ is some European contingent claim, that is, a derivative whose payoff $C_T$ at time $T$ is a function of $S_t$, $0 \le t \le T$, then its arbitrage-free price at time 0 is given by

$$C_0 = \tilde{\mathbb{E}}[B_T^{-1} C_T]. \tag{7.2}$$

When $C_T$ depends only on finitely many $S_{t_j}$, $j = 1, \dots, m$, then the expectation in (7.2) can be written as an $m$-dimensional integral, which is where QMC enters the game. The details of this will be given in Sect. 8.2.

In our continuous time model we assume that the option can be traded at any time prior to its maturity $T$. For this, the time $t$ analog of (7.2) is

$$B_t^{-1} C_t = \tilde{\mathbb{E}}[B_T^{-1} C_T], \tag{7.3}$$

or $C_t = B_t \tilde{\mathbb{E}}[B_T^{-1} C_T]$.

## Example 7.2

A European Call option on a share with price process $(S_t)_{t \geq 0}$ and with strike $K$ and maturity $T$ has payoff $C_T = \max(S_T - K, 0)$. The pricing equation (7.2) therefore gives the option price in the Black-Scholes model at time $t = 0$ as

$$C_0 = \exp(-rT) \tilde{\mathbb{E}}[\max(S_T - K, 0)].$$

Since

$$S_T = S_0 \exp\left(\left(r - \frac{\sigma^2}{2}\right) T + \sigma \tilde{W}_T\right),$$

and since $(r - \frac{\sigma^2}{2})T + \sigma \tilde{W}_T$ is a normal random variable with mean $(r - \frac{\sigma^2}{2})T$ and variance $\sigma^2 T$, we get

$$\begin{aligned} C_0 &= \frac{\exp(-rT)}{\sqrt{2\pi\sigma^2 T}} \int_{-\infty}^{\infty} \max(S_0 \exp(x) - K, 0) \exp\left(-\frac{(x - (r - \frac{\sigma^2}{2})T)^2}{2\sigma^2 T}\right) \mathrm{d}x \\ &= \frac{\exp(-rT)}{\sqrt{2\pi\sigma^2 T}} \int_{\log(\frac{K}{S_0})}^{\infty} (S_0 \exp(x) - K) \exp\left(-\frac{(x - (r - \frac{\sigma^2}{2})T)^2}{2\sigma^2 T}\right) \mathrm{d}x \,. \end{aligned}$$

It is a nice exercise to compute this integral. It turns out that its value is given by the famous *Black-Scholes option pricing formula*

$$C_0 = S_0 \Phi(d_1) - \exp(-rT) K \Phi(d_2), \tag{7.4}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2) \, \mathrm{d}t$,

$$d_1 = \frac{\log \frac{S_0}{K} + (r + \frac{\sigma^2}{2})T}{\sigma \sqrt{T}} \quad \text{and} \quad d_2 = \frac{\log \frac{S_0}{K} + (r - \frac{\sigma^2}{2})T}{\sigma \sqrt{T}} \,. \tag{7.5}$$

So in this case we get a closed-form formula and there is no need to apply simulation techniques. The price $C_t$ for $0 \leq t \leq T$ can be obtained from Eqs. (7.4) and (7.5) simply by substituting $S_t$ for $S_0$ and $T - t$ for $T$.

Another class of examples for which there often exist closed-form formulas are barrier- and lookback options, where the payoff depends on the maximum or minimum of the price of a single share over a given interval.

We move on to a somewhat harder example.

## Example 7.3

The payoff of an *Asian option* written on a share with price process $(S_t)_{t\in[0,T]}$ depends on the average price over some interval $[T_0, T]$, $T_0 < T$, where $T$ is the expiry date of the option. The payoff of a so-called *fixed strike* Asian call option is given by

$$C_T^{\text{fix}} = \max\left(\frac{1}{T-T_0}\int_{T_0}^{T} S_\tau \,\mathrm{d}\tau - K, 0\right),$$

and the payoff of a so-called *floating strike* Asian call option is given by

$$C_T^{\text{flt}} = \max\left(\frac{1}{T-T_0}\int_{T_0}^{T} S_\tau \,\mathrm{d}\tau - S_T, 0\right).$$

Up to now, nobody has found an explicit formula for either Asian option, but there are rather efficient methods using partial differential equations (PDE) to compute the value. Nevertheless, this example is a nice benchmark for simulation methods.

Because of its simplicity, the Black-Scholes model does not provide us with many interesting examples for simulation. One step towards demanding problems is to look at the $m$-dimensional Black-Scholes model.

Consider $m$ shares $S^1, \ldots, S^m$ whose price processes are given by

$$S_t^j = S_0^j \exp\left(\mu_j t + \sum_{l=1}^{k} \sigma_{jl} W_t^l\right)$$

$t \in [0, T]$, where $W^1, \ldots, W^k$ are $k$ independent Brownian motions and $\sigma = (\sigma_{jl})_{jl}$ is an $m \times k$ matrix. In this model neither the existence, nor the uniqueness of a probability measure that makes each process $(B_t^{-1} S_t^j)_{t\in[0,T]}$ a martingale is granted. In fact, every solution $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^\top$ with $\nu_j \in \mathbb{R}$ of the linear system

$$\sigma\boldsymbol{\nu} = r\mathbf{1} - \boldsymbol{\mu} - \frac{1}{2}\operatorname{diag}(\sigma\sigma^\top) \tag{7.6}$$

gives rise to such a measure, where $\mathbf{1}$ is the $m$-dimensional column vector with all entries equal to 1 and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$. To see this, consider that, if such a solution exists, we can write $\tilde{W}_t^j = W_t^j - \nu_j t$ and we can construct a measure equivalent to $\mathbb{P}$ such that $\tilde{W}^j$ is a Brownian motion under the new measure, just as in the one-dimensional case. In fact we can find one measure $\tilde{\mathbb{P}}$ such that $\tilde{W}^1, \ldots, \tilde{W}^k$ are $k$ independent Brownian motions. The Radon-Nikodým derivative is given by

$$\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}} = \prod_{j=1}^{k} \exp\left(\nu_j W_T^k - \frac{\nu_j^2}{2} T\right) .$$

By substitution we see that the price processes now take on the form

$$S_t^j = S_0^j \exp\left(\left(r - \frac{1}{2}(\sigma\sigma^\top)_{jj}\right) t + \sum_{l=1}^{k} \sigma_{jl} \tilde{W}_t^l\right) ,$$

and hence the discounted price processes remain constant on average, i.e., they are martingales. It can also be shown that no equivalent martingale measure exists if (7.6) has no solution, but we will not attempt this here.

The $m$-dimensional Black-Scholes model is interesting from the point of view of (optimal) portfolio selection, but it also provides us with practical high-dimensional integration problems through derivative pricing. An arbitrage-free price for a derivative with payoff $C_T$ can again be given as

$$C_0 = \exp(-rT)\tilde{\mathbb{E}}\left[C_T\right] ,$$

where $\tilde{\mathbb{E}}$ denotes expectation under a suitable equivalent martingale measure $\tilde{\mathbb{P}}$. As mentioned above, every solution $\nu$ of (7.6) gives rise to such a measure, and in general different measures will give different prices.

In practice, one has to make a reasonable choice for the pricing measure, and usually this is done by *calibrating* the model to market data. However, one must not forget that only one of many possible arbitrage-free prices is computed and that a different price in the market need not imply the existence of arbitrage.

## Example 7.4

A classical basket option on shares with prices $S^1, \ldots, S^m$ and weights $w^1, \ldots, w^m \in \mathbb{R}$ and strike $K$ has payoff

$$\max\left(w^1 S_T^1 + \cdots + w^m S_T^m - K, 0\right) .$$

More complicated dependencies on the price processes can be encountered in practice. In particular, the payoff may depend on the time averages of the price processes. Then the option also has some Asian characteristics. For basket options on several shares the PDE method for pricing soon becomes intractable. Here, we really have to use simulation.

## 7.4 SDE Models

In many models from financial mathematics, the share price process is not given explicitly, but is described via a *stochastic differential equation*, in short SDE.

For example, the one-dimensional Black-Scholes model can also be described by an SDE:

$$\begin{aligned} \mathrm{d}S_t &= \hat{\mu} S_t \,\mathrm{d}t + \sigma S_t \,\mathrm{d}W_t , \\ S_0 &= s_0 \,. \end{aligned}$$

The term $\mathrm{d}W_t$ is interpreted as an infinitesimal increment of a Brownian motion, where a finite increment over the time interval $[t, t+h]$ of Brownian motion is a Gaussian variable with mean 0 and variance $h$ (and thus with standard deviation $\sqrt{h}$). The almost surely unique solution[2] to this SDE with initial value $S_0$ is

$$S_t = s_0 \exp\left( \hat{\mu} t + \sigma W_t - \frac{\sigma^2}{2} t \right) ,$$

so that for $\hat{\mu} = \mu + \frac{\sigma^2}{2}$ we recover the price process from (7.1).

More generally, a model could be defined by an $m+1$-dimensional SDE

$$\begin{aligned} \mathrm{d}S_t &= \mu(t, S_t) \,\mathrm{d}t + \sigma(t, S_t) \,\mathrm{d}W_t , \\ S_0 &= s_0 . \end{aligned} \tag{7.7}$$

where $S = (S^0, \ldots, S^m)$ is an $m+1$-dimensional stochastic process and $s_0 = (s^0, \ldots, s^m) \in \mathbb{R}^{m+1}$. In this general model not all the components need to correspond to share prices or indeed to prices at all. For example, a coordinate process might be a model for an exchange rate or an interest rate. It is assumed that one coordinate is the price of an asset that can function as a numeraire in that it is always positive. Without loss of generality we may assume that $S^0$ is the numeraire. Discounting now means dividing by $S^0$, that is, the discounted process is the $m$-dimensional process $\tilde{S}$ given by

$$t \mapsto (S_t^0)^{-1} (S_t^1, \ldots, S_t^m) \,.$$

[2] This is a consequence of the famous Itô formula from stochastic analysis. In short, the Itô formula states that for a function $f$ which is $C^1$ in the first variable and $C^2$ in the second variable, we have

$$\mathrm{d}f(t, W_t) = \frac{\partial f}{\partial t}(t, W_t) \,\mathrm{d}t + \frac{\partial f}{\partial W}(t, W_t) \,\mathrm{d}W_t + \frac{1}{2} \frac{\partial^2 f}{\partial W^2}(t, W_t) \,\mathrm{d}t \,.$$

Only those components of $\tilde{S}$ that correspond to price processes of traded assets need to be martingales under the equivalent martingale measure.

## Example 7.5

Consider the so-called Heston model (already under an equivalent martingale measure):

$$\begin{aligned}
\mathrm{d}B_t &= rB_t\,\mathrm{d}t,\\
\mathrm{d}S_t &= rS_t\,\mathrm{d}t + \sqrt{V_t}S_t(\rho\,\mathrm{d}W_t^1 + \sqrt{1-\rho^2}\,\mathrm{d}W_t^2),\\
\mathrm{d}V_t &= \kappa(\theta - V_t)\,\mathrm{d}t + \xi\sqrt{V_t}\,\mathrm{d}W_t^1,\\
(B_0, S_0, V_0) &= (b_0, s_0, v_0)\,.
\end{aligned}$$

Here, $r, \kappa, \theta, \xi$ are positive constants, $\mu$ is a real constant, and $-1 < \rho < 1$ is a correlation coefficient. The third component of our process, $V$, is the so-called volatility of the share price and is not a traded asset. The first process $B$ models a bond price and thus serves as the numeraire. It is worth mentioning that, while no explicit solution is known for the SDE, there is a semi-exact formula for the price of a European call option in the Heston model using Laplace inversion.

We do not concern ourselves with the theory of SDEs, since this topic would easily fill another book, and indeed there are many excellent books available on SDEs and many of them cover applications in finance.

From the point of view of MC and QMC simulation it is mainly of interest to know that under suitable regularity requirements on the coefficients of the SDE there exists a unique solution and that under slightly stronger conditions this solution can be approximated by discrete algorithms.

Let $S_T$ be the solution to the SDE at time $T$ and let $\hat{S}_N$ be some approximation to $S_T$ computed on the time grid $0 = t_0 < t_1 < \cdots < t_N = T$ with fineness $\delta = \max_{1\leq k\leq N}(t_k - t_{k-1})$. We say that $\hat{S}_N$ converges to $S_T$ in the *strong sense* with order $\gamma$, if $\mathbb{E}[|S_T - \hat{S}_N|] = O(\delta^\gamma)$.

Sometimes it is enough to compute some characteristics of the solution, like $\mathbb{E}[f(S_T)]$ for a function $f$ belonging to some class $C$. This question is linked to the concept of *weak convergence* of numerical schemes. The benefit is that the weak order of an approximation scheme is usually higher than the strong order of the same scheme.

The most straightforward solution method for SDEs is the *Euler-Maruyama method*: given (7.7), we compute an approximate solution $\hat{S}$ on the time nodes $0, h, \ldots, nh = T$ via

$$\begin{aligned}
\hat{S}_0 &= S_0\\
\hat{S}_{k+1} &= \hat{S}_k + \mu(kh, \hat{S}_k)h + \sigma(kh, \hat{S}_k)\Delta W_{k+1}.
\end{aligned} \tag{7.8}$$

where $\Delta W_{k+1} := W_{(k+1)h} - W_{kh}$. It follows from the definition of Brownian motion, Definition 8.15, that $\Delta W_{k+1}$ is a normal random vector with expectation 0 and covariance matrix $\sqrt{h}I_{m+1}$, where $I_{m+1}$ is the $(m+1)\times(m+1)$ identity matrix. Frequently, (7.8) is therefore stated in the form

$$\hat{S}_{k+1} = \hat{S}_k + \mu(kh, \hat{S}_k)h + \sigma(kh, \hat{S}_k)\sqrt{h}Z_{k+1},$$

where $Z_1, Z_2, \dots$ is a sequence of standard normal vectors. However, we will prefer the original form when using QMC.

Under suitable regularity conditions (Lipschitz continuous in the second variable, sublinear growth with first variable, sufficient smoothness) on the coefficient functions $\mu, \sigma$ of the SDE, the Euler-Maruyama scheme converges in the strong sense with order $\frac{1}{2}$ and in the weak sense with order 1, such that, for sufficiently regular $f$, $\mathbb{E}[f(\hat{S}_N)]$ is a decent approximation to $\mathbb{E}[f(S_T)]$, for sufficiently small $h$.

We report two other schemes for solving autonomous SDEs numerically, which under appropriate conditions on the coefficients converge in the strong sense with order 1. The first is the *Milstein scheme*,

$$\hat{S}_{k+1} = \hat{S}_k + \mu(\hat{S}_k)h + \sigma(\hat{S}_k)\Delta W_{k+1} + \frac{1}{2}\sigma(\hat{S}_k)\sigma'(\hat{S}_k)(\Delta W_{k+1}^2 - h)\,,$$

where $\sigma'$ is the derivative of $\sigma$. The second is an example of a *Runge-Kutta scheme*, with the advantage of not requiring a derivative:

$$\hat{S}_{k+1} = \hat{S}_k + \mu(\hat{S}_k)h + \sigma(\hat{S}_k)\Delta W_{k+1} + \frac{1}{2}(\sigma(Y_k) - \sigma(\hat{S}_k))(\Delta W_{k+1}^2 - h)\frac{1}{\sqrt{h}},$$

where the supporting value $Y_k$ is given by $Y_k = \hat{S}_k + \sigma(\hat{S}_k)\sqrt{h}$.

A problem that can occur in practice is that the simulated path can leave the domain of definition, while the exact solution does not. For example, the approximate stock price and/or the volatility process may become negative, even if the theoretical process is always non-negative. Besides choosing small discretization intervals, there are sophisticated methods for treating this problem, but for which we have to refer the reader to the further reading section.

## 7.5 Further Reading and Exercises

### Further Reading

The use of Brownian motion as a model for stock price evolution dates back to the dissertation thesis of Louis Bachelier, “Théorie de la Spéculation”, from 1900. Bachelier is widely considered to be the founding father of financial mathematics.

The Black-Scholes pricing equation for European call options was obtained by Black, Scholes and Merton in the early 1970s, [9, 66]. They did not use the notions of martingales and equivalent martingale measures, an idea proposed in the late 1970s by Harrison and Kreps [32], but rather derived a partial differential equation (PDE) for the price by finding a replicating strategy for the option. The corresponding PDE and its modern generalizations provide an alternative framework for valuing derivatives that is usually preferred for low dimensions. For a comprehensive treatment of general arbitrage theory see Delbaen and Schachermayer [16]. A nice introduction to quantitative financial mathematics is provided by the book by Albrecher, Binder, Lautscham, and Mayer [2].

The standard reference for numerical treatment of SDEs is the book by Kloeden and Platen [53]. This book also covers stochastic analysis for continuous processes and the theory of SDEs. The regularity conditions needed for strong and weak convergence of the schemes provided in the text together with corresponding proofs can also be found there.

For the problem of possible negativity of the volatility process in the Heston model see again [53] and also [3].

The Black-Scholes model can also be generalized in a different direction: the "driving" stochastic process can be chosen to be a Lévy process, which is a generalization of Brownian motion allowing for jumps in prices. A good reference for Lévy models in finance is, for example, the book by Cont and Tankov [15].

## Exercises

7.1 A share can be viewed as a derivative on itself with maturity $T > 0$ and payoff equal to $S_T$. A bond with $B_0 = 1$ can be viewed as a derivative on a share with deterministic payoff $B_T = \exp(rT)$ at time $T$. The payoff of a portfolio of derivatives on a share with the same maturity is the sum of the single payoffs.

Draw the payoffs of the following derivatives/portfolios:

(a) 1 share;
(b) 1 share and $-K$ bonds;
(c) 1 share and $-1$ call option with strike $K$;
(d) 1 call option and $-1$ put option, both with strike $K$;
(e) 1 share, $-K$ bonds, $-1$ call option, 1 put option, both options with strike $K$.

7.2 Prove the so-called *Put-Call-parity*: Assume a market consisting of a share with price process $(S_t)_{t\in[0,T]}$, European put- and call options on the share with the same strike $K$ and same expiry $T$, and a bond with deterministic price process $(\exp(rt))_{t\in[0,T]}$, for some $r \in (0,\infty)$. Let $(P_t)_{t\in[0,T]}$ and $(C_t)_{t\in[0,T]}$ be the price processes of the put and call option, respectively.

Show that if for some $t \in [0,T)$ we have $S_t - \exp(-r(T-t))K \neq C_t - P_t$, then there exists an arbitrage opportunity in this market that can be realized using a simple buy-and-hold strategy. In other words: in an arbitrage-free market we have the put-call parity

$$S_t - \exp(-r(T-t))K = C_t - P_t \quad \forall t \in [0,T]\,.$$

7.3 A forward contract gives its holder the right and the obligation to buy one share at time $T$ in the future at a pre-arranged price, the "forward price".

(a) Draw the payoff of a forward contract.
(b) Show that the forward price must equal $\exp(rT)S_0$ by constructing a buy-and-hold arbitrage otherwise.

7.4 Using the definition $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp(-y^2/2)\,\mathrm{d}y$, and $\Phi(\infty) = 1$ compute the integrals

$$I_1 = \int_{-\infty}^{x} \exp(\theta y)\exp(-y^2/2)\frac{1}{\sqrt{2\pi}}\,\mathrm{d}y,$$

$$I_2 = \int_{x}^{\infty} \exp(\theta y)\exp(-y^2/2)\frac{1}{\sqrt{2\pi}}\,\mathrm{d}y,$$

$$I_3 = \int_{-\infty}^{\infty} \exp(\theta y)\exp(-y^2/2)\frac{1}{\sqrt{2\pi}}\,\mathrm{d}y.$$

*Hint:* complete the square.

7.5 Let $X$ be a standard normal variable, i.e., $X$ has probability density function $f(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$. Show that for any $\theta \in \mathbb{R}$ we have

$$\mathbb{E}[\exp(\theta X)] = \exp(\theta^2/2).$$

*Hint:* Use Exercise 7.4.

7.6 Deduce from Exercise 7.5 that in the Black-Scholes model

$$\mathbb{E}\left[S_{t_2}|S_{t_1}\right] = \exp(r(t_2 - t_1))S_{t_1}.$$

( $\mathbb{E}[X|Y]$ denotes conditional expectation of the random variable $X$ with respect to the random variable $Y$, which informally means that in evaluating the expectation $Y$ can be treated like a constant.)

7.7 Verify the Black-Scholes option pricing formula Eq. (7.4). *Hint:* use Exercise 7.4.

# Monte Carlo and Quasi-Monte Carlo Simulation 8

Is this chapter we will learn the basics of pricing derivatives using simulation methods. We will consider both Monte-Carlo and quasi-Monte Carlo but – of course – with a special emphasis on the latter. The aim of our exposition is not to provide a large toolbox for the quantitative analyst, but to help getting started with the topic. QMC-pricing is an active area of research by its own and the reader is encouraged to consult the specialized literature. We will, however, take a look at some popular examples that frequently serve as benchmarks for refined simulation techniques.

## 8.1 Non-Uniform Random Number Generation

Most random variables encountered in practical models are not uniformly distributed. We are therefore interested in methods for generating pseudo- or quasi-random numbers with a given distribution from their uniform counterparts.

The most straightforward method is the so-called inversion method, which will be presented in the first subsection.

We are also going to present the class of acceptance-rejection methods for generating random numbers with a given distribution. These methods have the reputation of being generally inapplicable for QMC. We try to give a more differentiated view on that topic.

### Inversion Method

We introduce this method for the special case of invertible cumulative distribution functions and defer the general method to the exercises.

Consider a real random variable $X$. Its *cumulative distribution function* (CDF) $F : \mathbb{R} \to [0, 1]$ is defined by $F(x) = \mathbb{P}[X \leq x]$ for all $x \in \mathbb{R}$. Let us assume that $F$

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6_8,

is invertible, that is, there exists a function $G : (0, 1) \to \mathbb{R}$ with $G(F(x)) = x$ for all $x \in \mathbb{R}$ and $F(G(u)) = u$ for all $u \in (0, 1)$.

Suppose now that the random variable $U$ is uniformly distributed on $(0, 1)$ and consider the new real random variable $Y = G(U)$. Then $Y$ has the same distribution as $X$. To see this, let $y \in \mathbb{R}$. Then

$$\mathbb{P}[Y \leq y] = \mathbb{P}[G(U) \leq y] = \mathbb{P}\big[F(G(U)) \leq F(y)\big] = \mathbb{P}[U \leq F(y)] = F(y),$$

as claimed.

A sufficient condition for a CDF to be invertible is that it has a positive probability density function (PDF) on $\mathbb{R}$ (see Exercise 8.1).

For some distributions the inversion of the CDF is almost trivial. For example, consider a Cauchy-distributed random variable $X$ with parameters $(x_0, \gamma)$, which has PDF

$$f(x) = \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2} \qquad \text{for } x \in \mathbb{R}.$$

The CDF is then given by

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(x)\, \mathrm{d}x = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}.$$

$F$ is obviously invertible, with inverse $F^{-1}(u) = x_0 + \gamma \tan(\pi(u - \frac{1}{2}))$. Thus if $U$ is uniformly distributed on $(0, 1)$, then $x_0 + \gamma \tan(\pi(U - \frac{1}{2}))$ is Cauchy distributed with parameter $(x_0, \gamma)$.

For other distributions the inversion procedure can only be done numerically.

**Definition 8.1**

A random variable $X$ is *normally distributed* with parameters $\mu$ and $\sigma > 0$ if it has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

We write $X \sim N(\mu, \sigma^2)$. If, in addition, $\mu = 0$ and $\sigma = 1$, then we call $X$ a standard normal variable.

The inverse of a normal CDF cannot be given by a formula using only algebraic combinations of elementary functions. However, there are highly accurate methods for approximating the standard normal CDF by rational functions, and implementations are available in every sufficiently popular programming language.

If $Z$ is a standard normal variable and $\mu, \sigma$ are real numbers with $\sigma > 0$, then $X = \sigma Z + \mu$ is a normal random variable with parameters $\mu$ and $\sigma$. So being able to sample from the standard normal distribution is all that is needed for sampling from general normals.

## Acceptance-Rejection Method

Inverting a CDF numerically can be computationally expensive. A very versatile and cheap alternative method for generating a random variable with prescribed PDF $f$ is the *acceptance-rejection method*. For its implementation we need another distribution for which it is cheap to sample from, e.g., via the inversion method. Let $g$ be the PDF of this distribution. Moreover, we need that, for some $0 < c \le 1$, $cf(x) \le g(x)$ for all $x \in \mathbb{R}$.

The algorithm is as follows:

**Algorithm 8.2** (Acceptance-rejection sampling). 1. Generate a sample $Y$ with PDF $g$ and an independent uniform random variable $U$.
2. If $U \le \frac{cf(Y)}{g(Y)}$, set $X = Y$ else go back to step 1.

We call $f$ the target PDF and $g$ the proposal PDF. We give a proof that the algorithm indeed gives a random variable with the desired distribution, and that $c$ should be as big as possible, so that the algorithm stops after as few steps as possible.

**Theorem 8.3.** *Let $X$ be the random variable generated by Algorithm 8.2 and let $N$ be the number of repetitions of Algorithm 8.2 before giving $X$. Then $X$ has PDF $f$, and $N$ has geometric distribution with parameter $c$, i.e. $\mathbb{P}[N = k] = c(1-c)^{k-1}$ for all $k \in \mathbb{N}$. In particular, the algorithm needs on average $\frac{1}{c}$ repetitions to complete.*

*Proof.* Let $Y_1, U_1, Y_2, U_2, \ldots$ be a sequence of independent random variables, where each $Y_j$ is distributed according to PDF $g$ and each $U_j$ is uniformly distributed on $[0, 1]$. Then for any measurable set $A \subseteq \mathbb{R}$

$$\mathbb{P}\left[Y_k \in A \text{ and } U_k \le \frac{cf(Y_k)}{g(Y_k)}\right] = \int_A \mathbb{P}\left[U_k \le \frac{cf(Y_k)}{g(Y_k)} \middle| Y_k = y\right] g(y)\,\mathrm{d}y$$
$$= \int_A \frac{cf(y)}{g(y)} g(y)\,\mathrm{d}y = c\int_A f(y)\,\mathrm{d}y.$$

In particular, $\mathbb{P}\left[U_k \le \frac{cf(Y_k)}{g(Y_k)}\right] = \mathbb{P}\left[Y_k \in \mathbb{R} \text{ and } U_k \le \frac{cf(Y_k)}{g(Y_k)}\right] = c$ for every $k$, so that the random variable $N = \min\left\{k \in \mathbb{N} : U_k \le \frac{cf(Y_k)}{g(Y_k)}\right\}$ has geometric distribution with parameter $c$.

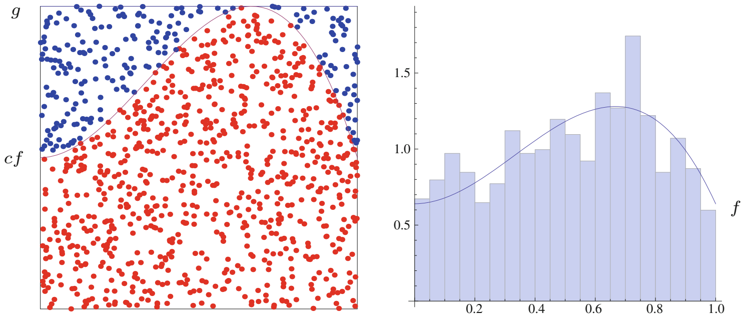

**Fig. 8.1** Illustration of the acceptance-rejection algorithm where $g \equiv 1$ and where 1000 random points in the unit square have been used. On the *right-hand side* we see the histogram of the random points so generated, compared with the probability PDF $f$

According to Algorithm 8.2 we have $X = Y_N$. Therefore,

$$\begin{aligned}
&\mathbb{P}[Y_k \in A \text{ and } N = k] \\
&= \mathbb{P}\left[Y_k \in A \text{ and } U_1 > \frac{cf(Y_1)}{g(Y_1)} \text{ and } \dots U_{k-1} > \frac{cf(Y_{k-1})}{g(Y_{k-1})} \text{ and } U_k \le \frac{cf(Y_k)}{g(Y_k)}\right] \\
&= \mathbb{P}\left[Y_k \in A \text{ and } U_k \le \frac{cf(Y_k)}{g(Y_k)}\right](1-c)^{k-1},
\end{aligned}$$

thanks to the independence of $U_1, \dots, U_k, Y_1, \dots, Y_k$. Now

$$\begin{aligned}
\mathbb{P}[X \in A] &= \sum_{k=1}^{\infty} \mathbb{P}[Y_k \in A \text{ and } N = k] \\
&= \sum_{k=1}^{\infty} c \int_A f(y)\,\mathrm{d}y\,(1-c)^{k-1} = \int_A f(y)\,\mathrm{d}y,
\end{aligned}$$

simply by computing the geometric sum.

The average number of repetitions is $\mathbb{E}[N] = \sum_{k=1}^{\infty} kc(1-c)^{k-1} = 1/c$, which is a simple exercise. □

The method can be illustrated by a picture like the left-hand side of Fig. 8.1. Here $g \equiv 1$ on the unit interval and $c$ is a constant such that $cf(u) \le 1 = g(u)$ for all $u \in [0, 1]$. Random points are sampled in the unit square and those points which lie below the graph of $cf$ are projected onto the $x$-axis. The right-hand side of Fig. 8.1 shows a normalized histogram of the projected points together with the PDF $f$.

## Example 8.4

The PDF of a gamma distributed random variable $X$ with parameters $(a, b)$, is given by

$$f(x) = \begin{cases} \frac{1}{\Gamma(a)} b^a x^{a-1} \exp(-bx) & \text{if } x > 0, \\ 0 & \text{if } x \leq 0, \end{cases}$$

where $\Gamma$ is Euler's gamma function,

$$\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x)\,\mathrm{d}x.$$

If we want to sample from a gamma distribution there is no loss of generality in restricting to the case $b = 1$, since if $X$ is gamma distributed with parameters $(a, 1)$ and $b > 0$, then $bX$ is gamma distributed with parameters $(a, b)$.

Suppose now that $a > 1$. Our idea is to sample from the exponential distribution, which is easy (Exercise 8.8). Thus we want to find some constants $\beta > 0$ and $0 < c \leq 1$ such that $cf(x) \leq \beta \exp(-\beta x)$ and take $g(x) = \beta \exp(-\beta x)$ in Algorithm 8.2.

We need $\frac{c}{\Gamma(a)} x^{a-1} \exp(-x) \leq \beta \exp(-\beta x)$ for all $x \geq 0$. For that, we maximize the function $h(x) = x^{a-1} \exp(-x)/(\Gamma(a)\beta \exp(-\beta x))$ using first-order conditions. We then set $c = h(x_0)^{-1}$, where $x_0$ is the maximizer of $h$. It can be shown that $\beta = a$ gives the lowest rejection rate, i.e., the biggest $c$.

The case where $a < 1$ is slightly more complicated and we leave it for the exercises. The choice of $g$ in the case $a > 1$ is not optimal in that the constant $c$ may be comparatively small. There are refined methods available for choosing good sampling densities for acceptance rejection algorithms.

### Box-Muller Method and Marsaglia-Bray Algorithm

In this section we will present two popular methods for generating standard normal vectors. But first recall the definition of a normal (or Gaussian) random vector:

**Definition 8.5**

A random vector $X = (X_1, \ldots, X_s)$ is said to be *normally distributed* with mean $\boldsymbol{\mu} \in \mathbb{R}^s$ and covariance matrix $\Sigma > 0$ if it has joint PDF

$$f_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^s \det(\Sigma)}} \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^\top}{2}\right), \quad (\boldsymbol{x} \in \mathbb{R}^s)$$

that is, for every measurable set $A \subseteq \mathbb{R}^s$

$$\mathbb{P}[X \in A] = \int_{\mathbb{R}^s} 1_A(x_1, \ldots, x_s) f_X(x_1, \ldots, x_s)\,\mathrm{d}x_1 \ldots \mathrm{d}x_s.$$

Here, $\Sigma > 0$ means that $\Sigma$ has to be positive definite, i.e., $\boldsymbol{x}\Sigma\boldsymbol{x}^\top > 0$ for all $\boldsymbol{x} = (x_1, \ldots, x_s) \in \mathbb{R}^s \backslash \{\boldsymbol{0}\}$.

Consider a 2-dimensional standard normal vector $(X, Y)$, i.e., $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I_s$, the identity matrix. Then

$$\begin{aligned}\mathbb{P}\left[\sqrt{X^2+Y^2} \le r\right] &= \frac{1}{2\pi}\int_{-r}^{r}\int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} \exp(-(x^2+y^2)/2)\,\mathrm{d}x\ \mathrm{d}y \\ &= \frac{1}{2\pi}\int_{0}^{r}\int_{0}^{2\pi} \rho\exp(-\rho^2/2)\,\mathrm{d}\varphi\ \mathrm{d}\rho \\ &= 1-\exp(-r^2/2).\end{aligned}$$

It follows that the modulus of $(X, Y)$ has CDF $F_R(r) = 1 - \exp(-r^2/2)$. But that means that we can generate a random radius by inversion of $F_R$, $F_R^{-1}(u) = \sqrt{-2\log(1-u)}$. Since the PDF has rotational symmetry, this gives us a method for generating independent pairs of standard normal variables.

**Algorithm 8.6** (Box-Muller). 1. Generate two independent $U[0, 1)$ random samples $U, V$;
2. Let $R = \sqrt{-2\log(1-U)}$;
3. Let $X = R\cos(2\pi V)$ and $Y = R\sin(2\pi V)$;
4. Return $(X, Y)$.

There is an acceptance-rejection-type variant of the Box-Muller method which is known as Marsaglia-Bray algorithm:

**Algorithm 8.7** (Marsaglia-Bray). 1. Generate two independent $U[0, 1)$ random samples $U, V$;
2. Let $U_1 = 2U - 1$ and $V_1 = 2V - 1$;
3. If $U_1^2 + V_1^2 \ge 1$ reject $(U, V)$ and start from the beginning;
4. Else let $S = U_1^2 + V_1^2$;
5. If $S = 0$ set $(X, Y) = (0, 0)$;
6. Else set $X = U_1\sqrt{-2\log(S)/S}$ and $Y = V_1\sqrt{-2\log(S)/S}$;
7. Return $(X, Y)$.

**Theorem 8.8.** *The vector $(X, Y)$ generated by either of the Algorithms 8.6 or 8.7 has standard normal distribution.*

We leave the proof that $(X, Y)$ are independent standard normal variables to the reader.

Both the Box-Muller method and the Marsaglia-Bray method are very appealing due to their elegance. We will argue later that using acceptance-rejection methods with QMC has its pitfalls. The Box-Muller method is a special kind of inversion method, but it does not map one coordinate to one coordinate; rather, it generates two coordinates at a time.

Though the approximation of the inverse CDF by rational functions is probably less elegant, and the corresponding algorithm is certainly harder to memorize, it is computationally only marginally more demanding.

## Importance Sampling

For some densities it is very hard – if not impossible – to invert the CDF exactly, and frequently it is very expensive to do so numerically.

On the other hand, it is not always necessary to generate exactly from the given distribution; rather, one samples from a distribution that is close (in some sense that remains to be made precise) to it and adjusts for the error made. This method is called *importance sampling* or, in the present context, *smooth rejection*.

We present the idea in a one-dimensional setup, the general case is straightforward. Consider a random variable $X$ with PDF $f_X$ and suppose we want to compute $\mathbb{E}[h(X)]$ for some function $h$. Let $F_X$ denote the corresponding CDF, $F_X(x) = \int_{-\infty}^{x} f_X(\xi)\,\mathrm{d}\xi$. Normally, we would compute

$$\mathbb{E}[h(X)] \approx \frac{1}{N}\sum_{n=1}^{N} h(F_X^{-1}(U_n))$$

using the inversion method, where $U_1, \dots, U_N$ is a uniform pseudo-random sequence or a low-discrepancy sequence.

Suppose now that we do not know how to (cheaply) invert $F_X$. However, assume that there is another PDF $g$ for which the corresponding CDF $G$, given by $G(x) = \int_{-\infty}^{x} g(\xi)\,\mathrm{d}\xi$, is easily inverted. Then

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} h(x)\frac{f_X(x)}{g(x)} g(x)\,\mathrm{d}x = \mathbb{E}\left[h(Y)\frac{f_X(Y)}{g(Y)}\right],$$

where $Y$ is a random variable with PDF $g$. Now the last expected value can be computed by sampling from the PDF $g$ using the inversion method:

$$\mathbb{E}\left[h(Y)\frac{f_X(Y)}{g(Y)}\right] \approx \frac{1}{N}\sum_{n=1}^{N} h\left(G^{-1}(U_n)\right)\frac{f_X(G^{-1}(U_n))}{g(G^{-1}(U_n))}.$$

When choosing the PDF $g$, one must take care not to make the integrand less regular. For example, suppose we want to compute $\mathbb{E}[\sin(X)]$, where $X$ has a hyperbolic distribution,

$$f_X(x) = \left(\int_{\mathbb{R}} \exp(-\sqrt{1+t^2})\,\mathrm{d}t\right)^{-1} \exp\left(-\sqrt{1+x^2}\right).$$

Let $Y$ have standard normal distribution, $g(y) = \exp(-\frac{y^2}{2})/\sqrt{2\pi}$. Then

$$\mathbb{E}[\sin(X)] = \mathbb{E}\left[\sin(Y)\frac{f_X(Y)}{g(Y)}\right],$$

but $f_X(y)/g(y)$ grows like $\exp(y^2/2)$ as $|y| \to \infty$. Thus we have sacrificed boundedness of our integrand. Taking a double exponential PDF instead, i.e., $g(x) = \frac{1}{2}\exp(|x|)$ preserves boundedness. In turn, this particular $g$ introduces a kink into the integrand, i.e., a place where the integrand is not differentiable. Usually that does less harm than unboundedness. If need arises, the kink can be dealt with by using a PDF made of three different parts.

► **Remark 8.9** When using Monte Carlo, one may also sample from the PDF $g$ using the rejection method. The goal of importance sampling is then to reduce the variance of the integrand to speed up convergence.

► **Remark 8.10** Importance sampling is particularly useful for sampling from a random vector whose components have a complicated correlation structure.

## Whether or Not – and How – to Use Acceptance-Rejection With QMC

We already mentioned that using acceptance-rejection algorithms with QMC is not straightforward. Indeed, the view that those algorithms should not (or cannot) be combined at all is not uncommon.

The question of whether or not acceptance-rejection algorithms can be used together with QMC depends, of course, on the context. Before we approach an answer to that question, we first discuss how the algorithm can be applied and – equally important – how it cannot.

But first consider MC simulation. We are given a pseudo-random number generator that gives us a sequence $(U_n)_{n\in\mathbb{N}_0}$ of numbers in $[0, 1)$ which are, ideally, indistinguishable from a truly random sequence of independent random variables with uniform distribution on $[0, 1)$. From the sequence $(U_n)_{n\in\mathbb{N}_0}$ we now compute a sequence $(X_n)_{n\in\mathbb{N}_0}$ of independent random variables with given distributions by using the acceptance-rejection algorithm with $cf \leq g$. According to Theorem 8.3 we need on average $N/c$ elements of the sequence $(U_n)_{n\in\mathbb{N}_0}$ to generate $X_0, \dots, X_{N-1}$. Which of the uniform distributions will be used in the generation of a particular $X_n$ cannot be known in advance (and is not important for MC). If in the simulation we need random $d$-dimensional vectors, this can be done simply be grouping $X_0, X_1, \dots$ into $d$-tuples.

For QMC the situation is quite different. We have a low-discrepancy sequence $(\boldsymbol{u}_n)_{n\in\mathbb{N}_0}$ in the $s$-dimensional unit cube. This sequence has inherently an $s$-dimensional structure, and when generating a $d$-dimensional vector with given distribution and independent components from it, we need to take care not to loose that structure. We provide different approaches as examples.

## Example 8.11

The most straightforward transfer of the corresponding MC scheme would be to use a one-dimensional low-discrepancy sequence, just like we use a one-dimensional random number generator for MC. This will give very bad results: consider, for example, the case where we want to construct two independent uniforms. Then the densities $f$ and $g$ in the acceptance-rejection algorithms coincide with $c = 1$.

Thus, no pair is rejected. Hence, applying the acceptance-rejection algorithm to the one-dimensional sequence $(x_n)_{n\in\mathbb{N}_0}$ will result in the two-dimensional sequence $((x_0, x_2), (x_4, x_6), (x_8, x_{10}), \ldots)$.

For example, if $(x_n)_{n\in\mathbb{N}_0}$ is the van der Corput sequence in base 2, then $x_{4k} \in [0, \frac{1}{4})$ and $x_{4k+2} \in [\frac{1}{4}, \frac{1}{2})$ for $k = 0, 1, 2, \ldots$. Thus all points of the 2-dimensional sequence will lie in one rectangle of area $2^{-4}$ and consequently are not even uniformly distributed.

## Example 8.12

Another straightforward transfer of the corresponding MC scheme suggests to use a $s = 2d$-dimensional low-discrepancy sequence, and check for every pair of coordinates whether it is rejected. If no pair is rejected, we get a vector with the desired distribution.

We provide a simple example. Let $f$ be the PDF of the gamma distribution with parameter $a$, $f(x) = x^{a-1}\exp(-x)/\Gamma(a)$, and let $g$ be the PDF of the exponential distribution with parameter $b$, $g(x) = b\exp(-bx)$. If $a = 1.2$ and $b = 0.85$, then $bf(x) \le g(x)$. We apply the rejection algorithm to some lattice rule in dimension 4, that is, the first two components are used to generate the first gamma variable while the last two components will be used to generate the second one. If rejection occurs in generating either of the components, the whole 4-dimensional point is rejected.

The resulting sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ will have the distribution of two independent $\Gamma(1.2, 0.85)$ variables, so applying the corresponding CDF to the components gives a sequence $(\boldsymbol{u}_n)_{n\in\mathbb{N}_0}$ which is uniform in the unit square. However, there is no reason why it should have any additional structure, like having low-discrepancy. Figure 8.2 compares a 2-dimensional point set generated from a 4-dimensional lattice with the first and third component of the same lattice. Of course, the whole number of points in the lattice must be greater than the number plotted so we can show an equal number of points in both plots.

It can be seen that, while the points on the left still bear some similarities to the lattice on the right, they also show some characteristics typical for random numbers, like the presence of clusters and holes. Nevertheless, we may hope that the result is not worse than when using MC, especially when $c$ is close to 1.

What else could we do? A key observation is that the acceptance-rejection algorithm does not require the densities $f$ and $g$ to be one-dimensional.

## Example 8.13

Suppose we want to generate a $d$-dimensional vector with independent components, each distributed according to the one-dimensional PDF $f$. So, the vector is distributed according to

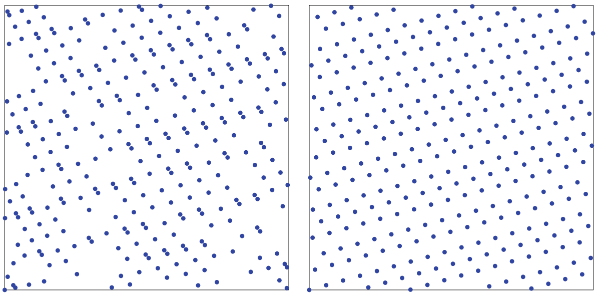

**Fig. 8.2** Comparison of rejected and original lattice points

the $d$-dimensional PDF $\boldsymbol{f}(x_1,\ldots,x_d) = \prod_{k=1}^{d} f(x_k)$. As usual, we assume that there exists a one-dimensional PDF $g$ from which it is easy to sample and for which there exists $c$ with $cf \leq g$. Then $\boldsymbol{g}(x_1,\ldots,x_d) = \prod_{k=1}^{d} g(x_k)$ is a $d$-dimensional PDF from which it is easy to sample (just do it coordinate-wise!) and for which $c^d\boldsymbol{f} \leq \boldsymbol{g}$. By "easy to sample from" we now explicitly mean that we have a function $H$ such that $H(U)$ is distributed according to $g$ for uniform $U$.

Thus we may apply Algorithm 8.2 in the following way: take a low-discrepancy sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ in the $s = (d+1)$-dimensional unit cube, compute $Y_n = (H(x_{n,1}),\ldots,H(x_{n,s}))$ for every $n$, and use only those points for which

$$x_{n,s} \leq \frac{c^d\boldsymbol{f}(Y_n)}{\boldsymbol{g}(Y_n)},$$

discard all others.

If we map the random vector obtained in this way back to the unit cube the picture is similar to Fig. 8.2 (for $d = 2$). The benefit is that we need only a $(d+1)$-dimensional low-discrepancy sequence instead of a $2d$-dimensional one.

Figure 8.3 shows graphs similar to those from Fig. 8.1, but with random points replaced by points from the Sobol' sequence. We see that the normalized histogram matches the PDF very well compared to the same histogram for random points. When the figures are compared, dismissing the use of QMC with acceptance-rejection altogether does not seem justified.

An issue with the acceptance-rejection method is that it sometimes makes the dependence of the result of a MC simulation on the model parameters less smooth. It is clear that the result of a true MC simulation is by definition stochastic. If one searches for model parameters which minimize (a function of) the integral that is computed, then this has the practical drawback that, for example, Newton's method cannot be used. In practice it is therefore common to fix the random sequence for the MC simulation, i.e., the random generator is started afresh for each set of parameters. In this sense the MC method becomes similar to QMC in that the point set is now deterministic.

However, if acceptance-rejection is used for the generation of random variables, then the integral as a function of the model parameters can still become noisy.

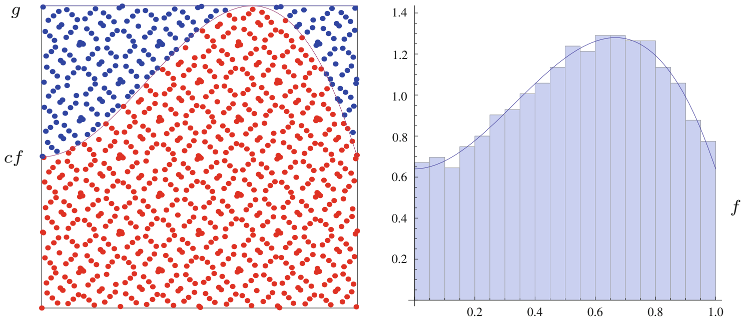

**Fig. 8.3** Illustration of the acceptance-rejection algorithm where $g \equiv 1$ and where 1000 points from the Faure sequence have been used instead of random ones. On the *right-hand side* we see the histogram of the points so generated, compared with the probability PDF function $f$. The histogram should be compared to that in Fig. 8.1

## Example 8.14

Let $(X_j^\lambda)_{j=1,\dots,s}$ be a sequence of i.i.d. Gamma$(\lambda, 1)$ random variables and $Y^\lambda = \sum_{j=1}^s X_j^\lambda$. Let further $\overline{\lambda} \in \mathbb{R}$ be fixed and $f(y) := y - \overline{\lambda}s$.

We want to approximate

$$\alpha(\lambda) = \mathbb{E}\left[f(Y^\lambda)\right]$$

by the estimator

$$\hat{\alpha}_N(\lambda) = \frac{1}{N}\sum_{j=0}^{N-1} f(Y_j^\lambda)$$

for different values of $\lambda$, $\lambda \in (\overline{\lambda} - \epsilon, \overline{\lambda} + \epsilon)$. We compare the following scenarios:

1. We use a MC method and acceptance-rejection with a suitable exponential distribution as dominating function. The pseudo-random number generator is restarted for every choice of $\lambda$, so that in fact we use the same sequence for every integral evaluation.
2. We use a low-discrepancy QMC sequence (here: a Sobol' sequence) together with the inverse transform method.

We plot the function $\hat{\alpha}_N$ with $s = 5, N = 1{,}024, \overline{\lambda} = 2$ and $\epsilon = 0.2$, where $\lambda$ changes in steps of 0.001. In the left plot of Fig. 8.4 one can see quite a bit of noise, while the right one is very smooth.

Smoothness is of importance if, for example, one wants to minimize $\alpha(\lambda)$. An application would be calibration of a financial model to market data.

Of course, this introduction of noise need not happen in any case. For example, when one generates $N(\mu, \sigma^2)$ random variables using Marsaglia-Bray, one in fact generates standard normal variables and transforms them into the required variables afterwards. Thus, different parameters $\mu, \sigma$ do not result in different input variables being rejected.

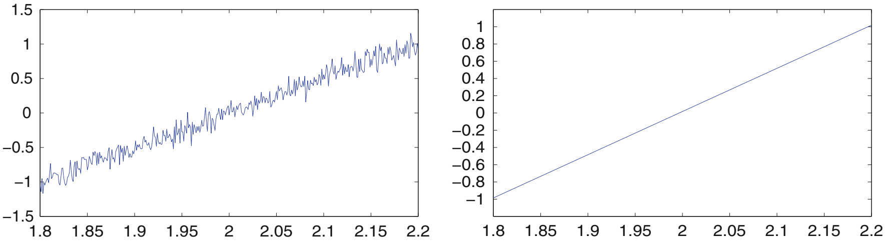

**Fig. 8.4** Acceptance-rejection method with a fixed MC point set compared to the inverse transform method with a QMC point set. $\hat{\alpha}_N$ is plotted as a function of $\lambda$

The observant reader will have noticed that the above reservation about the use of acceptance-rejection algorithms is not confined to QMC, but holds for MC as well.

So what is the answer to the questions posed in the title? Acceptance-rejection algorithms destroy some of the structure of low-discrepancy sets/sequences. There is no reason to believe that the resulting point set still has a particularly low discrepancy. So the recommendation is to avoid acceptance-rejection if the random variables can be generated by the inversion method with reasonable effort, or if importance sampling can be used. A situation where this may fail to be the case is when the PDF $f$ is not explicitly known, or when $f$ is only known up to a scalar factor. In many cases there is still a good chance that QMC performs better than MC.

## 8.2 Generation of Brownian Paths

Many problems from finance, but also from physics, encompass phenomena which are modeled by Brownian motion. In this section we give the basic definition and describe some methods for sampling from Brownian motion.

### Brownian Motion: Definition and Properties

**Definition 8.15**

A *standard Brownian motion* $W$ in $\mathbb{R}^m$ is a stochastic process in continuous time, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, having the following properties:

1. $W_0 = 0$ almost surely;
2. $W$ has *stationary increments*, that is, for any $t_1, t_2 \geq 0$ the random variables $W_{t_2} - W_{t_1}$ and $W_{t_2-t_1}$ have the same distribution;

3. $W$ has *independent increments*, that is, for any $d \in \mathbb{N}$ and any $t_1, \dots, t_d \in [0, \infty)$ with $t_0 := 0 < t_1 < t_2 < \dots < t_d$, the random variables $W_{t_1} - W_{t_0}, \dots, W_{t_d} - W_{t_{d-1}}$ are independent;
4. $\sqrt{1/t}\, W_t$ is a standard normal $\mathbb{R}^m$-valued random variable for every $t \geq 0$;
5. $W$ has *continuous paths*, that is, for each $\omega \in \Omega$ the mapping $t \mapsto W_t(\omega)$ is continuous.

We concentrate on the case of one-dimensional Brownian motion, i.e., $m = 1$. For applications we usually only need to evaluate the Brownian path at finitely many nodes $t_1, \dots, t_d$. We therefore define a *discrete Brownian path* with discretization $0 < t_1 < \dots < t_d$ as a Gaussian vector $(W_{t_1}, \dots, W_{t_d})$ with mean zero and covariance matrix

$$\big(\min(t_j, t_k)\big)_{j,k=1}^{d} = \begin{pmatrix} t_1 & t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & t_2 & \dots & t_2 \\ t_1 & t_2 & t_3 & \dots & t_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & t_3 & \dots & t_d \end{pmatrix}.$$

## Classical Constructions

There are three classical constructions of discrete Brownian paths:

- The *forward method*, also known as *step-by-step method* or *piecewise method*;
- the *Brownian bridge construction* or *Lévy-Ciesielski construction*;
- the *principal component analysis construction* (PCA construction).

The forward method is also the most straightforward one: given a standard normal vector $X = (X_1, \dots, X_d)$, the discrete Brownian path is computed inductively by

$$W_{t_1} = \sqrt{t_1} X_1 , \quad W_{t_{k+1}} = W_{t_k} + \sqrt{t_{k+1} - t_k} X_{k+1} .$$

Using that $\mathbb{E}[X_j X_k] = \delta_{jk}$, it is easy to see that $(W_{t_1}, \dots, W_{t_d})$ has the required correlation matrix. Besides its simplicity, the main attractivity of the forward method lies in the fact that it is very efficient: given that the values $\sqrt{t_{k+1} - t_k}$ are pre-computed, generation of a path takes only generation of the standard normal vector plus $d$ multiplications and $d - 1$ additions.

An alternative construction is the Brownian bridge construction, which allows the values $W_{t_1}, \dots, W_{t_d}$ to be computed in any given order. The main observation that makes this possible is the following lemma, the proof of which is left to the reader (Fig. 8.5).

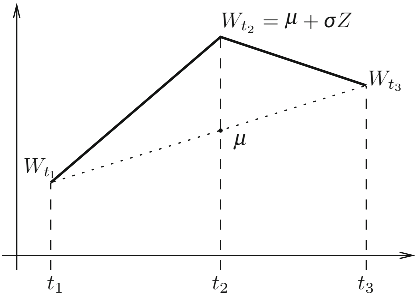

**Fig. 8.5** Illustration of Lemma 8.16. $Z$ is a standard normal variable

**Lemma 8.16.** *Let $W$ be a Brownian motion and let $t_1 < t_2 < t_3$. Then the conditional distribution of $W_{t_2}$ given $W_{t_1}, W_{t_3}$ is $N(\mu, \sigma^2)$, with*

$$\mu = \frac{t_3 - t_2}{t_3 - t_1} W_{t_1} + \frac{t_2 - t_1}{t_3 - t_1} W_{t_3} \text{ and } \sigma^2 = \frac{(t_3 - t_2)(t_2 - t_1)}{t_3 - t_1}.$$

Suppose the elements of $(W_{t_1}, \ldots, W_{t_d})$ need to be computed in the order $W_{t_{\pi(1)}}, W_{t_{\pi(2)}}, \ldots, W_{t_{\pi(d)}}$ for some permutation $\pi$ of $d$ elements. In computing $W_{t_{\pi(j)}}$ we need to take into account the previously computed elements, and at most two of those are of relevance, the one next to $\pi(j)$ on the left and the one next to $\pi(j)$ on the right: define for every $j \in \{1, \ldots, n\}$ two sets,

$$L(j) := \{k : k < \pi(j) \text{ and } \pi^{-1}(k) < j\},$$
$$R(j) := \{k : k > \pi(j) \text{ and } \pi^{-1}(k) < j\}.$$

Thus $L$ contains all the indices $k$ that are smaller than $\pi(j)$ and for which $W_{t_k}$ has already been constructed, while $R$ contains all the indices $k$ that are greater than $\pi(j)$ and for which $W_{t_k}$ has already been constructed. Now define

$$l(j) := \begin{cases} 0 & \text{if } L(j) = \emptyset, \\ \max L(j) & \text{if } L(j) \neq \emptyset, \end{cases}$$

$$r(j) := \begin{cases} \infty & \text{if } R(j) = \emptyset, \\ \min R(j) & \text{if } R(j) \neq \emptyset, \end{cases}$$

and set $W_{t_0} = 0$,

$$W_{t_{\pi(j)}} := \begin{cases} W_{t_{l(j)}} + \sqrt{t_{\pi(j)} - t_{l(j)}} X_j & \text{if } r(j) = \infty, \\ \frac{t_{r(j)} - t_{\pi(j)}}{t_{r(j)} - t_{l(j)}} W_{t_{l(j)}} + \frac{t_{\pi(j)} - t_{l(j)}}{t_{r(j)} - t_{l(j)}} W_{t_{r(j)}} & \\ \qquad + \sqrt{\frac{(t_{\pi(j)} - t_{l(j)})(t_{r(j)} - t_{\pi(j)})}{t_{r(j)} - t_{l(j)}}} X_j & \text{if } r(j) < \infty, \end{cases}$$

where $X = (X_1, \ldots, X_d)$ is a standard normal random vector.

It is easy to check that the vector $(W_{t_1}, \dots, W_{t_d})$ constructed in that way has again covariance matrix $(\min(t_j, t_k))_{j,k}$. The functions $l$ and $r$, as well as the coefficients of $W_{t_{l(j)}}$, $W_{t_{r(j)}}$, $X_j$, do not depend on the random vector $X$, so they can be pre-computed. In some special cases the functions $l$ and $r$ can be given explicitly, for example if the $\pi(t_j)$ are the first $n$ elements of the van der Corput sequence. Therefore the Brownian bridge construction is also very efficient: besides the generation of the vector $X$, computation of one sample uses at most $2d$ additions and $3d$ multiplications.

Moreover, we see that the forward construction is a special case of the Brownian bridge construction where $\pi$ is the identical permutation.

The PCA construction exploits the fact that the correlation matrix of $(W_{t_1}, \dots, W_{t_d})$ is positive definite and can therefore be written in the form $VDV^{-1}$ for a diagonal matrix $D$ with positive entries and an orthogonal matrix $V$. $D$ can be written as $D = D^{\frac{1}{2}} D^{\frac{1}{2}}$, where $D^{\frac{1}{2}}$ is the element-wise positive square root of $D$. Now the PCA construction from a standard normal random vector $X$ is given by

$$(W_{t_1}, \dots, W_{t_d})^\top = VD^{\frac{1}{2}} X^\top.$$

The proof that $(W_{t_1}, \dots, W_{t_d})$ is a discrete Brownian path on the nodes $t_1, \dots, t_d$ is an easy consequence of the following lemma:

**Lemma 8.17.** *Let $A, \Sigma$ be any $d \times d$ matrices with $AA^\top = \Sigma$ and let $X$ be a standard normal vector. Then $Y^\top = AX^\top$ is a normal vector with covariance matrix $\Sigma$, i.e., $\mathbb{E}[Y_j Y_k] = \Sigma_{jk}$.*

*Proof.* Since every linear combination of independent normal random variables is still normal, $AX^\top$ is normal. We compute the covariance matrix:

$$\begin{aligned}\mathbb{E}\left[(AX^\top)_j (AX^\top)_k\right] &= \mathbb{E}\left[\sum_{l=1}^{d} A_{jl} X_l \sum_{m=1}^{d} A_{km} X_m\right] \\ &= \sum_{l=1}^{d} \sum_{m=1}^{d} A_{jl} A_{km} \mathbb{E}\left[X_l X_m\right] \\ &= \sum_{l=1}^{d} A_{jl} A_{kl} = (AA^\top)_{jk} = \Sigma_{jk}.\end{aligned}$$

□

The disadvantage of the PCA for high-dimensional problems is that the matrix-vector multiplication, having computational complexity $O(d^2)$, becomes comparatively costly.

## What is Wrong About the Forward Construction?

We have provided three different constructions of Brownian paths with one standing apart in that it is clearly the most simple one. So why not use the forward construction for every application?

The answer is that theory predicts a big integration error for QMC if dimensions are big and the number of integration nodes is of realistic order, like a couple of millions only. But one can hope that if only a limited number of input parameters have significant importance for the result, then QMC might behave very similarly to a low-dimensional integration problem.

Figure 8.6 shows the influence of input parameters on the whole discrete path. We compare the forward construction on the left with the Brownian bridge construction on the right. In the two upper plots, all but the first input variables are held fixed. We see that the influence of the first variable on the overall behavior of the path (in an informal sense) is bigger for the Brownian bridge construction.

In the two lower plots all but the seventh input variables are held fixed. We see that in the forward construction only values of the path after the seventh node are influenced, but the overall influence is only slightly smaller than that of the first variable. In contrast, the influence of the seventh variable in the Brownian bridge construction is restricted to the third quarter and is much smaller than that of the first variable.

The above notion of "behaving like a low-dimensional problem" can be made precise using the notion of *effective dimension*. It must be added, though, that despite its popularity, the concept of effective dimension alone does not fully explain the success of the alternative constructions. An alternative approach is to use the concept of weighted spaces as presented in Chap. 4. One can argue that dimensions with higher index have a much smaller influence on the integrand than earlier dimensions, thus the norm of the integrand in a suitable weighted space (and thus the error of integration) is small. The problem is now to provide such a weighted space, together with suitable weights. The Korobov spaces described in Chap. 4 are spaces of periodic functions on the unit cube, while most problems in finance lead to non-periodic integrands that are either unbounded or defined on all of $\mathbb{R}^s$. We cannot provide such a space, and while there are many authors who investigated this problem, it remains largely unsolved at the present. So there is a gap in showing error bounds like those provided for weighted Korobov spaces for financial problems. Numerical experiments do however support the conjecture that similar bounds could be proven.

To answer the question posed in the header: there is nothing wrong with the forward construction, but for some classes of problems other constructions achieve lower errors, at least empirically. For other problems the forward construction may be just fine, presumably for the kind of problems for which the input variables already have different influence on the integrand. Yet some other problems may be such that they are intrinsically high-dimensional, i.e., any construction will yield an integrand that is ill-suited for QMC integration.

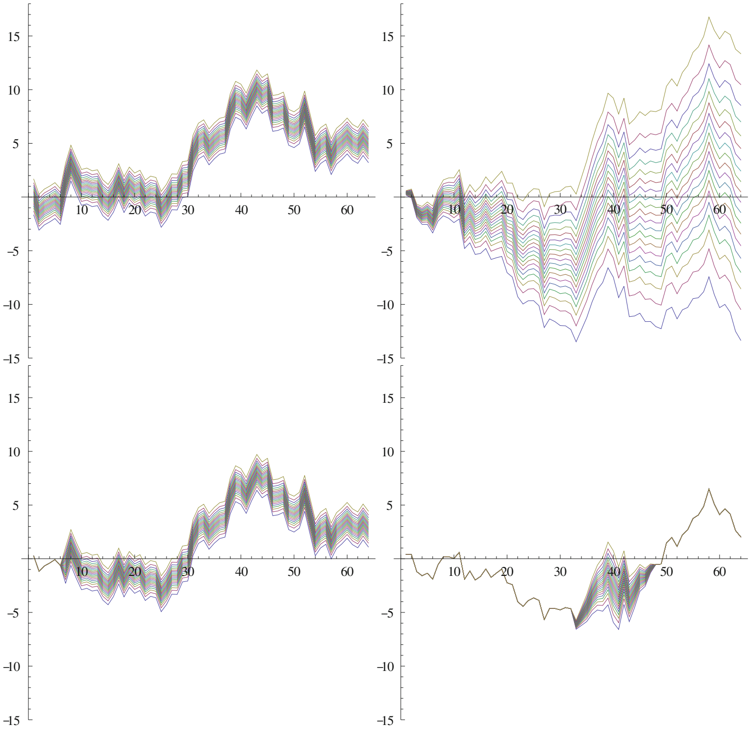

**Fig. 8.6** Paths of Brownian motion obtained by the forward construction (*left*) and the Brownian bridge construction (*right*). All but one parameters are fixed

## Evenly Spaced Discretization Nodes

The case where the $t_1, \dots, t_d$ are evenly spaced, i.e., where $t_j = \frac{jT}{d}$, $j = 1, \dots, d$, is of special interest. First, for many applications even spacing is natural. Second, as we will see below, we have a broader range of efficient constructions at our disposal for even spacing. There is no loss in assuming that $T = 1$.

In this setup the covariance matrix equals

$$\left(\frac{1}{d}\min(j,k)\right)_{j,k=1}^{d} = \frac{1}{d}\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2 & \dots & 2 \\ 1 & 2 & 3 & \dots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \dots & d \end{pmatrix}.$$

We will denote this matrix by $\Sigma^{(d)}$ or, if there is no danger of confusion, simply by $\Sigma$.

Note that we can compute the Cholesky decomposition of $\Sigma$ rather easily: $\Sigma^{(d)} = SS^\top$, where

$$S = S^{(d)} := \frac{1}{\sqrt{d}} \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix}.$$

Note that if $\boldsymbol{y} = (y_1, \dots, y_d)$ is a vector in $\mathbb{R}^d$, then $S\boldsymbol{y}^\top$ is the cumulative sum over $y$ divided by $\sqrt{d}$,

$$S\boldsymbol{y}^\top = \frac{1}{\sqrt{d}}(y_1, y_1 + y_2, \dots, y_1 + \dots + y_d)^\top .$$

We have the following two easy lemmas:

**Lemma 8.18.** *Let $A$ be any $d \times d$ matrix with $AA^\top = \Sigma$ and let $X$ be a standard normal vector. Then $W^\top = AX^\top$ is a discrete Brownian path with discretization* $\frac{1}{d}, \frac{2}{d}, \dots, \frac{d-1}{d}, 1$.

*Proof.* This is a special case of Lemma 8.17. □

**Lemma 8.19.** *Let $A$ be any $d \times d$ matrix with $AA^\top = \Sigma$. Then there is an orthogonal $d \times d$ matrix $V$ with $A = SV$. Conversely, $SV(SV)^\top = \Sigma$ for every orthogonal $d \times d$ matrix $V$.*

*Proof.* Suppose $AA^\top = \Sigma$, such that $AA^\top = SS^\top$. Note that $S$ is invertible and define $V = S^{-1}A$. Then

$$VV^\top = S^{-1}AA^\top(S^{-1})^\top = S^{-1}SS^\top(S^{-1})^\top = I_d ,$$

showing that $V$ is orthogonal. The converse follows from the fact that for orthogonal $V$ we have $V^\top = V^{-1}$. □

For evenly spaced discretization nodes the orthogonal matrices corresponding to the classical matrices can often be given explicitly. The orthogonal transform corresponding to the forward method is the identity mapping on $\mathbb{R}^d$. For $d = 2^k$, the orthogonal transform corresponding to the Brownian bridge construction where $W$ is computed in the order $W_1, W_{\frac{1}{2}}, W_{\frac{1}{4}}, W_{\frac{3}{4}}, W_{\frac{1}{8}}, W_{\frac{3}{8}}, W_{\frac{5}{8}}, \dots$, is given by the inverse Haar transform. For the PCA, the orthogonal transform can be computed using the fast discrete sine transform and thus the computation cost is $O(d \log d)$. The advantage of the representation of $A$ in Lemma 8.19 is that there are many orthogonal matrices that allow for fast matrix vector multiplication, that is, a path of length $d$ can be computed using $O(d \log d)$ operations. Examples include the Walsh transform, discrete sine/cosine transform, Hilbert transform and others.

In the case where the discrete Brownian path is used for solving an SDE, there is one more thing to be noted: only the increments of the Brownian path are needed. But for a discrete Brownian path $W$ on $\{\frac{1}{d}, \ldots, \frac{d}{d}\}$, computing the increments means multiplication by $S^{-1}$. Thus, if $W$ is generated using the orthogonal transformation $V$, i.e.,

$$(W_{\frac{1}{d}}, \ldots, W_{\frac{d}{d}})^\top = SV(X_1, \ldots, X_d)^\top ,$$

then

$$(W_{\frac{1}{d}}, W_{\frac{2}{d}} - W_{\frac{1}{d}} \ldots, W_{\frac{d}{d}} - W_{\frac{d-1}{d}})^\top = V(X_1, \ldots, X_d)^\top .$$

## Example 8.20

Consider the problem of valuating an Asian option, i.e., an option with payoff function $\max\left(\frac{1}{T-T_0}\int_{T_0}^{T} S_t \,\mathrm{d}t - K, 0\right)$, in the Heston model (cf. Example 7.5):

$$\begin{aligned}
\mathrm{d}B_t &= rB_t \,\mathrm{d}t,\\
\mathrm{d}S_t &= rS_t \,\mathrm{d}t + \sqrt{V_t}S_t(\rho \,\mathrm{d}W_t^{(1)} + \sqrt{1-\rho^2}\,\mathrm{d}W_t^{(2)}),\\
\mathrm{d}V_t &= \kappa(\theta - V_t)\,\mathrm{d}t + \xi\sqrt{V_t}\,\mathrm{d}W_t^{(1)},\\
(B_0, S_0, V_0) &= (b_0, s_0, v_0),
\end{aligned}$$

where $r, \rho, \kappa, \theta, \xi$ are positive constants and $W^{(1)}, W^{(2)}$ are independent standard Brownian motions. We solve the SDE using the simple Euler-Maruyama method given in (7.8), with $h = 2T/s$ ($s \in \mathbb{N}$ will be the dimension of the resulting integration problem, and we take $s$ to be even so we can use half of the coordinates for each Brownian path): $\hat{B}_0 = b_0, \hat{S}_0 = s_0, \hat{V}_0 = v_0$ and

$$\begin{aligned}
\hat{B}_{(k+1)} &= \hat{B}_k + r\hat{B}_k h = b_0(1+rh)^{k+1},\\
\hat{S}_{(k+1)} &= \hat{S}_k + r\hat{S}_k h + \sqrt{\hat{V}_k}\hat{S}_k\left(\rho\Delta W^{(1)}_{(k+1)} + \sqrt{1-\rho^2}\Delta W^{(2)}_{(k+1)}\right),\\
\hat{V}_{(k+1)} &= \hat{V}_k + \kappa(\theta - \hat{V}_k)h + \xi\sqrt{\hat{V}_{kh}}\Delta W^{(1)}_{(k+1)},
\end{aligned}$$

for $k = 1, \ldots, s/2$, where $\Delta W^{(j)}_{(k+1)} = W^{(j)}_{(k+1)h} - W^{(j)}_{kh}$, $j = 1, 2$. The payoff of the Asian option is approximated by replacing the integral by a corresponding sum, i.e.,

$$\max\left(\frac{1}{T-T_0}\int_{T_0}^{T} S_t \,\mathrm{d}t - K, 0\right) \approx \max\left(\frac{h}{T-T_0}\sum_{k=\lfloor\frac{T_0}{T}\frac{s}{2}\rfloor+1}^{s/2} \hat{S}_k - K, 0\right).$$

Next we need $\Delta W_k^{(j)}$ for $k = 1,\dots,\frac{s}{2}$, $j = 1,2$, the increments of the two Brownian motions. First we apply the inverse standard normal CDF component-wise to every point of an $s$-dimensional low-discrepancy sequence, giving us an $s$-dimensional sequence $(X_n)_{n\in\mathbb{N}_0}$. Now we can either partition each $X_n$ into two $\frac{s}{2}$-dimensional vectors and use each to generate the increments of one of the Brownian motions using some orthogonal transformation. For example, one could use the Brownian bridge construction to generate the first Brownian motion and the PCA to generate the secon one. Alternatively, and more generally, we can use one $s$-dimensional orthogonal transformation on $X_n$ and use the first $\frac{s}{2}$ components as the increments of the first Brownian motion and the remaining ones for the second one.

We test the different approaches numerically for a fixed set of parameters: let $s_0 = 100$, $v_0 = 0.3$, $r = 0.03$, $\rho = 0.2$, $\kappa = 2$, $\theta = 0.3$, $\xi = 0.5$, be the model parameters and let the option parameters be $K = 100$, $T_0 = 0$, $T = 1$. Note that the approximate payoff simplifies to

$$\max\left(\frac{2}{s}\sum_{k=1}^{s/2}\hat{S}_k - K, 0\right).$$

To perform the integration, we use the classical 64-dimensional Sobol' sequence with a random 64-dimensional shift added. We plot the $\log_2$ of the standard deviation over 64 integral evaluations each using $2^m$ points of the sequence, $m = 2,\dots,10$.

The left-hand graph of Fig. 8.7 shows the $\log_2$ of the standard deviation along $m$ for four different transforms: the identity, "Forward", the orthogonal transform corresponding to the Brownian bridge (i.e., the inverse Haar transform), "BB", the one corresponding to PCA and the Brownian bridge applied separately to the inputs of the two Brownian paths, "BB2". On the $x$-axis we plot the $\log_2$ of the number of integration points, i.e., $m$, while along the $y$-axis we plot the $\log_2$ of the standard deviation of the result over 64 runs.

We can see that, as in many practical examples, the PCA performs best. Maybe surprisingly, the idea of using two independent Brownian bridge constructions performs worse than the two combined transforms, but still much better than the identical transform.

We complement this graph with the corresponding one for a "ratchet option" , that is, an option with payoff

$$f(S_{\frac{T}{d}}, S_{\frac{2T}{d}},\dots,S_T) = \frac{1}{d}\sum_{j=1}^{d} 1_{[0,\infty)}\left(S_{\frac{jT}{d}} - S_{\frac{(j-1)T}{d}}\right) S_{\frac{jT}{d}}\,.$$

The errors are plotted on the right-hand side of Fig. 8.7. We can see that the orthogonal transforms that were so successful in the case of an Asian option now perform worse than the identity.

Thus the choice of a suitable orthogonal transform for generation of the Brownian paths depends on the payoff function. Exactly how this transform should be chosen, and for which types of payoffs it accelerates convergence, is still subject to research.

Coming back to the general case of unevenly spaced discretization nodes, we note the following: suppose you have nodes $0 < t_1 < \cdots < t_d$. We may compute an evenly spaced path $(W_{\frac{1}{d}}, W_{\frac{2}{d}},\dots,W_{\frac{d}{d}})$ using our favorite orthogonal transform, then compute

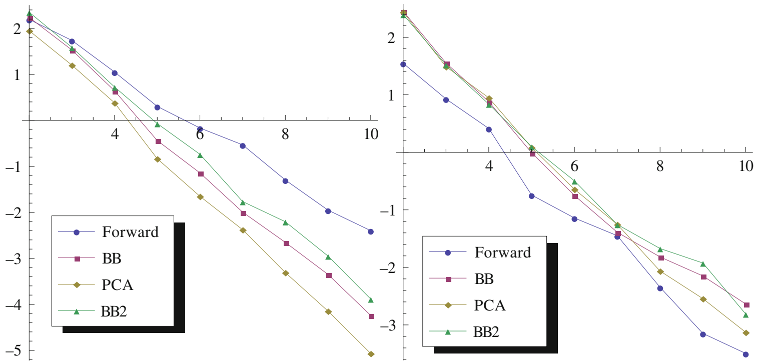

**Fig. 8.7** *Left*: Convergence of the price of an Asian option under different transforms. *Right*: same graph for the ratchet option

$$\tilde{W} = \sqrt{d}\Big(\sqrt{t_1}W_{\frac{1}{d}}, \sqrt{t_2 - t_1}(W_{\frac{2}{d}} - W_{\frac{1}{d}}), \ldots, \sqrt{t_d - t_{d-1}}(W_{\frac{d}{d}} - W_{\frac{d-1}{d}})\Big).$$

Then $\tilde{W}$ is a discrete Brownian path with discretization $0 < t_1 < \cdots < t_d$.

## 8.3 Multilevel (Quasi-)Monte Carlo

Suppose we want to approximate $\mathbb{E}[Y]$ for some random variable $Y$ which has finite expectation. Suppose further that we have a sequence of sufficiently regular functions $f^\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}$ such that

$$\lim_{\ell\to\infty} \mathbb{E}[f^\ell(X^\ell)] = \mathbb{E}[Y], \tag{8.1}$$

where for each $\ell \geq 0$, $X^\ell$ denotes a $d_\ell$-dimensional standard normal vector. In most cases the $f^\ell$ will be the discrete version of a function defined on the Brownian paths with $d_\ell$ discretization nodes, and typically $d_\ell = 2^\ell$. A standard example is provided by the fixed strike Asian option, which has payoff

$$f(W) := \max\left(\frac{1}{T}\int_0^T S_0 \exp\left(\sigma\sqrt{T}W_{t/T} + \left(r - \frac{\sigma^2}{2}\right)t\right) \mathrm{d}t - K, 0\right),$$

where $W$ is a standard Brownian motion, $S_0$ is the stock price at time 0, $K$ is the strike of the option, $\sigma$ is the volatility, and $r$ is the interest rate. $W$ will be approximated by a discrete path of the form $SV^\ell X^\ell$ where, for example, $V^\ell$ is the orthogonal transform corresponding to $d_\ell = 2^\ell$-dimensional PCA.

Equation (8.1) states that there exists a sequence of algorithms which approximate $\mathbb{E}[Y]$ with increasing accuracy. For example, if $f^\ell(X^\ell)$ has finite variance, we can approximate $\mathbb{E}[Y]$ by $\frac{1}{N}\sum_{k=0}^{N-1} f^{\ell}(X_k^{\ell})$ using sufficiently large $\ell$ and $N$, where $(X_k^{\ell})_{k\in\mathbb{N}_0}$ is a sequence of independent standard normal vectors.

Usually, evaluation of $f^{\ell}(X_k^{\ell})$ becomes more costly with increasing $\ell$. Multilevel methods sometimes help to save significant proportions of computation time by computing more samples for the coarser approximations, which need less computation time but have higher variance.

We have, for large $L$,

$$\begin{aligned}\mathbb{E}[Y] &\approx \mathbb{E}\left[f^{L}(X^{L})\right]\\ &= \mathbb{E}\left[f^{0}(X^{0})\right] + \sum_{\ell=1}^{L}\mathbb{E}\left[f^{\ell}(X^{\ell})\right] - \mathbb{E}\left[f^{\ell-1}(X^{\ell-1})\right]\\ &= \mathbb{E}\left[f^{0}(X^{0})\right] + \sum_{\ell=1}^{L}\mathbb{E}\left[f^{\ell}(X^{\ell})\right) - \mathbb{E}\left[f_c^{\ell-1}(X^{\ell})\right]\\ &= \mathbb{E}\left[f^{0}(X^{0})\right] + \sum_{\ell=1}^{L}\mathbb{E}\left[f^{\ell}(X^{\ell}) - f_c^{\ell-1}(X^{\ell})\right],\end{aligned} \tag{8.2}$$

where $(f_c^{\ell})_{\ell\in\mathbb{N}_0}$ is an arbitrary sequence of functions $f_c^{\ell} : \mathbb{R}^{d_{\ell+1}} \to \mathbb{R}$ with the property that $\mathbb{E}[f_c^{\ell-1}(X^{\ell})] = \mathbb{E}[f^{\ell-1}(X^{\ell-1})]$. The "$c$" in $f_c^{\ell}$ stands for "coarse level".

The most basic example for $f_c^{\ell}$ is given for $d_{\ell} = m^{\ell}$ by $f_c^{\ell} = f^{\ell} \circ C_{m,\ell}$, where $C_{m,\ell}$ is the linear map defined by the matrix

$$(C_{m,\ell})_{i,j} := \begin{cases} \frac{1}{\sqrt{m}} & \text{if } (i-1)m+1 \le j \le i\,m\,,\ 1 \le i \le m^{\ell},\\ 0 & \text{else.}\end{cases}$$

For example,

$$C_{2,\ell} := \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0\ldots & 0 & 0\\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0\ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & 0 & 0\ldots & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

In general, $f_c^{\ell}$ is chosen so as to get small variance for $f^{\ell}(X^{\ell}) - f_c^{\ell-1}(X^{\ell})$.

Equation (8.2) becomes useful if, as is often the case in practice, the expectation $\mathbb{E}\left[f^{\ell}(X^{\ell}) - f^{\ell-1}(C_{m,\ell}X^{\ell})\right]$ can be approximated to the required level of accuracy using less function evaluations $N_{\ell}$ for bigger $\ell$ while the costs $c_{\ell}$ per function evaluation increases. Suppose the error of approximation of $\mathbb{E}\left[f^{\ell}(X^{\ell}) - f_c^{\ell-1}(X^{\ell})\right]$ using $N_{\ell}$ points is $e_{\ell}(N_{\ell})$. We choose $N_0, \ldots, N_L$ so that

**Table 8.1** Multilevel (Q)MC using $2^{10}$ time steps ($L = 10$). The average and the standard deviation of the option price are based on 1000 runs. The average computing time is given in parentheses

| | Multilevel | | Multilevel QMC | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Monte Carlo | | Forward | | PCA | | Regression | |
| $N_L$ | Average | stddev | Average | stddev | Average | stddev | Average | stddev |
| 2 | 7.717 | $0.41 \times 10^{0}$ | 7.735 | $0.19 \times 10^{-1}$ | 7.736 | $0.16 \times 10^{-1}$ | 7.739 | $0.10 \times 10^{-1}$ |
| | (0.0057 s) | | (0.0057 s) | | (0.0088 s) | | (0.0069 s) | |
| 4 | 7.738 | $0.19 \times 10^{0}$ | 7.734 | $0.71 \times 10^{-2}$ | 7.736 | $0.44 \times 10^{-2}$ | 7.738 | $0.29 \times 10^{-2}$ |
| | (0.0074 s) | | (0.0074 s) | | (0.0118 s) | | (0.0091 s) | |
| 8 | 7.748 | $0.54 \times 10^{-1}$ | 7.737 | $0.30 \times 10^{-2}$ | 7.737 | $0.14 \times 10^{-2}$ | 7.736 | $0.10 \times 10^{-2}$ |
| | (0.0101 s) | | (0.0100 s) | | (0.0165 s) | | (0.0124 s) | |
| 16 | 7.746 | $0.40 \times 10^{-1}$ | 7.736 | $0.11 \times 10^{-2}$ | 7.737 | $0.69 \times 10^{-3}$ | 7.736 | $0.30 \times 10^{-3}$ |
| | (0.0157 s) | | (0.0157 s) | | (0.0279 s) | | (0.0194 s) | |
| 32 | 7.728 | $0.31 \times 10^{-1}$ | 7.736 | $0.49 \times 10^{-3}$ | 7.737 | $0.21 \times 10^{-3}$ | 7.736 | $0.10 \times 10^{-3}$ |
| | (0.0266 s) | | (0.0265 s) | | (0.0585 s) | | (0.0326 s) | |
| 64 | 7.739 | $0.81 \times 10^{-2}$ | 7.736 | $0.20 \times 10^{-3}$ | 7.737 | $0.69 \times 10^{-4}$ | 7.737 | $0.32 \times 10^{-4}$ |
| | (0.0486 s) | | (0.0484 s) | | (0.1202 s) | | (0.0583 s) | |

$$e_0(N_0) + \cdots + e_L(N_L) \leq \varepsilon$$

while minimizing the total cost

$$c = c_0 N_0 + \cdots + c_L N_L.$$

In this way the total computation cost is typically much lower than it would be if $\mathbb{E}[f^L(X^L)]$ would be computed directly.

We conclude with an example in which multilevel MC is combined with orthogonal transforms and QMC.

## Example 8.21

We compare multilevel MC with three multilevel QMC methods, forward, PCA, and the regression algorithm from [50] numerically. For that we choose the parameters in a Black-Scholes model as $r = 0.04$, $\sigma = 0.3$, $S_0 = 100$, and we aim to value an Asian call option with parameters $K = 100$ and $T = 1$. At the finest level we choose $2^{10}$ discretization points and at each coarser level the number of points is divided in by 2, i.e. $L = 10$ and $m = 2$. The number of sample points are doubled at each level starting with $N_L$ sample points at the finest level $L$. For the QMC approaches we take a Sobol' sequence with a random shift. In Table 8.1 we compare for different values $N_L$ both the average and the standard deviation of the price of the Asian call option based on 1000 independent runs. Moreover, the average computing time for one run is given in parentheses. As we can see, the QMC methods generally give better results in terms of accuracy in this simple example. In terms of computation times, the PCA suffers from its slightly more costly path construction and is not better than the forward method with double number of samples. The results for the regression method show that there is room for improvement beyond PCA.

## 8.4 Further Reading and Exercises

### Further Reading

The standard algorithm for sampling from the inverse normal CDF can be found in the article by Wichura [94], and the corresponding source code in FORTRAN can be found online. Translations into several other programming languages exist.

The standard reference for MC pricing of financial instruments is the book by Glasserman [31], which also contains some material on QMC simulation.

The topic of non-uniform random variate generation is treated in quite some detail in the book by Hörmann, Leydold, and Derflinger, [45]. There it is shown how the proposal PDF for a given target PDF can be computed automatically with very small rejection rate.

Dick and Zhu [21] study the combination of a special class of acceptance-rejection algorithms and QMC. They provide estimates for the discrepancy of the generated point set with respect to the target PDF and they find that it is of lower order than that of random points.

Example 8.14 is taken from the article [26] by Eichler, Leobacher, and Zellinger. There the problem of calibrating (financial) models is considered in some detail.

We noted in the text that the cost for generating Brownian paths using PCA is high due to the general matrix vector multiplication involved. Keiner and Waterhouse [52] describe an approximate PCA for which the cost of matrix-vector multiplication is $O(d \log d)$.

The notion of effective dimension was first formalized by Caflisch, Morokoff, and Owen in [12] using the concept of ANOVA decomposition. The latter has some interesting applications in complexity analysis, see for example [74].

In the article [63], Leobacher reviews a number of orthogonal transforms that can be computed using $O(d \log d)$ operations. For example, one can find these a detailed algorithm for the use of the fast discrete sine transform for fast computation of the PCA for evenly spaced nodes, which has originally been proposed by Scheicher, see [83].

The example for which the Brownian bridge and other transforms perform worse than the identical one, i.e., the ratchet option, was first considered in the context of QMC pricing by Papageorgiou in [80]. Papageorgiou concludes that the effectivity of a transform is intimately tied to the payoff structure. His results were formalized and generalized by Wang and Sloan [93]. In the articles [49, 50] by Leobacher and Irrgeher, it is tried to choose the orthogonal transform in a way that puts as much variance as possible into the dependence of the first input variable. To this end the payoff is approximated by a linear function $g$ ("regression") and an orthogonal (Householder-)transform $V$ is computed such that $g \circ V$ only depends on $X_1$. This $V$ is taken as the orthogonal transform for the original problem. See also Imai and Tan [47] for a different approach.

The idea of using orthogonal transforms on normal random vectors in order to obtain constructions for paths can to a certain degree be extended to Lévy paths. It was first used in [62] for the Brownian bridge construction and later, but independently, in [48] for general orthogonal transforms.

Multilevel MC is a technique for speeding up MC simulation, especially for SDE models. It has gained a lot of recognition over the last couple of years, starting with the pioneering work by Giles [30] and Heinrich [33].

One of the many topics from quantitative finance that has been omitted from this book, but which is very closely related to option pricing, is the computation of price sensitivities (the so-called "Greeks") with respect to model parameters. This concept and its significance is discussed in any textbook on finance, see, for example, Albrecher et al. [2]. We mention this topic because there are different ways for computing them using simulation, and usually the straightforward approach is not the best one. See Glasserman [31] for MC computation of sensitivities using the so-called likelihood ratio method. Fournier et al. provide an introduction to Malliavin calculus for computing sensitivities using MC [29].

## Exercises

8.1 Let $f : \mathbb{R} \to \mathbb{R}$ be a positive PDF, i.e.,

- $f$ is measurable;
- $f(x) > 0$ for all $x \in \mathbb{R}$;
- $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$.

Show that $F : \mathbb{R} \to (0,1)$, $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$ is bijective.

8.2 Suppose $f$ is a continuous positive PDF and $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$ for all $x \in \mathbb{R}$. Show that the inverse function $G$ of $F$ satisfies the ordinary differential equation $G'(u) = 1/f(G(u))$.

8.3 Let $F : \mathbb{R} \to [0,1]$ be a CDF, that is $F$ is monotone, right-continuous with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Define the *quasi-inverse* of $F$ by

$$F^{-}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\} \qquad \forall u \in (0,1).$$

Show that

(a) $F^{-}$ is non-decreasing;
(b) $F(F^{-}(u)) \geq u$ and $F^{-}(F(x)) \leq x$ for all $u \in (0,1)$, and all $x \in \mathbb{R}$;
(c) Conclude that for all $u \in (0,1)$, and all $x \in \mathbb{R}$, we have $F^{-}(u) \leq x$ if and only if $F(x) \geq u$;
(d) Conclude further that for any random variable $U$ which is uniformly distributed on $(0,1)$ we have that $F^{-}(U)$ is a real random variable with CDF $F$, i.e., $\mathbb{P}[F^{-}(U) \leq x] = F(x)$;
(e) Show that $F^{-}$ coincides with $F^{-1}$ if $F$ is invertible.

8.4 Another special case of great practical interest is where $X$ takes values in $\mathbb{Z}$. What does the quasi-inverse look like in that case? How can one write a computer program to sample from the distribution of $X$?

8.5 Write a simple computer program that generates a geometric random variable with parameter $q$ using a variant of the inversion method.

8.6 A doubly exponential random variable $X$ with parameters $\alpha, \beta$ has PDF

$$f(x) = \frac{1}{\frac{1}{\alpha} + \frac{1}{\beta}} \begin{cases} \exp(\alpha x) & \text{if } \alpha < 0, \\ \exp(-\beta x) & \text{if } \beta \geq 0. \end{cases}$$

(a) Show that $f$ is a PDF, i.e., $f$ is measurable, $f \geq 0$ almost everywhere, and $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$.
(b) Show that the CDF of $X$ is invertible.
(c) Compute the CDF of $X$.

8.7 Let $X$ be a real random variable with values in $I$ where $I$ is either $-(\infty, b]$, or $[a,b]$ or $[a,\infty)$. Suppose that the restriction of $F$ to $I$ is invertible. Show that the inversion method works analog to the case considered in the text.

8.8 The PDF of an exponential random variable $X$ is given by $f(x) = \lambda \exp(-\lambda x)$ for $x \geq 0$, $f(x) = 0$ for $x \leq 0$. Show that the CDF of $X$ is invertible on $[0,\infty)$ and compute its inverse.

8.9 The PDF of the arcsine distribution is given by $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$ for $x \in (0,1)$ and $f(x) = 0$ for $x \notin (0,1)$. Compute the corresponding CDF, show that it is invertible on $[0,1]$, and compute the inverse.

8.10 Show that the Marsaglia-Bray algorithm indeed generates a 2-dimensional standard vector.

8.11 Find a continuous PDF for sampling from a gamma distribution with parameter $a < 1$ using the acceptance-rejection method. *Hint:* first find a function $g_1 : (0,1] \to \mathbb{R}$ which is greater than the target PDF in its domain, then find a function $g_2 : [1,\infty) \to \mathbb{R}$ for the other part. $c^{-1}g$ equals $g_1$ on $(0,1]$ and $g_2$ on $(0,\infty)$, and $c$ has to be computed by integration.

8.12 Show how to construct a discrete Brownian path on the nodes $t_j = \frac{jT}{d}$ from a path on the nodes $\frac{1}{d}, \ldots, \frac{d}{d}$.

8.13 Prove Lemma 8.16.

# Bibliography

1. Aistleitner, C.: Covering numbers, dyadic chaining and discrepancy. J. Complex. **27**, 531–540 (2011)
2. Albrecher, H., Binder, A., Lautscham, V., Mayer, P.: Introduction to Quantitative Methods for Financial Markets. Birkhäuser, Basel (2013)
3. Andersen, L.: Simple and efficient simulation of the Heston stochastic volatility model. J. Comput. Finance **11**, 1–42 (2008)
4. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
5. Atanassov, E.I.: On the discrepancy of the Halton sequences. Math. Balkanica (N.S.) **18**, 15–32 (2004)
6. Bakhvalov, N.S.: Approximate computation of multiple integrals. Vestnik Moskov. Univ. Ser. Mat. Meh. Astr. Fiz. Him. **4**, 3–18 (1959). (Russian)
7. Beck, J., Chen, W.W.L.: Irregularities of Distribution. Cambridge University Press, Cambridge (1987)
8. Bilyk, D., Lacey, M.: The supremum norm of the discrepancy function: recent results and connections. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, pp. 23–28. Springer, Berlin/Heidelberg/New York (2013)
9. Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Pol. Econ. **81**, 637–654 (1973)
10. Brass, H., Petras, K.: Quadrature Theory: The Theory of Numerical Integration on a Compact Interval. Mathematical Surveys and Monographs, vol. 178. American Mathematical Society, Providence (2011)
11. Bykovskii, V.A.: The discrepancy of the Korobov lattice points. Izv. Math. **76**, 446–465 (2012)
12. Caflisch, R.E., Morokoff, W., Owen, A.: Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. J. Comput. Finance **1**, 27–46 (1997)
13. Chazelle, B.: The Discrepancy Method: Randomness and Complexity. Cambridge University Press, Cambridge (2000)
14. Cochrane, T.: Trigonometric approximation and uniform distribution modulo one. Proc. Am. Math. Soc. **103**, 695–702 (1988)
15. Cont, R., Tankov, P.: Financial Modelling with Jump Processes. Chapman & Hall/CRC, Boca Raton (2004)
16. Delbaen, F., Schachermayer, W.: The Mathematics of Arbitrage. Springer, Berlin (2006)
17. Dick, J., Niederreiter, H.: On the exact $t$-value of Niederreiter and Sobol' sequences. J. Complex. **24**, 572–581 (2008)
18. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
19. Dick, J., Pillichshammer, F.: Explicit constructions of point sets and sequences with low discrepancy. In: Kritzer, P., Niederreiter, H., Pillichshammer, F., Winterhof, A. (eds.) Uniform Distribution and Quasi-Monte Carlo Methods. Discrepancy, Integration and Applications, pp. 63–86. De Gruyter, Berlin (2014)

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6,

20. Dick, J., Pillichshammer, F.: Discrepancy theory and quasi-Monte Carlo integration. In: Chen, W.W.L., Srivastav, A., Travaglini, G. (eds.) Panorama of Discrepancy Theory, pp. 509–585. Springer (2014)
21. Dick, J., Zhu, H.: A discrepancy bound for a deterministic acceptance-rejection sampler. Electron. J. Statist. **8**, 678–707 (2014)
22. Dick, J., Kuo, F.Y., Sloan, I.H.: High dimensional integration – the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
23. Dick, J., Nuyens, D., Pillichshammer, F.: Lattice rules for nonperiodic smooth integrands. Numer. Math. **126**, 259–291 (2014)
24. Doerr, C., Gnewuch, M., Wahlström, M.: Calculation of discrepancy measures and applications. In: Chen, W.W.L., Srivastav, A., Travaglini, G. (eds.) Panorama of Discrepancy Theory, pp. 587–644. Springer (2014)
25. Drmota, M., Tichy, R.F.: Sequences, Discrepancies and Applications. Lecture Notes in Mathematics, vol. 1651. Springer, Berlin (1997)
26. Eichler A., Leobacher G., Zellinger, H.: Calibration of financial models using quasi-Monte Carlo. Monte-Carlo Methods Appl. **17**, 99–131 (2011)
27. Faure, H.: Discrépance de suites associées à un système de numération (en dimension $s$). Acta Arith. **41**, 337–351 (1982). (French)
28. Faure, H., Kritzer, P.: New star discrepancy bounds for $(t, m, s)$-nets and $(t, s)$-sequences. Monatsh. Math. **172**, 55–75 (2013)
29. Fournier, E., Lasry, J., Lebouchoux, J., Lions, P., Touzi, N.: Applications of Malliavin calculus to Monte Carlo methods in finance. Finance Stoch. **3**, 391–412 (1999)
30. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**, 607–617 (2008)
31. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
32. Harrison, J.M., Kreps, D.M.: Martingales and arbitrage in multiperiod securities markets. J. Econ. Theory **20**, 381–408 (1979)
33. Heinrich, S.: Multilevel Monte Carlo methods. Lect. Notes Comput. Sci. **2179**, 58–67 (2001)
34. Heinrich, S., Novak, E., Wasilkowski, G.W., Woźniakowski, H.: The inverse of the star discrepancy depends linearly on the dimension. Acta Arith. **96**, 279–302 (2001)
35. Hickernell, F.J.: Quadrature error bounds with applications to lattice rules. SIAM J. Numer. Anal. **33**, 1995–2016 (1996). Erratum: SIAM J. Numer. Anal. **34**, 853–866 (1997)
36. Hickernell, F.J.: Obtaining $O(n^{-2+\epsilon})$ convergence for lattice quadrature rules. In: Fang, K.T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 274–289. Springer, Berlin (2002)
37. Hickernell, F.J., Woźniakowski, H.: Tractability of multivariate integration for periodic functions. J. Complex. **17**, 660–682 (2001)
38. Hinrichs, A.: Covering numbers, Vapnik-Červonenkis classes and bounds for the star-discrepancy. J. Complex. **20**, 477–483 (2004)
39. Hinrichs, A., Novak, E., Ullrich, M., Woźniakowski, H.: The curse of dimensionality for numerical integration of smooth functions. Math. Comput. **83**, 2853–2863 (2014)
40. Hinrichs, A., Novak, E., Ullrich, M., Woźniakowski, H.: The curse of dimensionality for numerical integration of smooth functions II. J. Complex. **30**, 117–143 (2014)
41. Hinrichs, A., Novak, E., Woźniakowski, H.: The curse of dimensionality for monotone and convex functions of many variables. J. Approx. Theory **27**, 955–965 (2011)
42. Hlawka, E.: Funktionen von beschränkter Variation in der Theorie der Gleichverteilung. Ann. Mat. Pura Appl. **54**, 325–333 (1961). (German)
43. Hlawka, E.: Theorie der Gleichverteilung. Bibliographisches Institut, Mannheim (1979)
44. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
45. Hörmann, W., Leydold, J., Derflinger, G.: Automatic Nonuniform Random Variate Generation. Springer, Berlin (2004)
46. Hua, L.K., Wang, Y.: Applications of Number Theory to Numerical Analysis. Springer, Berlin/Heidelberg/New York (1981)

47. Imai, J., Tan, K.S.: A general dimension reduction technique for derivative pricing. J. Comput. Finance **10**, 129–155 (2007)
48. Imai, J., Tan, K.S.: An accelerating quasi-Monte Carlo method for option pricing under the generalized hyperbolic Lévy process. SIAM J. Sci. Comput. **31**, 2282–2302 (2009)
49. Irrgeher, C., Leobacher, G.: Fast orthogonal transforms for pricing derivatives with quasi-Monte Carlo. In: Laroque, C., Himmelspach, J., Pasupathy, R., Rose, O., Uhrmacher, A.M. (eds.) Proceedings of the 2012 Winter Simulation Conference, pp 385–398. Brussels (2012)
50. Irrgeher, C., Leobacher, G.: Fast orthogonal transforms for multilevel quasi-Monte Carlo simulation in computational finance. In: Vanmaele, W., et al. (eds.) Proceedings of Actuarial and Financial Mathematics Conference 2013, Interplay between Finance and Insurance, pp. 45–49. Berlin (2013)
51. Joe, S.: Component by component construction of rank-1 lattice rules having $O(n^{-1}(\ln(n))^d)$ star-discrepancy. In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 293–298. Springer, Berlin/Heidelberg/New York (2004)
52. Keiner, J., Waterhouse, B.J.: Fast principal components analysis method for finance problems with unequal time steps. In: L'Ecuyer, P., Owen, A.B. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2008, pp. 455–465. Springer, Berlin (2009)
53. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin/Heidelberg/New York (1992)
54. Knuth, D.E.: The Art of Computer Programming. Volume 2: Seminumerical Algorithms. Addison-Wesley, Reading/London/Don Mills (1969)
55. Korobov, N.M.: Number-theoretic methods in approximate analysis. Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow (1963). (Russian)
56. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Wiley, New York (1974)
57. Larcher, G.: On the distribution of sequences connected with good lattice points. Monatsh. Math. **101**, 135–150 (1986)
58. Larcher, G.: A best lower bound for good lattice points. Monatsh. Math. **104**, 45–51 (1987)
59. L'Ecuyer, P., Hellekalek, P.: Random number generators: selection criteria and testing. In: Hellekalek, P., Larcher, G. (eds.) Random and Quasi-Random Point Sets. Lecture Notes in Statistics, vol. 138. Springer, New York (1998)
60. L'Ecuyer, P., Munger, D.: Lattice builder: a general software tool for constructing rank-1 lattice rules (submitted for publication)
61. Lemieux, C.: Monte-Carlo and Quasi-Monte Carlo Sampling. Springer Series in Statistics. Springer, New York (2009)
62. Leobacher, G.: Stratified sampling and quasi-Monte Carlo simulation of Lévy processes. Monte-Carlo Methods Appl. **12**, 231–238 (2006)
63. Leobacher, G.: Fast orthogonal transforms and generation of Brownian paths. J. Complex. **28**, 278–302 (2012)
64. Lidl, R., Niederreiter, H.: Introduction to Finite Fields and Their Applications. Revision of the 1986, 1st edn. Cambridge University Press, Cambridge (1994)
65. Matoušek, J.: Geometric Discrepancy. Springer, Berlin (1999)
66. Merton, R.C.: Theory of rational option pricing. Bell J. Econ. Manage. Sci. **4**, 141–183 (1973)
67. Müller-Gronbach, T., Novak, E., Ritter, K.: Monte-Carlo Algorithmen. Springer, Berlin/Heidelberg (2012)
68. Niederreiter, H.: Point sets and sequences with small discrepancy. Monatsh. Math. **104**, 273–337 (1987)
69. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. Number 63 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia (1992)
70. Niederreiter, H., Xing, C.P.: Low-discrepancy sequences and global function fields with many rational places. Finite Fields Appl. **2**, 241–273 (1996)
71. Niederreiter, H., Xing, C.P.: Quasirandom points and global function fields. In: Finite Fields and Applications (Glasgow, 1995). London Mathematical Society Lecture Note Series, vol. 233, pp. 269–296. Cambridge University Press, Cambridge (1996)

72. Niederreiter, H., Xing, C.P.: Rational Points on Curves over Finite Fields: Theory and Applications. London Mathematical Society Lecture Note Series, vol. 285. Cambridge University Press, Cambridge (2001)
73. Novak, E.: Numerische Verfahren für Hochdimensionale Probleme und der Fluch der Dimension. Jahresber. Deutsch. Math. Verein. **101**, 151–177 (1999)
74. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume I: Linear Information. EMS, Zurich (2008)
75. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume II: Standard Information for Functionals. EMS, Zurich (2010)
76. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume III: Standard Information for Operators. EMS, Zurich (2012)
77. Nuyens, D.: Fast construction of good lattice rules. PhD thesis, Departement Computerwetenschappen, Katholieke Universiteit Leuven (2007)
78. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. Math. Comput. **75**, 903–920 (2006)
79. Nuyens, D., Cools, R.: Fast component-by-component construction, a reprise for different kernels. In: Niederreiter, H., Talay, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 373–387. Springer, Berlin (2006)
80. Papageorgiou, A.: The Brownian bridge does not offer a consistent advantage in quasi-Monte Carlo integration. J. Complex. **18**, 171–186 (2002)
81. Paskov, S.H., Traub, J.F.: Faster evaluation of financial derivatives. J. Portf. Manage. **22**, 113–120 (1995)
82. Rudin, W.: Real and Complex Analysis. McGraw-Hill, New York (1987)
83. Scheicher, K.: Complexity and effective dimension of discrete Lévy areas. J. Complex. **23**, 152–168 (2007)
84. Schmid, W.C.: Projections of digital nets and sequences. Math. Comput. Simul. **55**, 239–247 (2001)
85. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford University Press, New York/Oxford (1994)
86. Sloan, I.H., Reztsov, A.V.: Component-by-component construction of good lattice rules. Math. Comput. **71**, 263–273 (2002)
87. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals? J. Complex. **14**, 1–33 (1998)
88. Sloan, I.H., Woźniakowski, H.: Tractability of multivariate integration for weighted Korobov classes. J. Complex. **17**, 697–721 (2001)
89. Sobol', I.M.: Distribution of points in a cube and approximate evaluation of integrals. Ž. Vyčisl. Mat. i Mat. Fiz. **7**, 784–802 (1967)
90. Temlyakov, V.N.: Cubature formulas, discrepancy, and nonlinear approximation. J. Complex. **19**, 352–391 (2003)
91. Tezuka, S.: Uniform Random Numbers: Theory and Practice. The Kluwer International Series in Engineering and Computer Science, vol. 315. Kluwer Academic, Dordrecht (1995)
92. Traub, J.F., Werschulz, A.G.: Complexity and Information. Cambridge University Press, Cambridge (1998)
93. Wang, X., Sloan, I.H.: Quasi-Monte Carlo methods in financial engineering: an equivalence principle and dimension reduction. Oper. Res. **59**, 80–95 (2011)
94. Wichura, M.J.: Algorithm AS 241: the percentage points of the normal distribution. J. R. Stat. Soc. Ser. C (Appl. Stat.) **37**, 477–484 (1988)
95. Weyl, H.: Über die Gleichverteilung mod. Eins. Math. Ann. **77**, 313–352 (1916). (German)

# Index

G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-03425-6,