Random Numbers for Computer Simulation

Diplomarbeit
zur Erlangung des Magistergrades
an der Naturwissenschaftlichen Fakultät
der Universität Salzburg

eingereicht von
Hannes Leeb

Salzburg, im Jänner 1995

The core of this thesis originated from a series of discussions with my supervisor Peter Hellekalek and with Karl Entacher, Ferdinand Österreicher, and Maximilian Thaler.

I owe them.

Dr. Who: Hence this new device.

Romana: What is it?

Dr. Who: Ah, well, it's called a randomizer and it's fitted to the guidance system. It operates under a very complex scientific principle called potluck.

– Dr. Who, The Armageddon Factor

Contents

Chapter 1

Original sin

Anyone who considers arithmetical methods
of producing random digits
is, of course, in a state of sin.
– John von Neumann, 1951

(1.0.1) We will get involved into right this: computer algorithms for producing random numbers. Consider the C function declarations below.

```
double sin( double x );
int    rand();
```

The `sin()` function is well known, and implementations of `rand()` are what we are looking for. This might be not much of a problem – both `sin()` and `rand()` should approximate well-defined mathematical objects, the Sinus function and a sequence of independent random variables uniformly distributed on $\{0, \ldots, \mathrm{RAND\_MAX} - 1\}$[1]. The function `sin()` is easily compared to the mathematical object it should approximate; an implementation which returns `3.14` for `sin(0.0)` will certainly not be considered a well suited one. But what about an implementation of `rand()` which returns `12345, 1406932606, 654583775, 1449466924` on four successive calls?

(1.0.2) Ripley describes the common attitude towards this problem in [93, p.2] as follows: "The first thing needed for a stochastic simulation is

[1] RAND_MAX is a predefined integer depending on which system you are using.

a source of randomness. This is often taken for granted but is of fundamental importance." In [93, p.14], he continues: "Many users of simulation are content to remain ignorant of how such numbers were produced, merely calling standard functions to produce them. Such attitudes are dangerous, for random numbers are the foundations of our simulation edifice, and problems at higher levels are frequently traced back to faulty foundations."

(1.0.3) Ever since my first acquaintance with the subject of random number generation, I failed to cope with the apparent paradox of employing *deterministic* algorithms to produce *random* numbers. By the time, uneasyness turned from suspicion to the assertion that the whole concept per se did not make much sense. Finally, I was able to prove:

> In general, it is not possible to rate any finite sequence of numbers more 'random' than any other.

With this, an implementation of `rand()` which returns `0` all over again cannot be considered worse than your favorite 'random' number generator. But somehow intuition[2] as well as a whole world full of applications seem to oppose this attitude[3].

(1.0.4) It is common belief today that computers are capable of simulating almost everything – even quantities that are indeterminate and random. The crucial point in simulating randomness with a computer is that the latter is what the former is definitly not: *deterministic*. But, somehow, this concept obviously passed the test of application in practice. The various fields, ranging from physics to operations research and civil engineering, in which stochastic simulation is employed whitness the attraction of this method; almost every general-purpose programming language and even off-the-shelf spreadsheet programs like Microsoft Excel include a 'random' number generator.
Here, we cannot treat the numerous techniques of stochastic simulation (among the many books on this subject, we refer the reader to [8], [48], [57], or [93]); just let us observe that a method which proves to work well in so many applications can hardly be completely unfounded.

[2] You are likely to prefer your favorite algorithm for `rand()` to a 0-repeating `rand()`.

[3] There are exceptions like Zaremba [116] who considers the concept of simulating randomness on a computer as "spurious". But, somehow, this idea failed to become popular.

(1.0.5) The problem we face here is this: users expect mathematicians to propose deterministic algorithms – programs – 'well suited' for generating 'random' numbers. They have reason to do so, since mathematicians started using such algorithms in the first place.

To be precise: the whole story started in the 1940s, and von Neumann, Metropolis, Ulam, and Lehmer may be named among the pioneers in the field.
John von Neumann apparently conjectured the potential[4] of computers for stochastic simulation in 1945 when he wrote [84]: "It [the computer] will certainly open up a new approach to mathematical statistics; the approach by computed experiments ..." During the '40s, computer based stochastic simulation remained restricted to secret projects of the U.S. Department of Defense. The publication of 'The Monte Carlo method' by N. Metropolis and S.M. Ulam [81] in 1949 denotes the beginning of the 'official' history of this method (according to [87, p.11]). Two years later, D.H. Lehmer [73] proposed the linear congruential generator which – with a slight generalization by Thomson [110] and, independently, Rotenberg [97] – was to become today's most widely used method for random number generation.

(1.0.6) Before we actually propose some algorithms for generating 'random' numbers, let us state the problem a little more precise. The mathematical object we want to model is a sequence $(X_n)_{n\geq 0}$ of random variables which are stochastically independent and equidistributed on $[0,1[$ (or, in some examples, on $\{0,1\}$). We are looking for algorithms to produce a sequence of numbers $(x_n)_{n\geq 0}$, such that the x_n can be taken as realizations of the X_n. Since every stochastic simulation on a computer is expected to end in finite time, it can only consume a finite amount of 'random' numbers; we can therefore restrict our attention to the *finite* sequences or vectors $\mathbf{X} = (X_n)_{n=0}^{N-1}$ and $\mathbf{x} = (x_n)_{n=0}^{N-1}$.

Most stochastic models are based on random variables with distributions different from the uniform distribution on $[0,1[$; however, the latter is used as a 'point of common reference' from which the desired distribution is obtained by various transformation methods (which are, among others, discussed by Devroye in [23] or Ripley in [93, Chapter 3, 4]).

[4] It is interesting to note that the advent of computer based stochastic simulation happened to coincide with the advent of the first really bad 'random' number generator: the middle-square method has many undesirable properties which are discussed in detail by Knuth in [59, pp.3–8].

(1.0.7) In Chapter 2, we first reflect on how to assign some degree of randomness to a finite sequence of random numbers, restricting our considerations to numbers and random variables in $\{0,1\}$: we compare the numbers $\mathbf{x} = (x_n)_{n=0}^{N-1}$ to the mathematical object they should model, the random variables $\mathbf{X} = (X_n)_{n=0}^{N-1}$. In this way, we find that, *in general*, any sequence $\mathbf{x}$ is as 'good' as any other sequence $\mathbf{y}$; any random number generator is as 'good' as any other, no matter which sequence it produces. Therefore, it is impossible to find a generator which is 'good' in general.
Fortunately, the term '*in general*' is essential for the above statement to hold. If some information about the simulation problem or a class of simulation problems is available, we will be able to assign some degree of quality to generators, and different generators will differ in quality. Our intuitive preference for some generators to others[5] simply presupposes certain assumptions about our simulation problem. We develop a mathematical device which forces us to make these assumptions explicit and which enables us to make reasonable assessments of a random number generator's quality.

In Chapter 3, we generalize our observations on random numbers in $\{0,1\}$ to the more interesting case of random numbers in $[0,1[$.

Applications of our approach for assessing a generator's quality are given in Chapter 4. The applications mostly use existing mechanisms like statistical tests, but provide a deeper understanding of their relevance.

Finally, we present and study some random number generators in Chapter 5.

The two appendices contain some arguments and derivations needed by or being related to the main text. Both are entirely self-contained and can be understood without reading any of the other chapters. However, within the main text, we point out when an appendix might or should be read.

[5] Think of simulating games of roulette with a sequence of integers in $\{0,\ldots,36\}$: you will almost certainly rate a sequence which repeats 'Zero' all over again as 'worse' than any sequence without such obvious regularity.

Chapter 2

Running for randomness – heat 1

I would like, ah, if I may,
...to take you on a strange journey.
– Narrator, The Rocky Horror Picture Show

2.1 The notion of a sequence of random numbers

(2.1.1) Before we start searching for random number generators that produce a 'good' random number sequence $\mathbf{x} = (x_n)_{n=0}^{N-1}$, we give a precise definition of this notion. Everybody seems to have a clear idea of what random numbers are, but when asked to give an explicit definition, things get complicated. Knuth [59, p.2] observes: "People who think about this topic almost invariably get into philosophical discussions about what the word 'random' means. In a sense, there is no such thing as a random number; for example, is 2 a random number?"

(2.1.2) The mathematical object we are about to model, the vector of N random *variables* $\mathbf{X} = (X_n)_{n=0}^{N-1}$ is uniquely defined by the axioms of Kolmogorov [61]. But how to define a sequence $\mathbf{x}$ of random *numbers*? Knuth describes this problem in [59, p.142] as this: "The mathematical theory of probability and statistics carefully sidesteps the question; it refrains from making absolute statements, and instead expresses everything in terms of how much *probability* is to be attached to statements involving random sequences of events. The axioms of probability theory are set up so that abstract probabilities can be computed readily, but nothing is said about what probability really signifies, or how this concept can be applied meaningfully to the actual world." Similar opinions are expressed by Ripley in [93, p.14] and Schnorr in [103, p.8].

There are of course attempts to give the notion of a sequence of random numbers a precise meaning. We refer the interested reader to Kac [**?**] or Lagarias [65] for surveys and to Knuth [59, Section 3.1, 3.5] or Schnorr [103] for more thorough treatises. However, as Anderson notes in [3], be "warned that you will find many mathematical esoterica that, at present, have very little to do with the actual generation of random numbers on today's computers." In our opinion, the most interesting approach is given by Compagner in [12]; in particular, this approach complies with what we are going to find out in this and the following chapter. However, none of the various concepts succeeded in becoming a widely accepted standard so far.

The lack of a standard definition of a sequence of random numbers has led to some rather peculiar arguments for the quality of random number generators: in [41, p.177], Frederickson et. al. claim about their generators (which they call pseudo-random trees) that there is "no good reason to believe that any other family of pseudo-random trees offers any advantages."

(2.1.3) For our purposes, we will get along with a very simple definition of a random number sequence. Its simplicity is admissible because the sequences we are concerned with are computer-generated and therefore not random anyway. Moreover, it is admissible since we will show that every $\mathbf{x} \in [0,1[^N$ is as good in approximating the random vector $\mathbf{X}$ as any other[1] $\mathbf{y} \in [0,1[^N$.
Our definition is derived from Hammersley and Handscomb [48, p.25]: "The essential feature common to all Monte Carlo computations is that at some point we have to substitute for a random variable a corresponding set of actual values, having the statistical properties of the random variable. The values that we substitute are called *random* numbers [...]." We will treat the problem of the 'quality' of a random number sequence separately, so we can define it without demanding any statistical properties.

[1] This seems to be quite obvious for $N = 1$ as indicated by Knuth above; but for large N, say, $N = 10^6$, intuition plays a trick on us.

Definition 2.1 *A sequence* $\mathbf{x} = (x_n)_{n=0}^{N-1} \in [0,1[^N$ *is called a* finite sequence of random numbers, *if – while performing a stochastic simulation –* $\mathbf{x}$ *is substituted for a sequence* $\mathbf{X} = (X_n)_{n=0}^{N-1}$ *of independent and on* $[0,1[$ *equidistributed random variables.*

For convenience, we will drop the adjective 'finite' and refer to $\mathbf{x}$ as a sequence of random numbers, knowing it is finite anyway. Instead of numbers in $[0,1[$, numbers in $\{0,1\}$ or vectors in $[0,1[^s$ are required sometimes. We define a finite sequence of random numbers in $\{0,1\}$ or of random vectors in analogy to above (i.e., say, as N vectors in $[0,1[^s$ which – while performing a stochastic simulation – will be substituted for N independent random quantities which are equidistributed on $[0,1[^s$).

In the sense of this definition, a sequence of numbers is called 'random' simply because it is being used as substitute for a sequence of random variables. One may expect that this definition allows a sequence of random numbers to vary significantly in quality: intuitively, one would assign much less randomness to the sequence $(x_n = 0)_{n=0}^{N-1}$ than to most of the other conceivable sequences of length N. However, besides intuition, we have no means of assessing this quality so far. This is a major problem to be solved, because, as Niederreiter [87, p.2] notes: "The success of a Monte Carlo calculation often stands or falls with the 'quality' of the random samples that are used, where, by 'quality', we mean how well the random samples reflect true randomness." From what we have seen before, it is clear that a precise notion of quality is not easy to find. The above quotes from Hammersley and Handscomb as well as Niederreiter suggest that certain statistical properties will play a major role in measuring a random sequence's quality.

2.2 Scoring sequences of random numbers with statistical tests

(2.2.1) When assessing the quality of a sequence $\mathbf{x} = (x_n)_{n=0}^{N-1}$ of random numbers, we try to find out how closely it matches the mathematical object it should model, the sequence $\mathbf{X} = (X_n)_{n=0}^{N-1}$ of random variables. In this vein, we have to compare two things which defy direct comparison: random variables which are indeterminate by definition and pre-determined numbers. All we can do is check whether or not $\mathbf{x}$ gives a good representation of certain statistical properties of $\mathbf{X}$ or – putting it the other way round – whether the stochastic object $\mathbf{X}$ gives a good description of the behavior of the numbers $\mathbf{x}$. Stated formally, given $\mathbf{x}$, we have to choose one of the two alternatives below.

H_0 : The numbers x_n can be considered as realizations of the random variables X_n.

H_1 : The above statement does not hold.

This is the traditional setup for a statistical test.

Throughout the literature, random numbers are assumed to have statistical properties of random variables and to pass statistical tests. Lehmer, for example, defines a random sequence in [73] as "a vague notion embodying the idea of a sequence [...] whose digits pass a certain number of tests, traditional with statisticans and depending somewhat on the uses to which the sequence is to be put." Or with the words of Knuth [59, p.38]: "The theory of statistics provides us with some quantitative measures for randomness. [...] If a sequence behaves randomly with respect to tests $T_1, T_2, \ldots, T_n$, we cannot be *sure* in general that it will not be a miserable failure when it is subjected to a further test T_{n+1}; yet each test gives us more and more confidence in the randomness of the sequence." For more examples of this approach, see Hammersley and Handscomb [48, p.25] or Ripley [93, p.15]. We will see later in (2.3.7) and Section 4.1 if the confidence suggested by Knuth has a firmer foundation than bare intuition.

(2.2.2) Given two sequences $\mathbf{x}, \mathbf{y}$ of random numbers and a set of tests $\mathcal{F}$, we can identify the set $\mathcal{F}_{\mathbf{x}}$ of tests in $\mathcal{F}$ passed by $\mathbf{x}$ and the analogous set $\mathcal{F}_{\mathbf{y}}$ for $\mathbf{y}$. There are several possible ways to compare $\mathbf{x}$ and $\mathbf{y}$; we consider

Criterion 2.1 $\mathbf{x}$ *is at least as good as* $\mathbf{y}$ *if* $\mathbf{x}$ *passes at least as many tests as* $\mathbf{y}$*:*

$$\#\mathcal{F}_{\mathbf{y}} \leq \#\mathcal{F}_{\mathbf{x}}.$$

In addition, we try[2]

Criterion 2.2 $\mathbf{x}$ *is at least as good as* $\mathbf{y}$ *if* $\mathbf{x}$ *passes at least the tests passed by* $\mathbf{y}$*:*

$$\mathcal{F}_{\mathbf{y}} \subseteq \mathcal{F}_{\mathbf{x}}.$$

[2] We would like to thank Otmar Lendl for suggesting this one.

(2.2.3) To apply these criteria to specific $\mathbf{x}$ and $\mathbf{y}$, we need a tighter grip on the sets $\mathcal{F}$, $\mathcal{F}_{\mathbf{x}}$, and $\mathcal{F}_{\mathbf{y}}$ and therefore on the decision mechansim used by statistical tests in general. Let us study the basic function of statistical tests on a simple problem: given the results of N successive coin tosses, we are asked to decide if the coin is fair. Setting $x_n = 1$ for 'heads' on the n-th toss and $x_n = 0$ otherwise $(n = 0, \ldots, N-1)$, we can formalize the problem as follows.

Given the *sample* $\mathbf{x} = (x_0, \ldots, x_{N-1})$, we have to decide between the following two *hypotheses.*

H_0 : The x_n are realizations of random variables X_n which are independent and equidistributed on $\{0, 1\}$.

H_1 : The above statement does not hold.

Unfortunately, H_0 is so vague that it renders every 0-1-sequence of length N possible, each with probability $1/2^N$. Whatever values the x_n may have, each design is compatible with H_0. Whenever one decides in favor of H_1, misjudgement is therefore always possible. If the *possibility* of a wrong decision cannot be removed, one tries to decide so that at least the *probability* of a wrong decision is small. Thus: if H_0 holds, we want the event $\rightsquigarrow H_1$ of choosing H_1 to be quite improbable. Denoting 'quite improbable' by some probability α with $0 < \alpha \ll 1$, say, $\alpha = 0.05$, we want

$$P(\rightsquigarrow H_1 | H_0) = \alpha.$$

We can picture a statistical test as a decision rule T which, for every $\mathbf{x} \in \{0, 1\}^N$, either yields $T(\mathbf{x}) = 0$ if one decides in favor of H_0 or $T(\mathbf{x}) = 1$ if H_1 is favored. A decision for H_1 while H_0 holds is quite improbable if

$$P(T(\mathbf{X}) = 1) = \alpha.$$

Such a T is called *statistical test with level of significance* α; sometimes, we will refer to it as α–test, for short.

Deciding whether a coin is fair, one usually focuses on the number $S(\mathbf{x})$ of 'heads' which should be close to its expected value $N/2$. If the number of 'heads' deviates heavily from this, say, by more than some constant c, there might be reason to favor H_1 and regard the coin as unfair. It remains to choose c such that

$$P\left(\left|S(\mathbf{X}) - \frac{N}{2}\right| > c\right) = \alpha.$$

With this, the statistical test T with level of significance α is complete: we set $T(\mathbf{x}) = 0$ if $|S(\mathbf{x}) - N/2| \leq c$, $T(\mathbf{x}) = 1$ otherwise, and get $P(T(\mathbf{X}) = 1) = \alpha$.

(2.2.4) Observe that this framework of identifying an α–test with an event T with $P(T = 1) = \alpha$ is quite universal. As L'Ecuyer notes in [69, p.94], "any function of a finite set of i.i.d. uniform random variables can be used as a statistic to define a test of hypothesis, if its distribution is known." (The same argument is put forward by Marsaglia in [80, p.6]). This holds in particular for the events T we considered: every event T with probability α defines an α–test. On the other hand, every statistical test can be expressed as an event T with probability α. We can conclude that the set of all possible α–tests equals the set of events T with $P(T = 1) = \alpha$.

Picturing statistical tests as events is not very useful for actually performing a test. The reader should keep in mind that we are after sets of statistical tests, specifically $\mathcal{F}$, $\mathcal{F}_{\mathbf{x}}$, and $\mathcal{F}_{\mathbf{y}}$ which are needed to apply Criterion 2.1 and 2.2.

(2.2.5) Let us now consider a similar problem which is exactly what we are looking for: given the results $x_0, \ldots, x_{N-1}$ of N successive coin tosses simulated by a computer program, we are asked to decide if the x_n are sufficiently random, i.e. if they behave like realizations of random variables X_n which are statistically independent and equidistributed on $\{0, 1\}$. Again, we have to choose one out of two hypotheses [3].

H_0 : The x_n are realizations of random variables X_n which are independent and equidistributed on $\{0, 1\}$.

H_1 : The above statement does not hold.

And again, we search for a statistical test T with level of significance α. In the example above, we had $T(\mathbf{x}) = 0$ if $|S(\mathbf{x}) - N/2| \leq c$ and $T(\mathbf{x}) = 1$ otherwise, which was some sort of 'natural' choice; nobody being interested in the fairness of a coin would reasonably check if the nonoverlapping pairs $(x_0, x_1), (x_2, x_3), \ldots$ are equidistributed on $\{0, 1\} \times \{0, 1\}$ or if the number of 'runs' in the sequence[4] is close to its expected value $(N + 1)/2$.

[3] Note this problem is formally equivalent to the one above.

[4] The number of runs in the binary sequence $(x_n)_{n=0}^{N-1}$ is the number of maximum length blocks $x_n, \ldots, x_{n+k}$ in the sequence which consist entirely of 1s or entirely of 0s.

In the present context however, tests based on these aspects are perfectly reasonable[5]! We have no 'natural' choice for a statistical test T; for a user who takes the x_n as input for his simulation, *any* statistical aspect could be of fundamental importance. Having no information about the statistical aspects the user considers relevant, we have to treat all statistical tests as equally important. Concerning the two criteria, this means the set $\mathcal{F}$ should include *all* statistical tests with level of significance equal to[6] α.

If we consider all α–tests as equally important, we may take interest in the total number $\#\mathcal{F}$ of these tests. A statistical test with level of significance α is an event T on $\{0,1\}^N$ with $P(T=1)=\alpha$. Any event T on $\{0,1\}^N$ is, in turn, uniquely defined by the set $\mathbf{A}$ of those elements $\mathbf{x}$ for which $T(\mathbf{x})=1$; given $\mathbf{A}$, the event can be written as $T=1_{\mathbf{A}}$. For an event $1_{\mathbf{A}}$, we have $P(1_{\mathbf{A}}(\mathbf{X})=1)=\#\mathbf{A}/2^N$ (Note that demanding $P(1_{\mathbf{A}}(\mathbf{X})=1)=\alpha$ implicitly requires that α is rational and $\alpha 2^N$ is an integer). Hence there are as many α–tests as there are subsets of $\{0,1\}^N$ with $\alpha 2^N$ elements. The set $\mathcal{F}$ of all α–tests is

$$\mathcal{F}=\left\{1_{\mathbf{A}}:\ \mathbf{A}\text{ is a subset of }\{0,1\}^N\text{ with }\#\mathbf{A}=\alpha 2^N\right\},$$

and their number is

$$\#\mathcal{F}=\left(\begin{array}{c}2^N\\ \alpha 2^N\end{array}\right).$$

Now that we know $\mathcal{F}$, we can focus on the set $\mathcal{F}_{\mathbf{x}}$ of tests from $\mathcal{F}$ passed by $\mathbf{x}$. Since the sample $\mathbf{x}$ passes an α–test $T=1_{\mathbf{A}}$ if and only if $\mathbf{x}\notin\mathbf{A}$, we get

$$\mathcal{F}_{\mathbf{x}}=\{1_{\mathbf{A}}:\ 1_{\mathbf{A}}\in\mathcal{F},\,\mathbf{x}\notin\mathbf{A}\},$$

and elementary combinatorics yields

$$\#\mathcal{F}_{\mathbf{x}}=\left(\begin{array}{c}2^N-1\\ \alpha 2^N\end{array}\right).$$

Doesn't something strike you mind? ... $\#\mathcal{F}_{\mathbf{x}}$ is *independent* of $\mathbf{x}$! For any two samples $\mathbf{x}$ and $\mathbf{y}\in\{0,1\}^N$, we have

$$\#\mathcal{F}_{\mathbf{x}}=\#\mathcal{F}_{\mathbf{y}}.$$

[5] Among others, aspects like these are checked by Knuth's tests on random number generators in [59, Section 3.3].

[6] We could as well have allowed $\mathcal{F}$ to contain all tests with level of significance $\leq\alpha$ or with level of significance between two bounds α' and α''; the generalization is so obvious that it is left to the reader.

In words:

> Each binary sequence of length N passes exactly the same number of statistical tests.

What about the criteria we wanted to use for rating sequences of random numbers? Criterion 2.1 is obviously useless since it rates any $\mathbf{x}$ as 'good' as any other $\mathbf{y}$. Criterion 2.2 is useless as well: assume we rate $\mathbf{x}$ as actually better than $\mathbf{y}$ on the basis of Criterion 2.2. Then

$$\mathcal{F}_{\mathbf{y}} \subset \mathcal{F}_{\mathbf{x}}$$

in the sense that $\mathcal{F}_{\mathbf{y}}$ is a proper subset of $\mathcal{F}_{\mathbf{x}}$. This would imply

$$\#\mathcal{F}_{\mathbf{y}} < \#\mathcal{F}_{\mathbf{x}},$$

which – as we have seen above – cannot be.

One might argue that this phenomenon stems from our sampling from the simple set $\{0, 1\}$, but in vain. Sampling from $\{0, 1, 2, \ldots, M - 1\}$, we get the same results as above with just 2^N replaced by M^N in the formulas. Moreover, the same problem occurs if we sample from the continuum $[0, 1[$ instead of a finite set. The reader might believe this by taking the discrete case as an intuitive basis or see Chapter 3 for a formal proof.

(2.2.6) This is where we end up:

> There is, *in general*, no reason to regard any $\mathbf{x}$ as more random than any other. There is, *in general*, no reason to attribute randomness to any such $\mathbf{x}$.

The reader will now sense the reason in our rather unusual Definition 2.1 of a finite sequence of random numbers.

This problem is well known, of course; although seldom stated explicitly, it surfaces all through the literature. Niederreiter notes in [87, p.164]: "Early in the history of the Monte Carlo method, it already became clear that 'truly random' numbers are fictious from a practical point of view. Therefore users have resorted to pseudorandom numbers (abbreviated PRN) that can be readily generated in the computer by deterministic algorithms

with relatively few input parameters. [...] It should be clear that a deterministic sequence of numbers cannot perform well under *all* imaginable tests for randomness." These observations, although sobering, are particularly important for the computer–generated sequences we are searching for: concerning the 'quality' of a given $\mathbf{x}$, it is unimportant how the numbers were obtained. No matter how elaborate the physical source of randomness, no matter how sophisticated the random number generator, the produced sequence is as random as any other. Once the sample is *known*, all the randomness seems to evaporate in a puff of logic[7].
Knuth [59, p.2] observes: "In a sense, there is no such thing as a random number;" he continues [59, p.145]: "In a similar vein, one may argue that there is no way to judge whether a *finite* sequence is random or not; any particular sequence is just as likely as any other one."

Apart from the fact it is already being done rather successfully, we have reason to worry if substituting finite sequences of random numbers for random variables in stochastic simulations makes any sense at all: "the quality of a generator can never be *proven* by any statistical test.", as L'Ecuyer notes in [69, p.94], and we have seen this holds in a very general sense. Knuth [59, p.161] parries with intuition: "Still, nearly everyone would agree that the sequence 011101001 is 'more random' than 101010101, and even the latter sequence is 'more random' than 000000000." In fact, Knuth's parry is based on more than just a 'feeling': if you take the three sequences above to simulate a fair coin, the third sequence is quite likely to produce the worst result.

(2.2.7) All the trouble started when we regarded all α–tests as equally important; the reason for this was that we had *no information* about which statistical aspect might be relevant for the user in his simulation.

> The set $\mathcal{F}$ of *all* α–tests is simply too large to render a given sample superior in quality to any of the others.

If intuition and real-world experience teach us to prefer certain sequences $\mathbf{x}$ for certain simulation problems, we apparently base our preference on some *restricted* test set $\mathcal{F}$ or on some nonuniform *weighting of relevance* on

[7] It is discomfortingly easy to see that even the strong law of large numbers seizes to hold once they are determined, i.e. once the random variables are replaced by actual values.

this set. Considering some tests as very important and others as virtually irrelevant (for a given simulation problem) seems to be the key for finding 'good' random sequences. It is just as Niederreiter [87, p.164] recommends: "Therefore the user of PRN must be aware of the specific desirable statistical properties of the random samples in the current computational project and must choose PRN that are known to pass the corresponding statistical tests." In real-world applications, preferences based on such restrictions or weightings of relevance actually perform quite well: for most simulations you could conceive, your favorite random number sequence actually will behave more random than the sequence $(x_n = 0)_{n=0}^{N-1}$, although there is no reason for this in general.

Let us set out to find a criterion of quality which forces the hidden restrictions and weightings of relevance imposed by intuition to surface.

2.3 Two ways out

(2.3.1) Considering the set $\mathcal{F}$ of *all* statistical tests with level of significance α, we found this set to be too large to attribute any special status of 'randomness' to a finite sequence $\mathbf{x}$ of random numbers. The above quote from Niederreiter already suggests that, for choosing a 'good' $\mathbf{x}$, one should focus on "the specific desirable statistical properties of the random samples in the current computational project". This means we are free to ignore some tests (if they are based on statistical properties which are of no interest), while others should be more emphasized. But how to find the important tests, how to sort out the irrelevant ones?

An idea of Hammersley and Handscomb in [48, p.25] might help: "*one* of the tests that might have been applied is whether or not the random numbers yield an unbiased or a reliable answer to the Monte Carlo problem under study, and it is really only this test that interests us when we are ultimately concerned only with a final numerical solution to a particular problem. Taken in this second vein, the other tests are irrelevant". There is truth in this: what else could a user be interested in than in a good approximative answer to his simulation problem?

(2.3.2) A given 'Monte Carlo problem under study' is usually comprised of two successive approximations. The first one is pictured by Ripley [93, p.1] as follows: "In its technical sense simulation involves using a model to produce results, rather than experiment with the real system under study (which may not yet exist). For example, simulation is used to explore the extraction of oil from an oil reserve. If the model has a stochastic element, we have *stochastic simulation*". So the basic subject of the user's interest is some 'real system' T_R, which either cannot be described by a deterministic model, or whose deterministic model is far too complex to be tackled by today's mathematical methods directly. He therefore develops a stochastic model, i.e. a random variable[8] T_S coinciding with T_R 'in the mean', and focuses his attention on the expectation $E(T_S)$. This first approximation, the modelling of a real-world object by a mathematical, stochastic one is – for our purpose – of no interest; the quality of this modelling is solely in the user's responsibility.

Given T_S, Ripley [93, p.3] continues: "To make use of a model one has two choices:

1. To bring mathematical analysis to bear to try to understand the model's behavior. [...]

2. To experiment with the model."

If the user can derive $E(T_S)$ by analytical methods, he should do so and forget about any computer simulation[9]. Otherwise, he has to construct an *estimator* F for $E(T_S)$, which will be simulated on a computer. In general, an estimator F for $E(T_S)$ is a random variable with

$$E(F) = E(T_S)$$

and

$$V(F) \leq V(T_S).$$

To be simulated using a sequence of random numbers, F has to take the form $F(\mathbf{X})$ with $\mathbf{X} = (X_n)_{n=0}^{N-1}$, where the X_n are independent random variables, each equidistributed on $[0, 1[$. $\mathbf{X}$ will be substituted by some random numbers in the actual simulation. Observe that the movement from T_S to

[8] Any model which has a stochastic element depends on a more or less complex random quantity expressing the stochastic element. Since the model is governed by the random quantity, the model is a random quantity itself.

[9] Computer simulations of a stochastic model are often believed capable of revealing some extra information which could not be derived by analytical methods; but how can a computer–program make statements which are not contained in the mathematical model it is derived from?

F is no approximation since the random variables' expectations coincide. The second approximation is what we are concerned with: the user chooses some sequence of numbers $\mathbf{x}$, computes $F(\mathbf{x})$, and uses this value as an approximation of $E(F) = E(T_S)$ (Note that in our terminology, the computation of $F(\mathbf{x})$ already is the *whole* stochastic simulation[10]).

(2.3.3) When should we regard the numbers $\mathbf{x}$ as 'good' with respect to the simulation F? Well, obviously if

$$\Psi_{\mathbf{x}}(F) := |F(\mathbf{x}) - E(F)|$$

is small.

By introducing $\Psi_{\mathbf{x}}(F)$ to measure the quality of $\mathbf{x}$ in simulating F, we have preserved the idea that "it is really only this test that interests us when we are ultimately concerned only with a final numerical solution to a practical problem." Moreover, $\Psi_{\mathbf{x}}$ still does permit us to perform statistical tests. Let T be an α–test, i.e. $P(T = 1) = E(T) = \alpha$. If $\alpha < 1/2$ – which is not much of a restriction since nobody would reasonably apply a test with is likely to yield the wrong result – a sequence $\mathbf{x}$ passes T if and only if $\Psi_{\mathbf{x}}(T) < 1/2$. We can conclude that $\Psi_{\mathbf{x}}$ gives a measure more general and more flexible than statistical tests alone. Given F, $\Psi_{\mathbf{x}}(F)$ measures the deviation of the computed result from the desired result[11].

So far, we have de facto *eliminated* the notion of randomness; the problem is reduced to approximating $E(F)$ which in turn is an integral. However, if the user knows $\mathbf{x}$ to be good with respect to a given problem F, he may use these numbers to simulate F *acting as if they were random*; their behavior is – with respect to F – close to the expected behavior of random variables, and the elimination of randomness does not really matter. As Knuth [59, p.3] states (about computer–generated sequences): "The answer is that the sequence *isn't* random, but it *appears* to be. [...] Being 'apparently random' is perhaps all that can be said about any random sequence anyway."

(2.3.4) The measure of quality $\Psi_{\mathbf{x}}$ developed so far has one severe flaw: we can assess the quality of $\mathbf{x}$ with respect to *only one* problem F,

[10] For example, if T_S is a function of three random variables (the stochastic element in the model), i.e. $T_S = f(X_1, X_2, X_3)$, a quite popular choice for F is to fix a large number M, say, $M = 10^4$, $N \geq 3M$, and set $F(\mathbf{X}) := 1/M \sum_{n=0}^{M-1} f(X_{3n}, X_{3n+1}, X_{3n+2})$.

[11] A notion of quality similar in spirit but not formally is developed by Fuss in [42] using Petri nets to generate random decisions.

so all we can find is a good special-purpose sequence $\mathbf{x}$. This is not what people are looking for; they demand sequences of numbers which are good for simulating a variety of problems, a set $\mathcal{F}$ of problems F.
Given a set of problems $\mathcal{F}$ and a sequence $\mathbf{x}$ of random numbers, we consider

Criterion 2.3 $\mathbf{x}$ *is good with respect to* $\mathcal{F}$ *if*

$$\sup \Psi_{\mathbf{x}} := \sup \{\Psi_{\mathbf{x}}(F) : \ F \in \mathcal{F}\}$$

is small.

Based on this criterion, the existence of good sequences with respect to a certain set $\mathcal{F}$ of practical relevance is shown by the theory of good lattice points (see Hlawka [54] and Korobov [62]). The set $\mathcal{F}$ contains all functions of the form $F = 1/N \sum_{i=1}^{N} f$, where f is a function with in some sense rapidly decreasing Fourier coefficients. But, as Larcher and Traunfellner note in [67], "especially for dimensions $s \geq 3$, it turned out to be a challenge to give fast algorithms for finding good lattice points". A similar approach based on Haar instead of Fourier series was initiated by Sobol' (see [105, 106]) and developed to its full extent by Niederreiter (see [87, Chapter 4]). The shift from Fourier to Haar series has the advantage that methods to construct good samples for problems in higher dimensions have been found. The good samples $\mathbf{x}$ for this set $\mathcal{F}$ are called LP_τ–sets by Sobol' and $(t\text{-}m\text{-}s)$–nets by Niederreiter[12]. Shifting again, this time from Haar to Walsh series, Larcher and Traunfellner show in [66, 67] that $(t\text{-}m\text{-}s)$–nets are in some sense the *best possible choice* for approximating problems from the corresponding $\mathcal{F}$. In [101], Schmid gives an introduction to $(t\text{-}m\text{-}s)$–nets together with implementations for serial and parallel computer architectures.

For our present purpose, however, Criterion 2.3 is not flexible enough: the supremum tends to focus on 'local' properties of $\Psi_{\mathbf{x}}$. If, for example, $\mathbf{x} \neq \mathbf{y}$ and $\mathcal{F}$ is comprised of some statistical tests with level of significance $\alpha < 1/2$, then we have

$$\sup \Psi_{\mathbf{x}} < \sup \Psi_{\mathbf{y}}$$

if and only if

$$\begin{aligned} \forall T \in \mathcal{F} \quad &: \quad T(\mathbf{x}) = 0 \qquad \text{and} \\ \exists T \in \mathcal{F} \quad &: \quad T(\mathbf{y}) = 1. \end{aligned}$$

[12] Niederreiter's notion of a $(t\text{-}m\text{-}s)$–net is a generalization of Sobol's LP_τ–set.

The supremum assigns each sequence to one of two distinct sets: those sequences which pass all tests in $\mathcal{F}$ (the 'good' set) and those which fail at least one of them (the 'bad' set). This leaves us with no possibility to differentiate the 'good' set.
A possible remedy might be to extend $\mathcal{F}$ by additional problems or tests. But doing so, we stumble over the next inflexibility of Criterion 2.3: adding or removing just one problem from $\mathcal{F}$ can completely change our evaluation of $\sup \Psi_{\mathbf{x}}$!

(2.3.5) To get a more flexible criterion, recall that we have conjectured the necessity of $\mathcal{F}$ being *weighted* in (2.2.7). For a finite set of problems $\mathcal{F}$, let w be a nonnegative weighting function on $\mathcal{F}$. Excluding infinite weights, we may assume $\sum_{F\in\mathcal{F}} w(F) = 1$. With this, we regard $\mathbf{x}$ as 'good' with respect to $\mathcal{F}$ and w if

$$E_w(\Psi_{\mathbf{x}}) := \sum_{F\in\mathcal{F}} \Psi_{\mathbf{x}}(F)w(F)$$

is small – if $\Psi_{\mathbf{x}}$ is small *in the (weighted) mean.*
We generalize this to problem sets $\mathcal{F}$ of arbitrary size by employing a measure[13] on $\mathcal{F}$ in

Criterion 2.4 *Let $(\mathcal{F}, \mathcal{R}, \mu)$ be a probability space such that $\Psi_{\mathbf{x}}$ is a random variable.*
The finite sequence $\mathbf{x}$ of random numbers is good with respect to the problem set $\mathcal{F}$ and the probability measure μ on $\mathcal{F}$ if

$$E_\mu(\Psi_{\mathbf{x}}) := \int_{\mathcal{F}} \Psi_{\mathbf{x}} d\mu$$

is small.

Although Criterion 2.4 seems to be completely different from Criterion 2.3, we will show in Section 4.2 that under certain assumptions they are in fact equivalent.

In the spirit of this criterion, the search for good random number sequences becomes a two step process: first the user supplies some probabilistic description of the problem(s) he is interested in, and then we propose

[13] If you are not familiar with the notion of measure, just think of it as generalization of the weighting function w on finite $\mathcal{F}$. The weighted sum over a finite set is generalized to an integral over an infinite set.

a good number sequence according to the given information. Observe how much responsibility is loaded on the user this way. If his description is inadequate, so may be the proposed number sequence. But observe also that this is not just a mathematician's trick to avoid being blamed for the proposal of bad number sequences; since no finite sequence of numbers is random by itself, the user who wants to simulate a random variable using well-determined numbers has to take the burden of specifying which properties of randomness are relevant for him.

(2.3.6) There are some useful interpretations of $E_\mu(\Psi_\mathbf{x})$: as described above, μ can assign different 'grades of interest' to subsets of the problem set $\mathcal{F}$. Subsets the user considers vital are assigned high measure, while low measure is assigned to subsets which he thinks are interesting enough to be contained in $\mathcal{F}$ but not very important. If a sequence $\mathbf{x}$ is good with respect to $\mathcal{F}$ and μ in the above sense, the average weighted error by which $\mathbf{x}$ approximates the problems in $\mathcal{F}$ is small.
A second interpretation is from the mathematician's point of view. Given the set $\mathcal{F}$ and the measure μ, the mathematician faces the following problem. He has to propose a sequence $\mathbf{x}$ to the user. According to the probability measure μ, the user will randomly select a problem F from $\mathcal{F}$, estimate $E(F)$ by $F(\mathbf{x})$, and punish the mathematician (by lowering his reputation, cutting his salary, applying physical violence, or whatever else he considers appropriate) proportional to the resulting estimation error. $E_\mu(\Psi_\mathbf{x})$ is the punishment the mathematician can expect; choosing $\mathbf{x}$ as to minimize the expected punishment seems a very reasonable approach.

(2.3.7) The reader may object that measures on such complex sets as $\mathcal{F}$ are hard to find and that, even if an appropriate μ can be found, the computation of $E_\mu(\Psi_\mathbf{x})$ is – technically – not always possible. There is reason in this, of course. However, in all examples given in Chapter 4, we will get along with much less than the full-fledged measure μ, and even the precise extent of the problem set $\mathcal{F}$ will almost never be required. To demonstrate how this is possible, recall Niederreiter's recommendation from [87, p.164] to "be aware of the specific desirable statistical properties of the random samples in the current computational project and [...] choose PRN that are known to pass the corresponding statistical tests."
Suppose $\mathcal{G}$ is a set of estimators for a gambler's gain in successive rounds of roulette following one of several strategies[14]. $\Psi_\mathbf{x}(G)$ is the approximation

[14] There is quite a variety of ways to simulate this; just think of the various conceivable

error produced by $\mathbf{x}$ when simulating a given $G \in \mathcal{G}$. Next, suppose $\mathcal{H}$ is a set of statistical tests with level of significance $\alpha < 1/2$, each of which checks for the statistical properties desirable for simulating the gambler's gain[15], and let $\Phi_{\mathbf{x}}(T) := |T(\mathbf{x}) - \alpha|$ be the error[16] produced by $\mathbf{x}$ when subjected to a test $T \in \mathcal{H}$. Now let $\mathcal{F} := \mathcal{G} \times \mathcal{H}$ and let $(\mathcal{F}, \mathcal{R}, \mu)$ be a probability space such that $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ are random variables.

In general, precise knowledge of $(\mathcal{F}, \mathcal{R}, \mu)$ is required to compute $E_\mu(\Psi_{\mathbf{x}})$ and $E_\mu(\Phi_{\mathbf{x}})$.

Now suppose all we get to know is that the random quantities $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ are *positively correlated*. Informally, this means if $\mathbf{x}$ passes a test randomly selected from $\mathcal{H}$ (yielding $\Phi_{\mathbf{x}} < 1/2$), $\Psi_{\mathbf{x}}$ is likely to be small, too. Given only this information, we will show without much effort in Section 4.1 that

$$E_\mu(\Psi_{\mathbf{x}}|\Phi_{\mathbf{x}} < 1/2) \; < \; E_\mu(\Psi_{\mathbf{x}}) \; < \; E_\mu(\Psi_{\mathbf{x}}|\Phi_{\mathbf{x}} \geq 1/2)$$

for any probability space $(\mathcal{F}, \mathcal{R}, \mu)$ for which the positive correlation of $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ holds.

In this way, Knuth's argument [59, p.38] – which was based on no more than intuition in the light of Criterion 2.1 and 2.2 – does make sense: "If a sequence behaves randomly with respect to tests $T_1, T_2, \ldots, T_n$, we cannot be *sure* in general that it will not be a miserable failure when it is subjected to a further test T_{n+1}; yet each test gives us more and more confidence in the randomness of the sequence." When the hidden assumption of positive correlation is made explicit, the 'confidence' can be proven.

algorithms to transform the random numbers in $\mathbf{x}$ into lottery results.

[15] Again, there are many ways to check for these properties; think of varying the level of significance of the test, the sample size considered, the test statistic, or the region of acceptance.

[16] This is $\Phi_{\mathbf{x}}(T) < 1/2$ if $\mathbf{x}$ passes the test T and $\Phi_{\mathbf{x}}(T) \geq 1/2$ otherwise.

Chapter 3

Running for randomness – heat 2

Expect the unexpected!
– Douglas Adams, The Hitch Hiker's Guide to the Galaxy

We have seen that no finite 0-1-sequence is per se more random than any other. In this chapter, we will state this in a more general context, for any finite sequence of numbers in $[0,1[$. The reader who is convinced that analogous results hold for this more complex set and who does not want to get involved into mathematical details may skip this chapter. Here, we will use no ideas not already presented before, but we hope the mathematical generalization will provide a deeper insight in the phenomenon's nature.

3.1 Finite sequences of random numbers in $[0,1[$

(3.1.1) Our previous considerations were restricted to the particular set $\{0,1\}^N$. We have tried to find out if a sequence $\mathbf{x}$ of N numbers x_n in $\{0,1\}$ can be used to model a sequence $\mathbf{X}$ of N random variables X_n which are independent and equidistributed on $\{0,1\}$. The sequence $\mathbf{X}$ is a random variable with values in $\{0,1\}^N$ and – due to the independence of the X_n – it is equidistributed on $\{0,1\}^N$. Thus $\mathbf{X}$ is a random variable on the probability space $(\{0,1\}^N, \mathcal{P}(\{0,1\}^N), p)$, where $\mathcal{P}(\{0,1\}^N)$ is the family of all subsets of $\{0,1\}^N$ and p is the normalized counting measure on $\{0,1\}^N$.

Searching for a 'good' sequence of random numbers, we tried to model the random variable $\mathbf{X}$ by means of a specific sample $\mathbf{x} \in \{0,1\}^N$ (see (2.2.5)). Now we will do the same for the more complex probability space $([0,1[^N, \mathcal{B}_{[0,1[^N}, \lambda_N)$. $\mathcal{B}_{[0,1[^N}$ is the Borel σ-algebra on $[0,1[^N$ with respect to the standard topology and λ_N is the N-dimensional Lebesque measure restricted to $[0,1[^N$. This probability space has the following three helpful mathematical properties which we will use in the proofs.

(3.1.2) The real-valued random variables $X_1, \ldots, X_N$ are independent and equidistributed on $[0,1[$ if and only if the vector $\mathbf{X} = (X_1, \ldots, X_N)$ is a random variable on $[0,1[^N$ with probability measure λ_N. So the probability space $([0,1[^N, \mathcal{B}_{[0,1[^N}, \lambda_N)$ is exactly what we are concerned with when searching for a 'good' sequence of random numbers in $[0,1[$.

(3.1.3) For $\mathbf{x}, \mathbf{y} \in [0,1[^N$, we define $\mathbf{x} + \mathbf{y}$ as the component-wise sum of $\mathbf{x}$ and $\mathbf{y}$, reduced modulo 1: for $\mathbf{x} = (x_n)_{n=0}^{N-1}$ and $\mathbf{y} = (y_n)_{n=0}^{N-1}$, let

$$\mathbf{x} + \mathbf{y} := (x_n + y_n \bmod 1)_{n=0}^{N-1}.$$

Then $([0,1[^N, +)$ is an abelian group.

(3.1.4) The measure λ_N is invariant under translation: for any $\mathbf{A} \in \mathcal{B}_{[0,1[^N}$ and any $\mathbf{c} \in [0,1[^N$, we have

$$\lambda_N(\mathbf{A} + \mathbf{c}) = \lambda_N(\mathbf{A}).$$

Another way to state this is to observe that, for any $\mathbf{c} \in [0,1[^N$, the translation

$$T_{\mathbf{c}} : \mathbf{x} \longmapsto \mathbf{x} + \mathbf{c}$$

is a measure-preserving function.

The background of this is the fact that there is a topology τ on $[0,1[^N$ such that $([0,1[^N, +, \tau)$ is a *compact* abelian group. τ is very similar to the standard topology; in fact, the Borel σ-algebra with respect to τ is equal to $\mathcal{B}_{[0,1[^N}$. For any compact abelian group, the existence of an unique translation-invariant probability measure, the so called Haar measure can be proven. In our case this measure is just λ_N. For more information on compact abelian groups and Haar measures, see Cohn [11, Section 9.2].
We can, of course, state our problem of finding a 'good' $\mathbf{x}$ for a general compact abelian group instead of the specific group $([0,1[^N, +, \tau)$. Proposition 3.1 and 3.3 as well as Part 2 of Proposition 3.2, which we prove in this

chapter, can easily be proven in this general case. The proofs of Part 1 and 3 of Proposition 3.2 however, which are in fact the most interesting ones, rely heavily on the specific structure of $([0,1[^N,+,\tau)$.

(3.1.5) Just as we considered the set of all statistical tests on $\{0,1\}^N$ with level of significance α, we will now consider the analogous set of all α–tests on $[0,1[^N$ as well as more complex sets of problems. A problem set might be, say, for any given random variable Y, the set $\mathcal{F}$ of all estimators for $E(Y)$ or the set $\mathcal{F}$ of all such estimators whose variances are less than some $\delta > 0$, and so on. In general, any $\mathcal{F} \subseteq L^1([0,1[^N;\mathbf{R})$ can be taken as a problem set, this is any set of real-valued functions F whose expectations $E(F)$ exist and are finite[1].

We will show that no 'good' random numbers can be found as long as the test set $\mathcal{F}$ is *invariant under translation.* So if F is a test function from $\mathcal{F}$ (which serves some purpose like being a statistical test or estimating some $E(Y)$), then each translated test function $F \circ T_{\mathbf{c}}$ (which is a statistical test or an estimator for $E(Y)$ as well) is in $\mathcal{F}$, too. Let us state this formally.

Definition 3.1 *A set* $\mathcal{F} \subseteq L^1([0,1[^N;\mathbf{R})$ *is called* unbiased *if*

$$\forall F \in \mathcal{F}\ \forall \mathbf{c} \in [0,1[^N\colon\ F \circ T_{\mathbf{c}} \in \mathcal{F}.$$

It is easy to see that the set of all α-tests or the set of all estimators for a given random variable's expectation is unbiased because the translation $T_{\mathbf{c}}$ is measure-preserving.

(3.1.6) To evaluate the 'quality' of a given sample $\mathbf{x} \in [0,1[^N$, we proceed as in Section 2.3: for any simulation problem $F \in L^1([0,1[^N;\mathbf{R})$, we measure the 'quality' of $\mathbf{x}$ in simulating F by

$$\Psi_{\mathbf{x}}(F) := |E(F) - F(\mathbf{x})|,$$

where $E(F)$ is the integral of F with respect to λ_N. For a given test set $\mathcal{F}$ and a fixed error bound $\epsilon > 0$, we consider the set $\mathcal{F}_{\mathbf{x}}$ of those functions whose expectations $\mathbf{x}$ approximates by an error less than ϵ:

$$\mathcal{F}_{\mathbf{x}} := \{F \in \mathcal{F} :\ \Psi_{\mathbf{x}}(F) < \epsilon\}.$$

[1] The symbol $L^1([0,1[^N;\mathbf{R})$ denotes the set of all real-valued, measurable functions on $[0,1[^N$ for which the Lebesque integral $\int_{[0,1[^N} |F| d\lambda_N$ is well-defined and finite. Note that $E(F)$ is just defined as $\int_{[0,1[^N} F d\lambda_N$.

Note that $\mathcal{F}_{\mathbf{x}}$ implicitly depends on ϵ. For the rest of this chapter, we assume that ϵ is some fixed, positive error bound.
The quality of $\mathbf{x}$ with respect to $\mathcal{F}$ depends on the size of $\mathcal{F}_{\mathbf{x}}$. We will show that, for *unbiased* test sets $\mathcal{F}$, the size of $\mathcal{F}_{\mathbf{x}}$ *does not depend on* $\mathbf{x}$.

Before we can do so, we have to deal with a phenomenon which did not occur in the finite probability space on $\{0,1\}^N$ considered before: the existence of nonempty sets of measure zero. The function $1_{\emptyset}$, which assigns 0 to each $\mathbf{x}$, is a statistical test with level of significance 0 – it accepts every sample. This statistical test is of course meaningless. The function $1_{\{\mathbf{x}\}}$ is a statistical test with level of significance 0 as well, but it rejects the sample $\mathbf{x}$ since $1_{\{\mathbf{x}\}}(\mathbf{x}) = 1$. However, this test looks quite pathological, too. We refer to all functions which have the same property as trivial functions.

Definition 3.2 *A function $F \in L^1([0,1[^N;\mathbf{R})$ is called* trivial *if*

$$\lambda_N\left(\left\{\mathbf{x} \ \in\ [0,1[^N\colon\ \Psi_{\mathbf{x}} < \epsilon\right\}\right) \in \{0,1\}.$$

3.2 Three propositions

We have presented four criteria to measure the quality of a given sample $\mathbf{x}$ in Chapter 2. Here, we will show that whenever $\mathcal{F} \subseteq L^1([0,1[^N;\mathbf{R})$ is unbiased, the first three criteria do not allow the proposal of a 'good' sequence of random numbers $\mathbf{x} \in [0,1[^N$.

Proposition 3.1 *For any unbiased set $\mathcal{F} \subseteq L^1([0,1[^N;\mathbf{R})$ and any $\mathbf{x},\mathbf{y} \in [0,1[^N$, there exists a bijection Φ from $\mathcal{F}_{\mathbf{x}}$ onto $\mathcal{F}_{\mathbf{y}}$.*

Remark: If $\mathcal{F}$ is unbiased, Proposition 3.1 yields

$$\forall \mathbf{x},\mathbf{y} \in [0,1[^N\colon\ \#\mathcal{F}_{\mathbf{x}} = \#\mathcal{F}_{\mathbf{y}}.$$

In this case, Criterion 2.1 is useless.

Proof: Let $\mathcal{F} \in L([0,1[^N;\mathbf{R})$ be unbiased and let $\mathbf{x},\mathbf{y} \in [0,1[^N$ be fixed. We will construct a bijective and measure-preserving function $T : [0,1[^N \rightarrow [0,1[^N$, use T to define the mapping

$$\begin{array}{rcl}\Phi : \mathcal{F}_{\mathbf{x}} & \longrightarrow & \mathcal{F}_{\mathbf{y}} \\ F & \longmapsto & F \circ T,\end{array}$$

and show that Φ is bijective.

If $-\mathbf{y}$ is the inverse of $\mathbf{y}$ in $([0,1[^N,+)$, we define T as the translation by $\mathbf{x}-\mathbf{y}$:

$$T := T_{\mathbf{x}-\mathbf{y}}.$$

T is bijective, because $([0,1[^N,+)$ is an abelian group, and it is measure-preserving because λ_N is invariant under translation.

To see that Φ is well-defined, let $F \in \mathcal{F}_{\mathbf{x}}$:

$$\Psi_{\mathbf{x}}(F) < \epsilon.$$

Because $F \circ T(\mathbf{y}) = F(\mathbf{x})$ and because T is measure-preserving, we have

$$\Psi_{\mathbf{y}}(F \circ T) < \epsilon,$$

which means $\Phi(F) \in \mathcal{F}_{\mathbf{y}}$.

To show the injectivity of Φ, let $F, F' \in \mathcal{F}_{\mathbf{x}}$ and

$$F \circ T = F' \circ T.$$

Since T is bijective, it immediately follows that $F = F'$.

To show the surjectivity of Φ, let $G \in \mathcal{F}_{\mathbf{y}}$. Using the same argument as in showing that Φ is well-defined, we get

$$G \circ T_{\mathbf{y}-\mathbf{x}} \in \mathcal{F}_{\mathbf{x}}$$

and, due to the commutativity of $+$,

$$\Phi(G \circ T_{\mathbf{y}-\mathbf{x}}) = G$$

With this, the mapping Φ is a bijection from $\mathcal{F}_{\mathbf{x}}$ onto $\mathcal{F}_{\mathbf{y}}$. □

Of course, the existence of a bijection between two sets does not imply they are of equal size. There is, for example, a bijection from $[0,1/2[$ onto $[0,1[$. But since $[0,1/2[$ is a proper subset of $[0,1[$, $[0,1/2[\subset [0,1[$, the former is usually considered smaller than the latter[2]. We show that such an inclusion cannot occur between the sets $\mathcal{F}_{\mathbf{x}}$ and $\mathcal{F}_{\mathbf{y}}$ when $\mathcal{F}$ is unbiased and non-trivial.

[2] Observe that we use the symbol $A \subset B$ to denote that A is a proper subset of B, this is $A \subseteq B$ and $B \setminus A \neq \emptyset$.

Proposition 3.2 *Let $\mathcal{F} \subseteq L^1([0,1[^N; \mathbf{R})$ be unbiased and let $\mathbf{x} \in [0,1[^N$ be fixed.*

1. *For almost all $\mathbf{y} \in [0,1[^N$, the following holds: if*
$$\mathcal{F}_{\mathbf{x}} \subset \mathcal{F}_{\mathbf{y}},$$
then $\mathcal{F}$ is a set of trivial functions.
2. *For all $\mathbf{y} \in [0,1[^N$, the following holds: if*
$$\mathcal{F}_{\mathbf{x}} \subset \mathcal{F}_{\mathbf{y}},$$
then
$$\mathcal{F}_{\mathbf{y}} \subset \mathcal{F}_{\mathbf{x}+2(\mathbf{y}-\mathbf{x})} \subset \mathcal{F}_{\mathbf{x}+3(\mathbf{y}-\mathbf{x})} \subset \cdots.$$
3. *If $\mathbf{x}$ and $\mathbf{y} \in [0,1[^N$ are rational in each coordinate, then*
$$\mathcal{F}_{\mathbf{x}} \not\subset \mathcal{F}_{\mathbf{y}}.$$

Remark: If we take two samples $\mathbf{x}$ and $\mathbf{y}$ at random, and if we happen to find $\mathcal{F}_{\mathbf{x}} \subset \mathcal{F}_{\mathbf{y}}$, then Part 1 states that almost certainly $\mathbf{y}$ is superior to $\mathbf{x}$ only with respect to trivial functions. Part 2 states that whenever such an inclusion occurs, we can construct a sequence $(\mathbf{x} + n(\mathbf{y} - \mathbf{x}))_{n=0}^{\infty}$ of samples where each element is better than its predecessor; in this case, we cannot consider $\mathbf{y}$ as 'good' because we can actually compute an infinite number of samples which are 'better'. For computer simulations, Part 3 is the most interesting one. It states that for those sequences $\mathbf{x}$, $\mathbf{y}$ of random numbers which can be represented in a computer's finite-precision floating-point arithmetic, an inclusion of $\mathcal{F}_{\mathbf{x}}$ in $\mathcal{F}_{\mathbf{y}}$ cannot occur at all. As a corollary, this renders Criterion 2.2 useless whenever $\mathcal{F}$ is unbiased.

The Parts 2 and 3 are easy to prove, but Part 1 is not. To prove it, we use the bijection Φ from $\mathcal{F}_{\mathbf{x}}$ onto $\mathcal{F}_{\mathbf{y}}$ as constructed in the proof of Proposition 3.1: each $F \in \mathcal{F}_{\mathbf{x}}$ is translated to $F \circ T_{\mathbf{x}-\mathbf{y}}$, which is an element of $\mathcal{F}_{\mathbf{y}}$. We show that this translation of functions in $\mathcal{F}$ corresponds to the translation of some measurable subsets of $[0,1[^N$. The notion of ergodicity applied to these subsets is the key to Part 1. Recall the definition of ergodicity[3].

[3] For more than just the definition of this very interesting notion, we refer the reader to Parry [89], Petersen [91], or Walters [111].

Definition 3.3 *A measure-preserving mapping* $T : [0,1[^N \longrightarrow [0,1[^N$ *is called* ergodic *if, for any* $\mathbf{A} \in \mathcal{B}_{[0,1[^N}$, *the following implication holds:*

$$T^{-1}(\mathbf{A}) = \mathbf{A} \implies \lambda_N(\mathbf{A}) \in \{0,1\}.$$

Besides the notion of ergodicity, we need four lemmata to prove Proposition 3.2. The first two are used to state that the translation $T_{\mathbf{c}}$ is ergodic for almost all $\mathbf{c}$. The third shows that the translation of functions in $\mathcal{F}$ corresponds to the translation of certain measurable subsets of $[0,1[^N$. The fourth translates the concept of trivial functions to these measurable subsets.

Lemma 3.1 *Let* $\mathbf{c} = (c_n)_{n=0}^{N-1} \in [0,1[^N$. *Then* $T_{\mathbf{c}}$ *is ergodic if and only if the numbers* $c_0, \dots, c_{N-1}$, *and* 1 *are rationally independent.*

We do not prove this result here. The interested reader might consult Parry [89, Chapter 1]. Recall that the rational independence of $c_0, \dots, c_{N-1}, 1$ means that there are no integers $h_0, \dots, h_{N-1}, k$ satisfying $\sum_{n=0}^{N-1} h_n c_n = k$.

Lemma 3.2 *The translation* $T_{\mathbf{c}}$ *is ergodic for almost all* $\mathbf{c} \in [0,1[^N$.

Proof: We show that

$$\mathbf{A} := \{\mathbf{c} = (c_n)_{n=0}^{N-1} \in [0,1[^N\colon\ c_0, \dots, c_{N-1}, 1 \text{ are rationally dependent }\}$$

is a set of measure zero. The condition of rational dependence of the c_n and 1 can be read as '$\mathbf{c}$ lies on a hyperplane defined by the vector $\mathbf{h}$ and the scalar k, where $\mathbf{h}$ is a nonzero N-dimensional integer vector and k is an integer'. $\mathbf{A}$ is a subset of the union of all these hyperplanes. Since any hyperplane has measure zero and since the set of the above hyperplanes is countable, their union and its subset $\mathbf{A}$ have measure zero, too. □

Lemma 3.3 *Let* $\mathcal{F} \subseteq L^1([0,1[^N; \mathbf{R})$ *be unbiased. Then there is a set* $\mathcal{G} \subset \mathcal{F}$ *with*

$$\forall \mathbf{x} \in [0,1[^N\ \forall F \in \mathcal{G} :\ \exists \mathbf{C}_{\mathbf{x},F} \in \mathcal{B}_{[0,1[^N}$$

such that

1.

$$\mathcal{F}_{\mathbf{x}} = \bigcup_{F \in \mathcal{G}} \{F \circ T_{\mathbf{c}} :\ \mathbf{c} \in \mathbf{C}_{\mathbf{x},F}\}.$$

2. The sets in the above union are mutually disjoint.

3.
$$\forall \mathbf{x}, \mathbf{y} \in [0,1[^N \; \forall F \in \mathcal{G}: \; \mathbf{C}_{\mathbf{y},F} = \mathbf{C}_{\mathbf{x},F} + (\mathbf{x} - \mathbf{y}).$$

Proof: Let $\mathcal{F}$ be unbiased. We define an equivalence relation $\equiv$ on $\mathcal{F}$ as

$$F \equiv G \iff \exists \mathbf{c} \in [0,1[^N\!: \; F = G \circ T_{\mathbf{c}}.$$

It is easy to see that $\equiv$ is in fact an equivalence relation: it is reflexive since there is a neutral element, the zero-vector in the group $([0,1[^N, +)$; it is symmetric since, for each $\mathbf{c}$, there is an inverse element in the group, too; finally, it is transitive also because $([0,1[^N, +)$ is a group.
In the following argument, we define $\mathcal{G}$ as a set of representatives[4] of $\mathcal{F}$ with respect to $\equiv$, and construct the sets $\mathbf{C}_{\mathbf{x},F}$ accordingly.

Since $\mathcal{G}$ is a set of representatives and since $\mathcal{F}$ is unbiased, we have

$$\mathcal{F} = \bigcup_{F \in \mathcal{G}} \left\{ F \circ T_{\mathbf{c}} : \; \mathbf{c} \in [0,1[^N \right\},$$

and this union is taken over disjoint sets. Now we represent the set $\mathcal{F}_{\mathbf{x}}$ in a similar form:

$$\begin{aligned} \mathcal{F}_{\mathbf{x}} &= \{F \in \mathcal{F} : \; \Psi_{\mathbf{x}}(F) < \epsilon\} \\ &= \bigcup_{F \in \mathcal{G}} \left\{ F \circ T_{\mathbf{c}} : \; \mathbf{c} \in [0,1[^N, \; \Psi_{\mathbf{x}}(F \circ T_{\mathbf{c}}) < \epsilon \right\} \\ &= \bigcup_{F \in \mathcal{G}} \left\{ F \circ T_{\mathbf{c}} : \; \mathbf{c} \in \left\{ \mathbf{c} \in [0,1[^N\!: \; \Psi_{\mathbf{x}}(F \circ T_{\mathbf{c}}) < \epsilon \right\} \right\}. \end{aligned}$$

As above, the union is taken over disjoint sets. We define the sets $\mathbf{C}_{\mathbf{x},F}$ as

$$\mathbf{C}_{\mathbf{x},F} := \left\{ \mathbf{c} \in [0,1[^N\!: \; \Psi_{\mathbf{x}}(F \circ T_{\mathbf{c}}) < \epsilon \right\}, \tag{1}$$

which are measurable because F is measurable itself. With this, the Parts 1 and 2 are proved.

For Part 3, recall the mapping Φ we used in the proof of Proposition 3.1: Φ is a bijective function from $\mathcal{F}_{\mathbf{x}}$ to $\mathcal{F}_{\mathbf{y}}$, which maps F to $F \circ T_{\mathbf{x}-\mathbf{y}}$. Let $F \in \mathcal{G}$. On one hand, we have

$$\mathcal{F}_{\mathbf{y}} = \bigcup_{F \in \mathcal{G}} \{ F \circ T_{\mathbf{c}} : \; \mathbf{c} \in \mathbf{C}_{\mathbf{y},F} \},$$

[4] Which does exist if we assume the Axiom of Choice.

and on the other, since Φ is bijective,

$$\begin{aligned}\mathcal{F}_{\mathbf{y}} &= \Phi(\mathcal{F}_{\mathbf{x}}) \\ &= \bigcup_{F\in\mathcal{G}} \Phi\left(\{F\circ T_{\mathbf{c}} :\ \mathbf{c}\in \mathbf{C}_{\mathbf{x},F}\}\right) \\ &= \bigcup_{F\in\mathcal{G}} \left\{F\circ T_{\mathbf{c}'} :\ \mathbf{c}'\in \mathbf{C}_{\mathbf{x},F}+(\mathbf{x}-\mathbf{y})\right\}.\end{aligned}$$

Because the sets $\mathbf{C}_{\mathbf{x},F}$ and $\mathbf{C}_{\mathbf{y},F}$ are uniquely determined by (1), we have

$$\mathbf{C}_{\mathbf{y},F} = \mathbf{C}_{\mathbf{x},F} + (\mathbf{x}-\mathbf{y}).$$

□

Lemma 3.4 *Let $\mathcal{F}\subseteq L^1([0,1[^N;\mathbf{R})$ be unbiased and let the sets $\mathcal{G},\mathbf{C}_{\mathbf{z},F}$ ($\mathbf{z}\in[0,1[^N$, $F\in\mathcal{G}$) be defined as in Lemma 3.3.*
If

$$\exists\mathbf{x}\in[0,1[^N\ \forall F\in\mathcal{G}:\ \lambda_N(\mathbf{C}_{\mathbf{x},F})\in\{0,1\}, \tag{2}$$

then $\mathcal{F}$ is a set of trivial functions.

Proof: Let $\mathcal{F}\subseteq L^1([0,1[^N;\mathbf{R})$ be unbiased, let $\mathbf{x}\in[0,1[^N$, and suppose (2) holds for the given $\mathbf{x}$. It is sufficient to show that $\mathcal{G}$ is a set of trivial functions because $\mathcal{G}$ is a set of representatives of $\mathcal{F}$ with respect to the translations $T_{\mathbf{c}}$ and because, if $F\in\mathcal{G}$ is trivial, each $F\circ T_{\mathbf{c}}$ is trivial, too.

Let $F\in\mathcal{G}$. Due to (1) we have

$$\begin{aligned}\mathbf{C}_{\mathbf{x},F} &= \left\{\mathbf{c}\in[0,1[^N:\ \Psi_{\mathbf{x}}(F\circ T_{\mathbf{c}})<\epsilon\right\} \\ &= \left\{\mathbf{c}\in[0,1[^N:\ \Psi_{\mathbf{x}+\mathbf{c}}(F)<\epsilon\right\} \\ &= \left\{\mathbf{c}'-\mathbf{x}\in[0,1[^N:\ \Psi_{\mathbf{c}'}(F)<\epsilon\right\} \\ &= \left\{\mathbf{c}'\in[0,1[^N:\ \Psi_{\mathbf{c}'}(F)<\epsilon\right\}-\mathbf{x} \\ &=:\ \mathbf{A}-\mathbf{x}.\end{aligned}$$

Since $\lambda_N(\mathbf{C}_{\mathbf{x},F})\in\{0,1\}$ and since λ_N is invariant under translation, we get that $\lambda_N(\mathbf{A})\in\{0,1\}$ and that F is trivial. □

Now we are ready to prove Proposition 3.2. We show Part 1 to 3 in reverse order because the argument runs smoother this way.

Proof of Proposition 3.2: Let $\mathcal{F} \subseteq L^1([0,1[^N;\mathbf{R})$ be unbiased, $\mathbf{x} \in [0,1[^N$, and let $\mathcal{G}$ and the sets $\mathbf{C}_{\mathbf{z},F}$ ($\mathbf{z} \in [0,1[^N$, $F \in \mathcal{G}$) as in Lemma 3.3. For a fixed $\mathbf{y} \in [0,1[^N$, let $\Phi : \mathcal{F}_{\mathbf{x}} \to \mathcal{F}_{\mathbf{y}}$ be defined as in the proof of Proposition 3.1: F is mapped to $\Phi(F) = F \circ T_{\mathbf{x}-\mathbf{y}}$. Finally, let $\mathbf{y} \in [0,1[^N$ such that

$$\mathcal{F}_{\mathbf{x}} \subset \mathcal{F}_{\mathbf{y}}. \tag{3}$$

For Part 3, let both $\mathbf{x}$ and $\mathbf{y}$ be rational in each coordinate. Then the same holds for $\mathbf{c} := \mathbf{x} - \mathbf{y}$. Let $\mathbf{0} = (0,\ldots,0)$ be the zero-vector in $[0,1[^N$. Since $\mathbf{c}$ is rational, there is an integer $n > 0$ such that $n\mathbf{c} = \mathbf{0}$ in the abelian group $([0,1[^N,+)$. For $F \in \mathcal{G}$, (1) and (3) imply[5] that $\mathbf{C}_{\mathbf{x},F} \subseteq \mathbf{C}_{\mathbf{y},F}$. Because of Lemma 3.3, Part 3, for any $F \in \mathcal{G}$, we have

$$\begin{aligned} \mathbf{C}_{\mathbf{x},F} &\subseteq \mathbf{C}_{\mathbf{y},F} = \mathbf{C}_{\mathbf{x},F} + \mathbf{c} \\ &\subseteq \mathbf{C}_{\mathbf{x},F} + 2\mathbf{c} \\ &\vdots \\ &\subseteq \mathbf{C}_{\mathbf{x},F} + n\mathbf{c} = \mathbf{C}_{\mathbf{x},F}. \end{aligned}$$

With Lemma 3.3, this yields $\mathcal{F}_{\mathbf{x}} = \mathcal{F}_{\mathbf{y}}$, which is a contradiction to (3). Hence (3) cannot hold whenever $\mathbf{x}$ and $\mathbf{y}$ are rational in each coordinate. This proves Part 3.

To show Part 2, note that $\Phi(\mathcal{F}_{\mathbf{x}}) = \mathcal{F}_{\mathbf{y}}$, so (3) can be written as

$$\mathcal{F}_{\mathbf{x}} \subset \Phi(\mathcal{F}_{\mathbf{x}}) = \mathcal{F}_{\mathbf{x}+(\mathbf{y}-\mathbf{x})}.$$

Inductively, this yields

$$\mathcal{F}_{\mathbf{x}+n(\mathbf{y}-\mathbf{x})} \subset \mathcal{F}_{\mathbf{x}+(n+1)(\mathbf{y}-\mathbf{x})}$$

for any integer $n \geq 0$ and proves Part 2.

To prove Part 1, let $\mathbf{y} \in [0,1[^N$ such that $T_{\mathbf{x}-\mathbf{y}}$ is ergodic. Observe that this holds for almost all $\mathbf{y}$[6]. With Lemma 3.4, it is sufficient to show that,

[5] If we can choose $\mathbf{c} \in \mathbf{C}_{\mathbf{x},F} \setminus \mathbf{C}_{\mathbf{y},F}$, then (1) gives $\Psi_{\mathbf{x}}(F \circ T_{\mathbf{c}}) < \epsilon$ and $\Psi_{\mathbf{y}}(F \circ T_{\mathbf{c}}) \geq \epsilon$. But this means $F \circ T_{\mathbf{c}} \in \mathcal{F}_{\mathbf{x}} \setminus \mathcal{F}_{\mathbf{y}}$, which is a contradiction to (3).

[6] Due to Lemma 3.2, $T_{\mathbf{y}}$ is ergodic for almost all $\mathbf{y} \in [0,1[^N$. Since the translation $T_{\mathbf{x}-2\mathbf{y}}$ is measure-preserving, the mapping

$$T_{\mathbf{x}-\mathbf{y}} = T_{\mathbf{y}} \circ T_{\mathbf{x}-2\mathbf{y}}$$

is ergodic for almost all $\mathbf{y}$, too.

for any $F \in \mathcal{G}$, the measure $\lambda_N(\mathbf{C}_{\mathbf{x},F})$ is either 0 or 1.
Because of (1) and (3), the following relations between $\mathbf{C}_{\mathbf{x},F}$ and $\mathbf{C}_{\mathbf{y},F}$ are possible[7]:

$$\mathbf{C}_{\mathbf{x},F} = \mathbf{C}_{\mathbf{y},F}$$

or

$$\mathbf{C}_{\mathbf{x},F} \subset \mathbf{C}_{\mathbf{y},F}.$$

In the first case, $\mathbf{C}_{\mathbf{x},F} = \mathbf{C}_{\mathbf{y},F}$, Lemma 3.3, Part 3 yields

$$\mathbf{C}_{\mathbf{x},F} \ = \ \mathbf{C}_{\mathbf{y},F} \ = \ \mathbf{C}_{\mathbf{x},F} + (\mathbf{x} - \mathbf{y})$$

or, equivalently,

$$\mathbf{C}_{\mathbf{x},F} = T^{-1}_{\mathbf{x}-\mathbf{y}}\left(\mathbf{C}_{\mathbf{x},F}\right).$$

Since $T_{\mathbf{x}-\mathbf{y}}$ is ergodic, $\lambda_N(\mathbf{C}_{\mathbf{x},F})$ is either 0 or 1.
In the second case, $\mathbf{C}_{\mathbf{x},F} \subset \mathbf{C}_{\mathbf{y},F}$, the already proven Part 2 of this proposition yields that, for $\mathbf{c} := \mathbf{x} - \mathbf{y}$, we have

$$\forall n \geq 0: \ \mathbf{C}_{\mathbf{x},F} + n\mathbf{c} \subset \mathbf{C}_{\mathbf{x},F} + (n+1)\mathbf{c}.$$

Because λ_N is invariant under translation, the measure of each $\mathbf{C}_{\mathbf{x},F} + n\mathbf{c}$ is equal to $\lambda_N(\mathbf{C}_{\mathbf{x},F})$. Now we define

$$\mathbf{C} := \bigcup_{n=0}^{\infty}\left(\mathbf{C}_{\mathbf{x},F} + n\mathbf{c}\right).$$

Then, because the measure λ_N is finite and therefore continuous from below[8], we have

$$\lambda_N(\mathbf{C}) = \lambda_N(\mathbf{C}_{\mathbf{x},F}).$$

Since $\mathbf{C}_{\mathbf{x},F} + n\mathbf{c} \subset \mathbf{C}_{\mathbf{x},F} + (n+1)\mathbf{c}$, we have $\mathbf{C} = \cup_{n=1}^{\infty}(\mathbf{C}_{\mathbf{x},F} + n\mathbf{c})$. Hence

$$\mathbf{C} + \mathbf{c} = \mathbf{C}$$

or, equivalently,

$$T_{\mathbf{c}}^{-1}(\mathbf{C}) = \mathbf{C}.$$

Since $T_{\mathbf{c}}$ is ergodic, the measure $\lambda_N(\mathbf{C}) = \lambda_N(\mathbf{C}_{\mathbf{x},F})$ is either 0 or 1. □

The failure of Criterion 2.3 when $\mathcal{F}$ is unbiased is stated in

[7] As we have seen in the proof of Part 3, the case $\mathbf{C}_{\mathbf{x},F} \setminus \mathbf{C}_{\mathbf{y},F} \neq \emptyset$ is impossible.

[8] A measure μ is continuous from below if, for arbitrary measurable sets A_n and A with $\cup_{n=1}^{\infty} A_n = A$, the relation $lim_{n\to\infty}\mu(\cup_{i=1}^{n} A_n) = \mu(A)$ holds.

Proposition 3.3 *Let $\mathcal{F} \subseteq L^1([0,1[^N;\mathbf{R})$ be unbiased. Then*

$$\forall \mathbf{x},\mathbf{y} \in [0,1[^N\colon\ \sup_{\mathcal{F}} \Psi_{\mathbf{x}} = \sup_{\mathcal{F}} \Psi_{\mathbf{y}}.$$

Proof: Let $\mathcal{F} \in L^1([0,1[^N;\mathbf{R})$ be unbiased and let $\mathbf{x},\mathbf{y} \in [0,1[^N$. Recall the bijective mapping

$$\begin{aligned} \Phi : \mathcal{F}_{\mathbf{x}} &\longrightarrow \mathcal{F}_{\mathbf{y}} \\ F &\longmapsto F \circ T_{\mathbf{x}-\mathbf{y}} \end{aligned}$$

we used in the proof of Proposition 3.1. By the same argument as used to show that Φ is well-defined, we get

$$\Psi_{\mathbf{x}}(F) = \Psi_{\mathbf{y}}(F \circ T_{\mathbf{x}-\mathbf{y}}),$$

which, because of the bijectivity of Φ, completes the proof. □

3.3 A critical remark

(3.3.1) Note that the condition of unbiasedness we demanded for $\mathcal{F}$ is quite restrictive. We could make use of Criterion 2.1, 2.2, and 2.3 by considering sets $\mathcal{F}$ which are not unbiased (for an example of this, see our remark concerning Criterion 2.3 in (2.3.4)).

Anyway, all we wanted to say was that, *in general*, no finite sample can be considered more random than any other (see (1.0.3)). With the example of unbiased sets $\mathcal{F}$, and in particular since $L^1([0,1[^N;\mathbf{R})$ is unbiased, this is shown.

Moreover, searching for good general-purpose random number sequences to run stochastic simulations on computers, one almost invariably stumbles over unbiased sets $\mathcal{F}$. Think of a user who wants to simulate Y, a [your-simulation-problem-here] on a computer. In particular, if Y is a complex quantity, there are numerous possible estimator for $E(Y)$, and each of them can be implemented on a computer in numerous different ways. The user does not want a separate random number sequence for each individual estimator F, but one which performs well for all F[9]. Although additional

[9] Using the same sequence with different estimators is quite useful for debugging the code and comparing the estimators.

knowledge about the relevant estimators may be available, the information which actually *reaches* the mathematician is often no more than that 'one searches for a sequence well suited for simulating Y'. In this case, the corresponding set $\mathcal{F}$ is

$$\mathcal{F} = \{F \in L^1([0,1[^N; \mathbf{R}) : \; E(F) = E(Y)\}$$

which is, of course, unbiased.

As already hinted in (2.3.6) and as we will see in the next chapter, the use of a probability space over $\mathcal{F}$ provides us with a quite handy tool for making justified proposals of which sequence to use.

Chapter 4

Getting a 'good one'

With a bit of a mind flip
– Riff Raff, The Rocky Horror Picture Show

(4.0.2) We have seen that it is impossible to attribute a special status of randomness to any fixed sequence $\mathbf{x} = (x_n)_{n=0}^{N-1}$ of N numbers in general. On the other hand, we are looking for good random numbers for computer simulation. Such numbers can be found only if some information about the simulation problem is available. We have presented Criterion 2.3 and 2.4 to take such information into account. Criterion 2.3 defines good sequences with respect to a restricted set[1] of simulations $\mathcal{F}$. For practical applications, the problem with restricted sets $\mathcal{F}$ is twofold.
First, those $\mathcal{F}$, for which the quality of a sequence $\mathbf{x}$ with respect to Criterion 2.3 can actually be proven, are comprised of rather primitive simulation problems (see (2.3.4)). Most of the interesting problems (think of, say, discrete event simulations) are of a more complex nature. In principle, it might be possible to find good sequences for more complex $\mathcal{F}$; unfortunately there are many technical difficulties on the way, which are very hard to surmount even for rather primitive $\mathcal{F}$.
Second, even if we can prove the quality of a sequence $\mathbf{x}$ with respect to Criterion 2.3 and $\mathcal{F}$, the question of whether a given simulation problem actually belongs to $\mathcal{F}$ or not is mostly a matter of guessing[2].

[1] To those who have read Chapter 3: with problem sets $\mathcal{F}$ which are biased, Criterion 2.1 and 2.2 will also work.

[2] Think of, say, the problem set $\mathcal{F}$ described in (2.3.4), for which the $(t\text{-}m\text{-}s)$–nets are

In this chapter, we will apply Criterion 2.4 in some common situations. To apply this criterion, we need some information about the simulation problem just as in the application of Criterion 2.3. The advantage of Criterion 2.4 is that it requires much less information and that the assumptions on the simulation problem are less restrictive and mostly intuitively convincing. But nothing comes for free, of course. The assessments of quality based on Criterion 2.4 are much weaker than those based on Criterion 2.3. Whereas the latter proposes sequences which are good 'with certainty', the proposals based on the former are good only 'in the mean'.
We present four examples of applying Criterion 2.4. The aim is not to find completely new arguments for a sequence's quality, but to use existing arguments found in the literature, and to point out under which assumptions they are reasonable. You may view these examples as plug-ins, saying:

> If the user can supply us with this-or-that information about his simulation problem, then we advise him to use this-or-that sequence.

The point is that the advice can actually be *derived* from the given information. Thus the above if-then-argument should better read:

> If the user is willing to make this-or-that assumption about his simulation problem, then these assumptions *imply* the preference of this-or-that sequence.

4.1 Pre-testing

(4.1.1) A sequence $\mathbf{x}$ is often believed to be good if it behaves well in certain statistical tests. We have already encountered this attitude in (2.2.1). If a sequence passes, say, ten statistical tests, you cannot be sure it will be a good choice for your simulation, but usually the ten tests it passed will increase your confidence in the sequence.

Suppose a user wants to simulate the flow of aerosol particles in the human lung. More precisely, he is interested in how particles are deposited in

proven to be good. $\mathcal{F}$ is made up of functions $F = 1/N \sum_{n=1}^{N} f$, where f is a function with in some sense rapidly decreasing Walsh coefficients. The rate of decrease of a given function's Walsh coefficients is not easy to estimate let alone to compute!

bronchial airway bifurcations. The model is roughly described as follows[3]. The positions of a number of particles are randomly selected at the inlet of the bifurcation. The number of particles is chosen large enough so that the resulting deposition patterns can be considered significant. Each particle enters the bifurcation at the selected position in a stream of inhaled air. Its trajectory within the bifurcation is governed by the rules of intertial impaction, interception, gravitational setting, and Brownian motion. The particle either leaves the bifurcation at one of its outlets or is deposited somewhere within. This model is used with varying bifurcation geometries, varying flow profiles at the inlet boundary, and varying particle sizes.
Note that this is a stochastic model. Apart from randomly selecting particle positions at the inlet, the particle trajectories have a stochastic element. The first three rules governing each trajectory are obeyed by solving the corresponding differential equations numerically. To obey the fourth rule in a computer simulation, random numbers are used. The Brownian motion of a particle is modeled by incrementally changing its position by some three-dimensional random vector Δ after a fixed time interval. Independent realizations $\delta_0, \delta_1, \ldots$ of the random vector Δ are obtained from a sequence $\mathbf{x}$ of random numbers and an algorithm to compute the δ_n given $\mathbf{x}$.
Now think of a statistical test which checks if the $\delta_0, \delta_1, \ldots$ behave as independent realizations of Δ should, i.e. a test whether the realizations are correctly distributed. Learning $\mathbf{x}$ behaves well in this test might well increase the user's confidence in the sequence.

We will see that this increase of confidence is not only convincing on an intuitive basis, but can in fact be derived using Criterion 2.4 from a moderate assumption about the test's *relevance* for the simulation.

(4.1.2) Let $\mathcal{G}$ be a set of simulations, let $\mathcal{H}$ be a set of statistical tests with level of significance α $(0 < \alpha < 1/2)$, let $\mathcal{F} := \mathcal{G} \times \mathcal{H}$, and let $(\mathcal{F}, \mathcal{R}, \mu)$ be a probability space.

To fix ideas, think of $\mathcal{G}$ being the set of all simulations of a model of particle-flow in bronchial bifurcations with varying bifurcation geometries, flow profiles, and particle sizes. Or think of $\mathcal{G}$ containing many different models for the particle-flow, all of which are mathematical images of the same real-world phenomenon. The set $\mathcal{H}$ might be thought of containing all statistical tests based on the random jumps $\delta_0, \delta_1, \ldots$ computed from a

[3] For a more complete description of the model and various simulation results, see Balásházy and Hofmann [4].

sequence of random numbers or just one statistical test with different sample sizes, different levels of significance, and so on.

For a pair $(F, T) \in \mathcal{F}$ and a sequence $\mathbf{x}$,

$$\Psi_{\mathbf{x}}(F) := |F(\mathbf{x}) - E(F)|$$

measures the quality of $\mathbf{x}$ in simulating F and

$$\Phi_{\mathbf{x}}(T) := |T(\mathbf{x}) - E(T)|$$

measures the behavior of $\mathbf{x}$ in the test T[4]. Additionally, we assume that the σ-algebra $\mathcal{R}$ on $\mathcal{F}$ is such that these quantities are real-valued random variables (We have already encountered this setup in (2.3.7)).
It is of course not easy to describe the mathematical object $(\mathcal{F}, \mathcal{R}, \mu)$ explicitly, especially for a user who just wants to simulate the flow of particles in the bronchiae. Fortunately, we do not need a complete, explicit description but just one property of μ.

Suppose the user assumes that the random quantities $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ are *positively correlated.* The positive correlation can be thought of expressing that the tests in $\mathcal{H}$ check for certain properties which are *relevant* for the simulations in $\mathcal{G}$. We have the following result.

Whenever the positive correlation of $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ holds, we have

$$E_\mu(\Psi_{\mathbf{x}}|\Phi_{\mathbf{x}} < 1/2) \; < \; E_\mu(\Psi_{\mathbf{x}}) \; < \; E_\mu(\Psi_{\mathbf{x}}|\Phi_{\mathbf{x}} \geq 1/2).$$

Thus, if $\mathbf{x}$ passes a test randomly selected from $\mathcal{H}$, the user *has to be* more optimistic about the error $\mathbf{x}$ is expected to produce in simulating a problem randomly selected from $\mathcal{G}$.

Proof: Recall that the positive correlation of $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ means that their covariance is positive:

$$E_\mu(\Psi_{\mathbf{x}}\Phi_{\mathbf{x}}) - E_\mu(\Psi_{\mathbf{x}})E_\mu(\Phi_{\mathbf{x}}) > 0. \tag{1}$$

Now let $\Upsilon_{\mathbf{x}}$ be the event of $\mathbf{x}$ passing a test from $\mathcal{H}$.

$$\begin{aligned} \Upsilon_{\mathbf{x}}(T) \;\; &:= \;\; \begin{cases} 1 & \text{if } \Phi_{\mathbf{x}}(T) < 1/2, \\ 0 & \text{otherwise} \end{cases} \\ &= \;\; \frac{1 - \alpha - \Phi_{\mathbf{x}}(T)}{1 - 2\alpha}. \end{aligned} \tag{2}$$

[4] Recall our observation from (2.3.3): if T is a test with level of significance $\alpha < 1/2$, then $\mathbf{x}$ passes the test if and only if $\Phi_{\mathbf{x}}(T) < 1/2$.

To verify (2), recall that $\Phi_{\mathbf{x}}(T) = \alpha < 1/2$ if and only if $T(\mathbf{x}) = 0$.

Since $\Psi_{\mathbf{x}}$ is positively correlated to $\Phi_{\mathbf{x}}$, it is negatively correlated to $\Upsilon_{\mathbf{x}}$. To derive this formally, simply use (2) to express $\Phi_{\mathbf{x}}$ by means of $\Upsilon_{\mathbf{x}}$, and substitute this in (1); this yields

$$E_\mu(\Psi_{\mathbf{x}}\Upsilon_{\mathbf{x}}) - E_\mu(\Psi_{\mathbf{x}})E_\mu(\Upsilon_{\mathbf{x}}) < 0.$$

This is in fact equivalent to the positive correlation of $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ since $\Upsilon_{\mathbf{x}}$ was obtained from $\Phi_{\mathbf{x}}$ by a linear transformation.
Next, observe that for the above negative correlation to hold, the terms $E_\mu(\Psi_{\mathbf{x}})$ and $E_\mu(\Upsilon_{\mathbf{x}})$ both must be positive[5]. With this, we can transform the above equation to

$$\frac{E_\mu(\Psi_{\mathbf{x}}\Upsilon_{\mathbf{x}})}{E_\mu(\Psi_{\mathbf{x}})E_\mu(\Upsilon_{\mathbf{x}})} < 1. \tag{3}$$

Now we have

$$\begin{aligned} E_\mu(\Psi_{\mathbf{x}}\Upsilon_{\mathbf{x}}) &= E_\mu(\Psi_{\mathbf{x}}\Upsilon_{\mathbf{x}}|\Upsilon_{\mathbf{x}} = 1)P_\mu(\Upsilon_{\mathbf{x}} = 1) + \\ &\quad E_\mu(\Psi_{\mathbf{x}}\Upsilon_{\mathbf{x}}|\Upsilon_{\mathbf{x}} = 0)P_\mu(\Upsilon_{\mathbf{x}} = 0) \\ &= E_\mu(\Psi_{\mathbf{x}}|\Upsilon_{\mathbf{x}} = 1)P_\mu(\Upsilon_{\mathbf{x}} = 1). \end{aligned} \tag{4}$$

Employing the fact that $E_\mu(\Upsilon_{\mathbf{x}}) = P_\mu(\Upsilon_{\mathbf{x}} = 1)$, we can substitute (4) in (3) to get

$$\frac{E_\mu(\Psi_{\mathbf{x}}|\Upsilon_{\mathbf{x}} = 1)}{E_\mu(\Psi_{\mathbf{x}})} < 1$$

which is equivalent to

$$E_\mu(\Psi_{\mathbf{x}}|\Upsilon_{\mathbf{x}} = 1) < E_\mu(\Psi_{\mathbf{x}}).$$

The proof of the second inequality is analogous, now with the auxiliary event $\Upsilon_{\mathbf{x}} := 1$ if $\mathbf{x}$ fails the test and 0 otherwise. □

(4.1.3) A similar argument can be applied to choose between two sequences $\mathbf{x}$ and $\mathbf{y}$ when the user cannot decide which one he should prefer. If he considers the quantities $\Psi_{\mathbf{z}}$ and $\Phi_{\mathbf{z}}$ as negatively correlated for any sequence $\mathbf{z}$, we advise him to use that sequence which passes the test (if one

[5] All involved random variables are nonnegative and, therefore, their expectations are also nonnegative. Hence $E_\mu(\Upsilon_{\mathbf{x}})$ and $E_\mu(\Psi_{\mathbf{x}})$ must both be positive, since otherwise the inequality of negative correlation cannot hold.

of them fails).
This is because his indifference about which sequence to use means

$$E_\mu(\Psi_{\mathbf{x}}) = E_\mu(\Psi_{\mathbf{y}}).$$

Assuming that **y** fails the test and **x** passes it, we get

$$E_\mu\left(\Psi_{\mathbf{x}}|\Phi_{\mathbf{x}} < 1/2\right) < E_\mu\left(\Psi_{\mathbf{y}}|\Phi_{\mathbf{y}} \geq 1/2\right).$$

Moreover, even if the test is evaluated only for one of the samples, a reasonable decision is possible. If, say, we only know that $\Psi_{\mathbf{x}} < 1/2$, we get

$$E_\mu\left(\Psi_{\mathbf{x}}|\Phi_{\mathbf{x}} < 1/2\right) < E_\mu(\Psi_{\mathbf{y}}).$$

(4.1.4) There is of course no need to restrict the pre–testing to sets $\mathcal{H}$ of statistical tests only; any set of functions whose evaluation amounts to answering a yes-no question, functions which assume only two possible values, can be used. If, say, $\mathcal{H}$ is a set of functions T which can only take the values $T(\mathbf{x}) = 0$ or $T(\mathbf{x}) = 1$ (where $T(\mathbf{x}) = 0$ is interpreted as **x** passing the 'test' T) and if positive correlation between the corresponding random variables $\Psi_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$ is assumend, then pre–testing is applicable.

(4.1.5) So this is the justification for applying statistical tests to random number sequences.

> If the user considers a test as relevant for his simulation, his confidence in a sequence necessarily increases if it passes the test.

This argument is especially useful since we can apply a battery of statistical tests to a sequence *before* the user actually starts looking for a good one. The selection of these tests is of course more or less arbitrary, but this does not matter as long as the user can find a test which checks for a property he considers relevant.

The first simple tests for random numbers were designed "in the early days of making tables of random digits" [80, p.6] in the work of Kendall and Babington-Smith [58]. These and a few more tests suggested by MacLaren and Marsaglia in [76] were compiled in Knuth's [59, Section 3.3]. An application of (slightly modified) tests from Knuth's battery to specific sequences was performed by L'Ecuyer in [68].

Marsaglia [80] designed a battery of what he calls "*stringent* tests, because they seem more difficult to pass than the mild tests that have become standard." He gives a description of his tests and the results of applying them to various random number sequences.

The most extensively applied test battery is due to Fishman and Moore [38]. In addition to presenting individual tests, they also present an "omnibus test" which is the conjunction of the individual tests. In [39], they apply their tests to a whole class of random number generators: out of over 534 million candidate generators, they select the 414 'best' with respect to some reasonable criterion[6], subject them to their tests, and present the results and the five best generators. The same procedure is applied by Fishman in [37] to the whole of another class of random number generators and to parts of yet a third.

(4.1.6) However, the utility of test batteries is restricted by the requirement that the tests must check for properties which the user considers relevant for his simulation. If he cannot find a test which checks for relevant properties, probably the best advice for using the argument presented in this section is given by Marsaglia [80, p.6]: "If the RNG [random number generator] is to be used for a particular problem, one should try to create a test based on a similar problem for which the underlying distributions are known ...". With the considerations of (4.1.4), we add that *any* test will do, statistical or not, if only it is relevant.

4.2 Employing known defects

(4.2.1) By a *defect* of a sequence $\mathbf{x}$, we understand a simulation problem F which $\mathbf{x}$ approximates 'very badly'. In our terminology, this is to say that

$$\Psi_{\mathbf{x}}(F) \geq \epsilon$$

for some error bound ϵ determining the defect's severeness.

[6] Their selection is based on a normalized version c/ν_s of the spectral test $1/\nu_s$. We introduce the spectral test informally in (4.2.4) and formally in Section B.3. c depends on the generator and the tested dimension such that results of different generators are comparable.

We have already seen that any two samples $\mathbf{x}$ and $\mathbf{y}$ are 'good' in approximating about the same number of simulation problems[7]. It is clear that the same holds for the number of simulation problems in which $\mathbf{x}$ and $\mathbf{y}$ are 'bad'[8]. Thus:

> Any sample is as defective as any other.

We know $\mathbf{x} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and $\mathbf{y} = (1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0)$ have the same number of defects, so – in general – $\mathbf{x}$ is 'as good as' $\mathbf{y}$. However, if we are going to simulate the average number of 'heads' in 12 tosses of a fair coin (taking 1 as 'head' and 0 as 'tail'), we will tend to prefer $\mathbf{y}$.
Although this preference for $\mathbf{y}$ is quite obvious, it exemplifies one thing: the crucial point about a sequence's defects is not their *presence* but their *relevance* for the simulation problem at hand. The more we know about the defects of $\mathbf{x}$, the better we can decide for which simulations $\mathbf{x}$ should not be used. If no defect of $\mathbf{x}$ is explicitly known, we may – simulating a given problem F – stumble right over one of its worst deficiencies. In this sense, a sound knowledge of the defects of $\mathbf{x}$ holds in favor of the sequence.

(4.2.2) Concerning known defects, the study of deterministic sequences of random numbers, produced by so-called *random number generators*[9], can be very valuable. In such a sequence $\mathbf{x} = (x_n)_{n=0}^{N-1}$, each number x_n is obtained by an explicit computational rule which – for virtually all random number generators in use today – is some sort of recursion, say,

$$\begin{aligned} x_0 &:= 0, \\ x_n &:= f(x_{n-1}) \qquad (1 \le i < N). \end{aligned}$$

[7] We saw this with respect to statistical tests and finite binary sequences in Chapter 2 and with respect to arbitrary unbiased test sets and finite sequences in $[0, 1[$ in Chapter 3.

[8] To those who have read Chapter 3: for a problem set $\mathcal{F}$, those functions in which a sample $\mathbf{x}$ is defective with respect to ϵ form just the set $\mathcal{F} \setminus \mathcal{F}_{\mathbf{x}}$. Since the sets $\mathcal{F}_{\mathbf{x}}$ and $\mathcal{F}_{\mathbf{y}}$ are 'of the same size' for almost all $\mathbf{x}$ and $\mathbf{y}$ whenever $\mathcal{F}$ is unbiased and non-trivial, (Proposition 3.1 and 3.2), the same holds for their complements.

[9] If the finite sequence $\mathbf{x}$ is generated by a deterministic algorithm, some authors refer to $\mathbf{x}$ as a sequence of pseudo-random numbers and to the algorithm as a pseudo-random number generator. In accordance with Definition 2.1, and because we have seen that there is no probabilistic way to separate the finite sequences of 'truly'-random numbers from those of pseudo-random numbers, we omit the adjective 'pseudo' in this text. Any finite sequence $\mathbf{x}$ of numbers is a sequence of random numbers if it is used as a substitute for random variables in a simulation. If $\mathbf{x}$ is generated by a deterministic algorithm, we refer to this algorithm as a *random number generator*.

The output $\mathbf{x}$ of a random number generator has one important advantage over a nondeterministic $\mathbf{y}$. $\mathbf{x}$ is completely determined by f and x_0; any defect of $\mathbf{x}$ can – in principle – be *derived* from f and x_0. On the other hand, the defects of $\mathbf{y}$ depend on all the $y_0, \ldots, y_{N-1}$; if there is no mathematical structure in the numbers y_n, its defects cannot be derived but just *observed*[10]. As Knuth notes in [59, p.75]: "Although it is always possible to test a random number generator using the [statistical] methods in the previous section, it is far better to have 'a priori tests,' i.e., theoretical results that tell us in advance how well those tests will come out. Such theoretical results give us much more understanding about the generation methods than empirical, 'trial-and-error' results do."

Although it is not always easy to infer structural deficiencies of a random number generator's output from its algorithm, it can be done in some cases.

> For some random number generators, we are able to *prove* they are defective for a whole class of simulation problems, so we know in advance that they will fail the corresponding tests.
> Conversely, for some other random number generators, we are able to prove that they are not defective for a whole class of problems; they are known in advance to pass the corresponding tests.

In the rest of this section, we describe these generators and their defects along with how they can be used in conjunction with Criterion 2.4. For the sake of simplicity, this section is rather informal, focusing on the defects' consequences rather than on their cause. The precise and formal treatise is postponed to Chapter 5.

(4.2.3) Today, the most widely used generator is in fact the one with

[10] There is yet another important advantage of using a random number generator to produce $\mathbf{x}$. Since its output is repeatable, so is the computed simulation result $F(\mathbf{x})$. The ability to obtain reproducible results is appreciated in scientific work in general and in computing in particular, as noted by Ripley in [95, p.153]: "For example, Ripley and Kirkland [96] [...] show some summaries of the simulation of a Markov random field which show an abrupt change at one point in the supposedly converging iterative process. Because a repeatable sequence was used, the process could be run up to just before that point and stopped, so the critical phase could be examined in detail." A similar statement is given by Halton in [47, p.2]. Without the ability to repeat the sequence $\mathbf{x}$, one is unable to decide whether a given simulation result $F(\mathbf{x})$ is valid or simply due to some programming error!

the apparently most serious defect: the linear congruential generator[11] and its lattice structure.
A linear congruential generator (LCG, for short) with integer parameters M, a, b, and y_0 computes its values by the recursion

$$\begin{aligned} x_0 &:= \frac{y_0}{M}, \\ x_n &:= a x_{n-1} + \frac{b}{M} \pmod{1} \qquad (0 < n < N), \end{aligned}$$

where the period N of the generator depends on M, a, and b. We refer the reader to Definition 5.1 for an exact definition and to (5.2.2) for conditions to achieve maximal period length.

As can be seen from the definition of the x_n, the LCG's recursion is defined by a straight line which is wrapped around the unit square by the modulo-operation; for an LCG with $M = 64$, $a = 5$, and $b = 21$, the recursion f has the following form[12]:

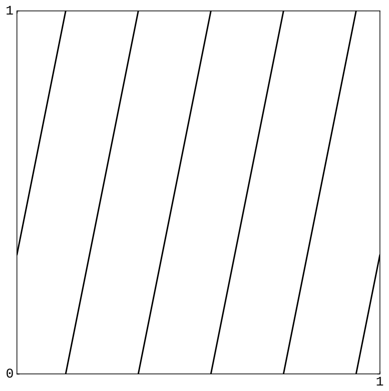


Figure 1

If we plot the points $(x_0, x_1), (x_1, x_2), \ldots, (x_{N-2}, x_{N-1}), (x_{N-1}, x_0)$ produced by this LCG, they lie on this wrapped line:

[11] See (1.0.5) for who proposed this generator.

[12] If we used another LCG with larger values of M or a, the basic appearance of f would be the same but it would have significantly more (and maybe steeper) linear branches which are hard to plot.

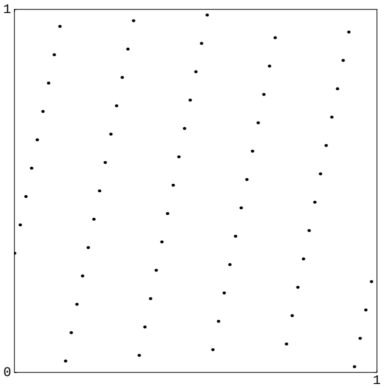


Figure 2

Linear patterns like this were known for quite a time[13] without being recognized as an inherent flaw of this kind of generator. Even eighteen years after the first proposal of the LCG, Chambers [10] argued: "With [linear congruential] generators having sufficiently long periods such patterns are no longer evident."[14]

The first proof that linear patterns are inherent to the LCG was given by Marsaglia [78] in 1968:

> "if n-tuples $(u_1, u_2, \ldots, u_n)$, $(u_2, u_3, \ldots, u_{n+1}), \ldots$ of uniform variates produced by the [linear congruential] generator are viewed as points in the unit cube of n dimensions, then *all* the points will be found to lie in a relatively small number of parallel hyperplanes."

This structure of overlapping tuples[15] occurs with any LCG in any dimension. Moreover, the points do not only fall into quite a few parallel hyperplanes, but also exhibit a regular behavior within these. As Beyer et al. [5] and Smith [104] showed in 1971, the s-dimensional points produced by an LCG form a (shifted) *lattice*[16] in $[0,1[^s$ (Intermediate cases, where the points are merely contained in a (shifted) lattice, or form or are contained

[13] See, for example, Greenberger [43, 44].

[14] There is of course reason in Chambers' observation. If the period is large and if we plot only a small fraction of all the points, then indeed the linear patterns are no longer evident; see Figure 4b.

[15] See also Marsaglias papers [79, ?].

[16] See Figure 2 to take a look at the lattice for $s = 2$.

in the union of several (shifted) lattices exist; see [1, Section 3.3]. Handling these intermediate cases would add some technical difficulties but not more clarity. Therefore, we only consider the standard case of points forming a (shifted) lattice in $[0,1[^s$).

Without going into detail, it is clear that this is definitely not the expected behavior of the random variables which the LCG should approximate. Thus: let T be any test which checks a sequence $\mathbf{z}$ of random numbers for the presence of lattice or hyperplane structures in some dimension $s > 1$. We know than any $\mathbf{x}$ produced by an LCG will fail T. If the user considers T as *relevant* for his simulation problem, Criterion 2.4 can be applied just as in Section 4.1, yielding the result that $\mathbf{x}$ should not be used.

(4.2.4) The lattice structure of the LCG enables us not only to avoid it in certain circumstances but also to select a 'good' LCG, i.e. parameters M, a, and b such that the corresponding LCG is 'better' than others. If you look at the lattice produced by the following two LCG[17],

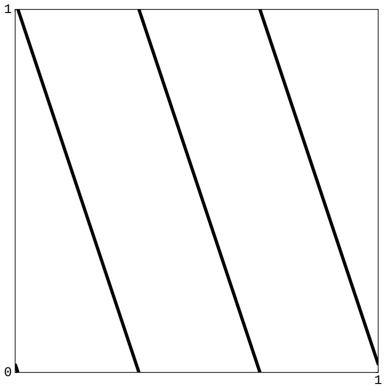


Figure 3a

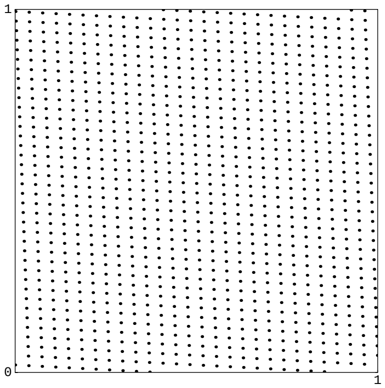


Figure 3b

you are likely to prefer the second generator. Both of them produce a regular pattern, but the first generator seems to be much worse.

Marsaglia's work in 1968 [78] has triggered a series of papers on how to describe the LCG's lattice in higher dimension. Several figures of merit

[17] Figure 3a was produced by an LCG with $M = 1024$, $a = 1021$, and $b = 21$; Figure 3b was produced by an LCG with the same parameters except for a, which was set to 997. The points in Figure 3a are packed so closely on just four line-segments that they cannot be perceived as individual points any more.

have been developed to facilitate finding an LCG whose points fall on more parallel hyperplanes, whose parallel hyperplanes are packed more closely, or whose lattice's unit cell is more cube-like. Considering the generators shown in Figure 3, think of a (not necessarily statistical) test T which 'rejects' a sample $\mathbf{z}$ if the two dimensional points of $\mathbf{z}$ are covered by 10 or less parallel lines and 'accepts' $\mathbf{z}$ otherwise. If we consider T as relevant for whatever problem we are going to simulate, we have reason to prefer the generator in Figure 3b to the one in Figure 3a.
To consider a test T as 'relevant', one has to understand the defects T can detect and the properties of randomness T tests for. The better these properties are understood, the better one can judge whether or not they are relevant for a given simulation problem. The following figures of merit are quite easy to understand.

- The maximal distance $1/\nu_s$ of parallel hyperplanes which cover the LCG's lattice in $\mathbf{R}^s$; this is called the spectral test.
  One interpretation of this is quite obvious: the larger $1/\nu_s$, the further are the LCG's hyperplanes apart, and so the larger is the parallelepiped in $[0,1[^s$ which is *never* hit by one of the points. If the simulation $F(\mathbf{x})$ depends on the distribution of the s-dimensional points produced by the sequence $\mathbf{x}$ and if the LCG $\mathbf{x}$ has a large value of $1/\nu_s$, then $F(\mathbf{x})$ will ignore large areas of $[0,1[^s$. If those areas happen to be important for F, the simulation $F(\mathbf{x})$ will ignore important aspects of the simulation problem F.
  A quite but not entirely unlike interpretation is given by Knuth in [59, p.90]: "If we take truly random numbers between 0 and 1, and round or truncate them to finite accuracy so that each is an integer multiple of $1/\nu$ for some given number ν, then the t-dimensional points [...] we obtain will have an extremely regular character when viewed through a microscope. [...] We shall call ν_2 the two-dimensional *accuracy* of the random number generator, since the pairs of successive numbers have a fine structure that is essentially good to one part in ν_2."
  See (5.2.5) and Section B.3 for a formal specification of $1/\nu_s$ and for how to compute this quantity.

  It is interesting to note that Coveyou and MacPherson introduced the spectral test in 1967 [14] without actually employing the s-dimensional lattice structure. "Instead of working with the grid structure of successive points, they considered random number generators as sources of t-dimensional 'waves'. The [inverse of the distance of parallel adjacent hyperplanes covering the lattice] [...] in their original treatment were the wave 'frequencies', or points in the 'spectrum' defined by the random number generator, with low-frequency waves being the most damaging to randomness; hence the name *spectral test*" (from [59, p.110])[18].

The approach of Coveyou and MacPherson suggests that the spectral test might be used as a general figure of merit for rating sequences of random numbers. But when the sequence $\mathbf{x}$ has no lattice structure, Niederreiter observes in [87, p.168] that the "difficulty here is to find a convincing quantitative formulation of this idea." An alternative approach which conserves the basic concept of Coveyou and MacPherson and which is applicable for *any* sequence $\mathbf{x}$ is presented by Hellekalek in [50]: the diaphony (due to Zinterhof [117]; see also the corresponding footnote on page 62) can be viewed as a *weighted spectral test*. While the spectral test computes the lowest wave frequency, the diaphony computes a weighted sum of all wave frequencies where lower frequencies are more emphasized.

- The relation q_s of the shortest to the longest edge of the unit cell of the lattice in $[0,1[^s$, called the Beyer-quotient.
  Due to Beyer [5], the Beyer-quotient gives a good description of the unit cell of the lattice. If we look at Figure 3, we tend to prefer the generator for which q_2 is closer to 1.
  One interpretation of q_s is given by L'Ecuyer in [69, p.90]: "A unit cell of the lattice is determined by the vectors of a *Minkowski-reduced lattice base* (MRLB) [...]. It is traditionally accepted that 'better' generators are obtained when the unit cells of the lattice are more 'cubic-like' (i.e. when the vectors of the MRLB have about the same size). The ratio q_t of the sizes of the shortest and the longest vectors of a MRLB is called the *Beyer-quotient*."
  Another interpretation can be derived from Knuth's concept of s-dimensional accuracy. Suppose we take "truly random" points in $[0,1[^s$ and represent them with coordinates relative to the MRLB. If we round or truncate each coordinate to an integer value (performing some wraparound at the unit cube's boundaries), the resulting rounded points

[18] This 'indirect' introduction of $1/\nu_s$ is probably one of the reasons why this quantity was misunderstood for quite a long time. See, for example, [?, p.278] for an unquotable statement on the spectral test.

will lie just on the lattice spanned by the MRLB. A Beyer-quotient q_s close to 1 means that we loose about the same amount of accuracy in each coordinate, while a small value of q_s means that one coordinate looses significantly more accuracy than some other.
We refer to (5.2.5) and Section B.2 for a more complete treatise of q_s and Minkowski-reduced lattice bases and for how to compute these numerical quantities.

Some more figures of merit describing the s-dimensional lattice of an LCG have been proposed, but we do not present them here since we could not find reasonable interpretations by which the user can judge their relevance for his simulation problem. We refer the interested reader to [28], [78], [**?**], [86], or [92].

(4.2.5) Whenever lattice tests or tests for the presence of hyperplane structures are considered relevant, we can not only advise the user not to use an LCG but also which kind of generator to use instead: an *inversive generator.* In fact, there are two kinds of inversive generators, the inversive congruential generator (ICG) and the explicit inversive congruential generator (EICG).

The ICG, proposed in 1986 by Eichenauer and Lehn [31], computes its numbers by means of a recursion just as the LCG, but its recursion is based on a nonlinear function. For a prime number p and suitable integer parameters a, b, and y_0, the ICG computes the sequence $\mathbf{x} = (x_n)_{n=0}^{p-1}$ as

$$\begin{array}{rcll} x_0 & := & \frac{y_0}{p}, & \\ x_n & := & f(p,a,b,x_{n-1}) & \quad (0 < n < p), \end{array}$$

where $f(p,a,b,x)$ is a 'highly' nonlinear function in x. We give a more complete description of the ICG in Section 5.3.
The EICG, proposed in 1993 by Eichenauer-Hermann [34], uses no recursion but, given the desired index n, facilitates the explicit computation of x_n. Just as the ICG, it uses a prime number p and integer parameters a, b, and n_0 to compute $\mathbf{x} = (x_n)_{n=0}^{p-1}$ as

$$x_n \quad := \quad g(p,a,b,n_0,n) \qquad (0 \le n < p).$$

A more complete description of this generator can be found in Section 5.4.

The output of inversive generators has a property which renders them *the alternative* to the LCG whenever hyperplane defects are considered relevant[19].

> The s-dimensional points produced by an inversive generator avoid the hyperplanes in $\mathbf{R}^s$.

So with respect to hyperplane structures, inversive generators behave exactly as random variables are expected to.
Since points avoiding the hyperplanes cannot form a lattice, the ICG and the EICG are proven to pass exactly those tests which the LCG fails!

(4.2.6) Considering the undesirable lattice structure of the LCG and the lack of this structure in inversive generators, one might argue as Marsaglia does in [**?**, p.250]: "The conclusion is that [linear] congruential random number generators are not suitable for precision Monte Carlo use." This statement is not completely valid. Of course, inversive generators perform better than the LCG – if lattice or hyperplane structures are concerned. Without this 'if', there is no reason to reject the LCG. Knuth warns in [59, p.90] not to overestimate the LCG's lattice structure: "At first glance we might think that such a systematic behavior is so nonrandom as to make [linear] congruential generators quite worthless; but more careful reflection, remembering that [the period] m is quite large in practice, provides a better insight." If a simulation problem consumes only a small fraction of a generator's output using too few points for the LCG's lattice to show up, it is rather hard to judge if an LCG or an inversive generator should be used[20]:

[19] For a precise statement of this result, see (5.3.5) for the ICG and (5.4.5) for the EICG.

[20] Figure 4a is a plot of the points $\{(x_n, x_{n+1}) : 0 \leq i < 1000\}$ obtained from the EICG with $M = 2^{31} - 1$, $a = 1$, $b = 0$, and $y_0 = 0$; Figure 4b is a plot of the same set of points obtained from the LCG with $M = 2^{31} - 1$, $a = 742938285$, $b = 0$, and $y_0 = 1$, which Fishman and Moore analyzed in [39].

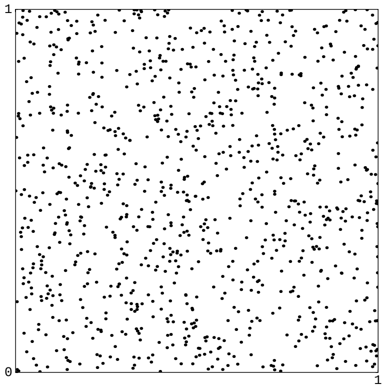


Figure 4a

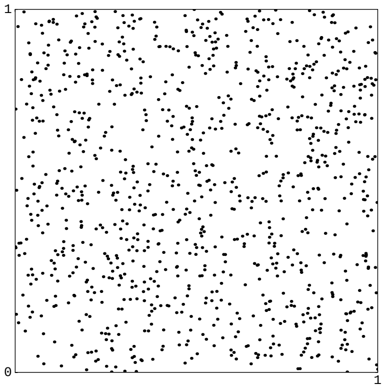


Figure 4b

Even in their first proposal of the ICG in [31, p.322], Eichenauer and Lehn stress that "one cannot recommend the application of nonlinear congruential generators [...] instead of linear congruential generators [...] in general. But they should be applied if one has the feeling that there is something wrong with the simulation results and one suspects that this is caused by the lattice structure of the linear congruential generator." Careful examination of the simulation problem is necessary to judge whether a lattice structure can be considered relevant or not.

There are simulation problems whose examination does in fact indicate that lattice structures might be highly relevant. Ripley [92] considers the minimal distance of k independent points in $[0,1[^2$. If such points are simulated by an LCG, they lie on a regular lattice. The shortest possible distance of two lattice points is just the length $||\mathbf{m}||$ of the shortest side of the lattice's unit cell. Although the minimal distance is expected to decrease as k increases, an LCG-based simulation will *never* yield minimal distances below $||\mathbf{m}||$.
Simulations of this minimal distance, where an ICG scores significantly better than an LCG, were performed by Eichenauer and Lehn in [31]. However, these simulations have to be taken with a grain of salt. The involved generators have a very short period of less than 300000, while most real-world applications require period lengths of about at least 2^{31}. Although it is, in principle, possible to repeat the experiment in [31] for generators with larger period lengths, sampling more points for the LCG's lattice to show up, the computational complexity of the test-statistic – which is $O(\text{sample size}^2)$ – hindered us from doing so. In short, we could not produce results as devastating as those in [31] for LCGs with large period lengths.

Simulations where inversive generators are found to be always at least as good as and sometimes significantly better than even the best LCG proposed in [39] (whose period is $2^{31}-2$) have been performed by Entacher [35, 36], Leeb [70, 71, 72], and Wegenkittl [112].

(4.2.7) A particularly serious defect common to many random number generators is the presence of a special kind of *long-range correlations.* If we use a sequence $\mathbf{x} = (x_n)_{n=0}^{N-1}$ to form the points (x_n, x_{n+s}) for a fixed shift s, these points should be randomly scattered in the unit square; in particular, the numbers x_n and x_{n+s} should be empirically uncorrelated[21]. For a variety of generators, the converse is true.

> For specific integers s called critical distances, the points (x_n, x_{n+s}) concentrate on a few parallel lines in the unit square.

The presence of this kind of long-range correlation in a sequence $\mathbf{r} = (r_n)_{n=0}^{N-1}$ can be devastating in a simulation which uses more than s random numbers. As De Matteis and Pagnutti note in [21, p.67]: "In any case, the event simulated by means of r_n will be correlated with that simulated by r_{n+s} for every n and wherever one starts in the sequence, i.e. each event will keep a memory of what happened s events before."

Just as the LCG's lattice structure, this gives reason not to use generators which exhibit critical distances – if this defect is considered relevant. Again, this 'if' is important. De Matteis and Pagnutti remark on the long-range correlations in [21, p.6]: "How harmful this may be depends on the particular application and also on the quantity of numbers required." The user is responsible to analyze his simulation problem and to evaluate the relevance of the long-range correlation defect.

The first indication of this defect was noticed in the LCG in the early '60s by Coveyou [13], Greenberger [43], and Peach [90]. It is interesting how many years passed until it was analyzed to its full extent. Various long-range correlations were of course observed since then (like, say, by De Matteis and Faleschini [17] in 1963, Dieter [24] in 1968, Ahrens, Dieter and Grube [2] in 1970, Dieter and Ahrens [27] in 1971, Neuman and Merrick [83] in 1976, Neuman and Martin [82] in 1976, Holmild and Rynefors [56] in 1978, and

[21] $\mathbf{x}$ is used to model the sequence $\mathbf{X} = (X_n)_{n=0}^{N-1}$ of independent random variables. Since X_n and X_{n+s} are stochastically independent, they are uncorrelated and so should be x_n and x_{n+s}.

Hill [53] in 1979), but even in 1987 Bowman and Robinson [7] noticed the defect merely by "Detailed examination of examples ...". Finally, in 1988, De Matteis and Pagnutti *proved* the presence of critical distances in a class of LCGs for specific values of s in [18] (see also [19]), a result which was generalized for arbitrary s in 1989 by Eichenauer-Hermann and Grothe in [33]. One may suspect this defect to be a direct consequence of the simple linear recursion of the LCG and that using more complicated or nonlinear recursions might be a remedy. Unfortunately, long-range correlations were observed also in Tausworthe generators[22] by Neuman and Martin in [82], and the existence of critical distances s in Wichmann-Hill generators[23] was shown by De Matteis and Pagnutti in [21]. Moreover, in [20], the above authors gave a quite restrictive but nevertheless sufficient condition under which *any* generator with a recursion of order one, be it linear or nonlinear, has this defect.
We show the existence of critical distances in the LCG in (5.2.6) and for any congruential generator forming a lattice in higher dimensions in Section B.4.

(4.2.8) For sequences which exhibit critical distances, these can be used to select one whose long-range correlation between the n-th and the $(n+s)$-th number takes shape later, for larger values of s. To be specific, let $\mathbf{x} = (x_n)_{n=0}^{N-1}$ and $\mathbf{y} = (y_n)_{n=0}^{N-1}$ be sequences which exhibit critical distances. For some integer l, say, $l = 1000$, we define[24]

$$s^\star(\mathbf{x}) := \min\{s : \text{ the } (x_n, x_{n+s}) \text{ are covered by at most } l \text{ parallel lines}\}$$

for $\mathbf{x}$ and $s^\star(\mathbf{y})$ as the analogous number for $\mathbf{y}$ (note that both sequences are finite so the minima are well-defined). If we find $s^\star(\mathbf{x}) > s^\star(\mathbf{y})$, we know that the specific long-range correlation in $\mathbf{x}$ forms later than in $\mathbf{y}$. If we consider this concentration on at most l parallel lines as relevant, we can apply Criterion 2.4 to prefer $\mathbf{x}$ since it will pass a corresponding test which $\mathbf{y}$ will fail.

(4.2.9) If the user considers the mere presence of critical distances

[22] Proposed by Tausworthe in [109]; see also [87, Chapter 9].

[23] The Wichmann-Hill generator, due to Wichmann and Hill [114, 115], computes its numbers by combining the output of (usually) three LCG. This technique was supposed to perturb the LCGs' simple lattice structure.

[24] An algorithm for computing the quantity $s^\star$ for LCGs with $b = 0$ is given by De Matteis, Eichenauer-Hermann, and Grothe in [16] and for Wichmann-Hill generators (composed of three LCGs whose additive constant b is zero) by De Matteis and Pagnutti in [21].

as relevant for his simulation problem[25], he should use an EICG. For this generator, the absence of critical distances s (except for the trivial one, $s = 0$) is proven by the following result.

> The points (x_n, x_{n+s}) produced by an EICG avoid the lines in the unit square.

This property, which is a consequence of Niederreiter's [85, Theorem 1], is stated more properly in (5.4.6). A similar but slightly weaker result for the ICG is given in (5.3.6).

(4.2.10) All the random number generators presented so far yield, when *all* their numbers are produced, a one-dimensional grid equal or similar to $\{1/N, \dots, N - 1/N\}$, sometimes including 0 (see (5.2.1), (5.3.2), and (5.4.2) for the exact form of this grid for the LCG, ICG, and EICG, respectively). As with the lattice produced by the LCG, it is clear that producing such excessive uniformity is not the behavior one would expect from random variables.

If a user's simulation problem will consume, say, k random numbers, and if he considers excessive uniformity as a relevant defect, he has reason not to use a generator whose period N is close to k (if the generator exhibits excessive uniformity).

It is not easy to assess which fraction of a generator's period N can be used without risking excessive uniformity. A common advice is to use no more than $\sqrt{N}$ of a total of N random numbers, but as MacLaren objects in [77, p.47]: "This rule appears to be passed primarily by word of mouth, because nobody seems to know either where it originated or a reliable reference to its reason."
MacLaren observes excessive uniformity in a number of generators in simulating a certain quantity when using more than $N^{2/3}$ numbers. However, he is cautious not to recommend this fraction as a definite safety limit: "This does not mean that all analyses will start to give incorrect results at that point, but that at least some will do so." Besides the fractions $\sqrt{N}$ and $N^{2/3}$, using no more than $N/100$ is suggested by Deák in [22, p.156], $N/1000$ by De Matteis and Pagnutti in [18, p.607], and an argument for using no more than $\sqrt{N/200}$ is given by Ripley in [93, p.26].

[25] As he may, for example, in certain forms of stochastic simulation on parallel computers; see [18] and [19].

Although we do not feel able to favor one of these fractions, we can conclude that a generator with large period length might be preferred to another whose period is significantly shorter if the user expects his simulation to require a sufficiently large amount of random numbers.
To get a feeling of how many numbers certain simulations consume today and thus how large the period of the employed generator should be, we quote Halton [47, p.3]: "In a typical conventional particle-transport calculation, using nonbranching random walks, we may compute some $10^3 - 10^8$ random walks, averaging perhaps $10^2 - 10^6$ steps each, with every step requiring around 10 random numbers; this adds up to a need for something of the order of $10^6 - 10^{15}$ random numbers."

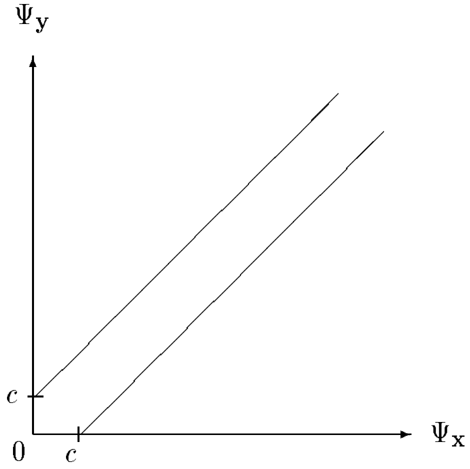
4.3 Cross-testing

(4.3.1) Donald Knuth ends his treatise on random numbers in [59, p.173] with the following advice: "The most prudent policy for a person to follow is to run each Monte Carlo program at least twice using quite different sources of random numbers, before taking the answers of the program seriously". This recommendation seems to be quite sensible and convincing. Suppose we simulate some problem F twice, using two sequences $\mathbf{x}$ and $\mathbf{y}$ of random numbers. If we find $F(\mathbf{x})$ and $F(\mathbf{y})$ to be completely different, we are in the unpleasant situation of knowing that one result is significantly better than the other ... without knowing which. Conversely, if we find $F(\mathbf{x})$ to be close to $F(\mathbf{y})$, this seems to give rise to some optimism that both simulation results are rather good. This optimism is well exemplified by Dudewicz and Ralley in [29, p.3]: "it is often desired to re-run at least part of a simulation study using a quite different generator (after which, if the two sets of results agree, one has somewhat more confidence that generator regularities were not of such a serious nature as to vitiate the simulation study)."

As long as the term 'quite different generator' is unspecified, having 'somewhat more confidence' is completely unfounded. All that can be derived from $F(\mathbf{x})$ being close to $F(\mathbf{y})$ is that

$$\begin{aligned}
|F(\mathbf{x}) - F(\mathbf{y})| &= |(F(\mathbf{x}) - E(F)) - (F(\mathbf{y}) - E(F))| \\
&\geq ||(F(\mathbf{x}) - E(F))| - |(F(\mathbf{y}) - E(F))|| \\
&= \left|\Psi_{\mathbf{x}}(F) - \Psi_{\mathbf{y}}(F)\right| . \qquad (1)
\end{aligned}$$

If $c := |F(\mathbf{x}) - F(\mathbf{y})|$ is small, this only implies that that the pair $(\Psi_{\mathbf{x}}(F), \Psi_{\mathbf{y}}(F))$, viewed as point in the plane, lies withing a diagonal stripe of width $c\sqrt{2}$:



Without further assumptions, the point $(\Psi_{\mathbf{x}}(F), \Psi_{\mathbf{y}}(F))$ can be – albeit within this stripe – arbitrarily far from the origin, so both approximation errors $\Psi_{\mathbf{x}}(F)$ and $\Psi_{\mathbf{y}}(F)$ can be arbitrarily large.

(4.3.2) As with the pre–testing presented in Section 4.1, the above optimism is based on a *hidden assumption.* The samples $\mathbf{x}$ and $\mathbf{y}$ are considered 'quite different' in the sense that they are unlikely to be both simultaneously 'bad' in a simulation. Given this assumption, we will see that the above optimism is not only well-founded but in fact necessary.

Let $\mathcal{F}$ be a set of simulation problems. For $F \in \mathcal{F}$, let

$$\Psi_{\mathbf{x}}(F) := |F(\mathbf{x}) - E(F)|$$

and let $\Psi_{\mathbf{y}}(F)$ be defined analogously. Furthermore, for a fixed $\delta > 0$, let

$$\Upsilon_{\delta}(F) := \begin{cases} 1 & \text{if } |\Psi_{\mathbf{x}}(F) - \Psi_{\mathbf{y}}(F)| \leq \delta, \\ 0 & \text{otherwise.} \end{cases}$$

The event Υ_{δ} is equal to 1 if and only if the point $(\Psi_{\mathbf{x}}(F), \Psi_{\mathbf{y}}(F))$ lies within the diagonal stripe of width $\delta\sqrt{2}$.

Let $(\mathcal{F}, \mathcal{R}, \mu)$ be a probability space such that $\Psi_{\mathbf{x}}$, $\Psi_{\mathbf{y}}$, and Υ_δ are random variables. Then we have the following result.

> Whenever the random variables $\Psi_{\mathbf{x}}$ and Υ_δ are *negatively correlated*, then
>
> $$E_\mu(\Psi_{\mathbf{x}}|\Upsilon_\delta = 1) \, < \, E_\mu(\Psi_{\mathbf{x}}) \, < \, E_\mu(\Psi_{\mathbf{x}}|\Upsilon_\delta = 0). \tag{2}$$

Note that the negative correlation of $\Psi_{\mathbf{x}}$ and Υ_δ expresses the assumption that $\mathbf{x}$ and $\mathbf{y}$ are unlikely to be simultaneously both 'bad' in a simulation problem. Note, too, that this 'unlikelyness' depends not only on $\mathbf{x}$ and $\mathbf{y}$ alone, but also on the set of problems $\mathcal{F}$ and their 'weighting' μ as well.

The proof of (2) is simply a transcription of the proof of our statement on pre-testing given in (4.1.2) just with the positive correlation assumed there replaced by a negative one; there is no need to repeat it here.

Now what is the consequence of this if we observe that our simulations $F(\mathbf{x})$ and $F(\mathbf{y})$ produce approximately the same result? If

$$\delta \geq |F(\mathbf{x}) - F(\mathbf{y})|,$$

then (1) implies that

$$\Upsilon_\delta(F) = 1.$$

Hence, assuming the negative correlation of $\Psi_{\mathbf{x}}$ and Υ_δ and observing $|F(\mathbf{x}) - F(\mathbf{y})| \leq \delta$, we are *forced* to be optimistic about the approximation error produced by $\mathbf{x}$. In this sense, Knuth's recommendation is indeed the "most prudent policy for a person to follow".

(4.3.3) One might ask if the assumption of negative correlation can be replaced by a weaker one. In general, this is not possible. If we set $\delta := |F(\mathbf{x}) - F(\mathbf{y})|$ and run the proof of (2) backwards, we get that (2) implies the negative correlation of $\Psi_{\mathbf{x}}$ and Υ_δ: let

$$E_\mu(\Psi_{\mathbf{x}}|\Upsilon_\delta = 1) < E_\mu(\Psi_{\mathbf{x}}).$$

If we assume that $P_\mu(\Upsilon_\delta = 1)$ is defined and positive, this implies

$$\frac{E_\mu(\Psi_{\mathbf{x}}|\Upsilon_\delta = 1) P_\mu(\Upsilon_\delta = 1)}{E_\mu(\Psi_{\mathbf{x}}) P_\mu(\Upsilon_\delta = 1)} < 1.$$

Using the same argument as in (4.1.2) yields

$$\frac{E_\mu(\Psi_{\mathbf{x}}\Upsilon_\delta)}{E_\mu(\Psi_{\mathbf{x}})E_\mu(\Upsilon_\delta)} < 1,$$

which means that $\Psi_{\mathbf{x}}$ and Υ_δ are negatively correlated. In this sense, the negative correlation of $E_\mu(\Psi_{\mathbf{x}})$ and Υ_δ and the optimism expressed in (2) are equivalent.

4.4 Employing deterministic error bounds

(4.4.1) One might question the applicability of deterministic error bounds when dealing with a probabilistic notion of goodness. Deterministic error bounds seem to render the application of deterministic notions of quality such as Criterion 2.3 more appropriate, and we have already remarked in (4.0.2) that Criterion 2.3 is in some sense much stronger than Criterion 2.4. So why using the strong bounds with the weak criterion?
In this section we present two examples of applying Criterion 2.4 in conjunction with a deterministic error bound from which we draw two interesting conclusions. The first example shows that whenever Criterion 2.3 is applicable, Criterion 2.4 is applicable as well, and both criteria suggest the same sequences as being 'good'. In this sense, the weak, probabilistic Criterion 2.4 is *consistent* with the strong, deterministic Criterion 2.3. The second example gives a scenario where a deterministic error bound is available but no 'good' sequence can be found with Criterion 2.3. In this case, Criterion 2.4 will turn out to be applicable.

In both examples we will use the following result which is introduced, stated formally, and proved later on, in Appendix A.

> Let $\mathbf{T}$ be a mapping from $\mathcal{F}$ into $\mathbf{R}^s$ and let $\mathbf{t}$ be a point in $\mathbf{R}^s$. There exists a probability space $(\mathcal{F}, \mathcal{R}, \mu)$ with $E_\mu(\mathbf{T}) = \mathbf{t}$ (taking expectations component-wise) if and only if $\mathbf{t}$ lies in the convex hull of the set $\mathbf{T}(\mathcal{F})$.

With this, the problem of proving the existence of a probability space $(\mathcal{F}, \mathcal{R}, \mu)$ with $E_\mu(\mathbf{T}) = \mathbf{t}$ is equivalent to solving a relatively simple geometric problem. We know that using a result before proving it is not the way to run a smooth argument; regrettably the proof of the above is so completely off the track of thought we have followed so far that we cannot help but do just this.

(4.4.2) In both examples, we will use a deterministic error bound known as Koksma's inequality[26]. This inequality concerns functions F_f of the form

$$F_f(\mathbf{x}) = \frac{1}{N}\sum_{n=0}^{N-1} f(x_n),$$

where f is a real-valued function on the closed unit interval[27] with bounded variation[28].

Recall that approximative solutions of the problem F_f are approximations of the integral of f. If f is integrable and $\mathbf{X} = (X_n)_{n=0}^{N-1}$ is a sequence of N stochastically independent random variables each of which is equidistributed on $[0,1[$, then the expectation of $F_f(\mathbf{X})$ equals the expectation of $f(X_0)$, which in turn is $\int f(X_0)d\lambda$: by means of Fubini's theorem and the independence of the X_n, we have

$$\begin{aligned} E(F_f(\mathbf{X})) &= \int_{[0,1]^N} F_f(\mathbf{X})d\lambda_N \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\int_{[0,1]^N} f(X_n)d\lambda_N \end{aligned}$$

[26] Other deterministic error bounds are available, and we might have used them instead. A particularly interesting alternative is an error bound based on the diaphony which was introduced by Zinterhof in [117]; see also Stegbuchner [107]. The advantage of this notion is that, in contrary to the discrepancy on which Koksma's inequality is based, the diaphony of N points in $\mathbf{R}^s$ can actually be *computed* with reasonable effort for $s \geq 1$. Moreover, as pointed out on page 51, the diaphony has an interesting relation to the spectral test.

The reason for choosing Koksma's inequality is twofold. First of all, Koksma's inequality is widely known and – mostly in the work of Niederreiter – ratings for various types of random number generators with respect to this inequality are available. The second, personal reason is that the concept of Criterion 2.4 was perceived in the contemplation of Koksma's inequality.

[27] A generalization for the more interesting case of f being a function of more than one variable is available. Since we intend to use Koksma's inequality primarily for illustrative purpose, the special case of f depending on just one variable will suffice. Moreover, the generalization has all the properties we require from the one-dimensional case, so all we are going to show for functions f of one variable applies also to functions f of many variables.

[28] This is a technical requirement which we will describe in the following.

$$\begin{aligned} &= \frac{1}{N}\sum_{n=0}^{N-1}\int_{[0,1]} f(X_n)d\lambda \\ &= \int f(X_0)d\lambda. \end{aligned}$$

For these problems F_f, Koksma's inequality (due to Koksma [60]; see also [87, Theorem 2.9]) provides an upper bound for the approximation error $\Psi_{\mathbf{x}}(F_f) := |F_f(\mathbf{x}) - E(F_f)|$. If f has bounded variation $V(f)$ on $[0,1]$, then

$$\Psi_{\mathbf{x}}(F_f) \leq V(f)D_N^\star(\mathbf{x}). \tag{1}$$

We do not prove Koksma's inequality here. For a proof of it and its generalizations for functions of more than one variable, the Koksma-Hlawka inequality, the reader is refered to Niederreiter [87, Theorem 2.9 and 2.11].

Before we describe the quantities on the right side of Koksma's inequality in detail, observe its formal appearance. The error $\Psi_{\mathbf{x}}(F_f)$ depends on both $\mathbf{x}$ and f simultaneously. With Koksma's inequality, the error is bounded by the product of two quantities $D_N^\star(\mathbf{x})$ and $V(f)$, where the first depends only on $\mathbf{x}$ and the second only on f. The contributions of $\mathbf{x}$ and f to the error bound are – in this sense – independent!

The quantity $D_N^\star(\mathbf{x})$, the *star-discrepancy*, is defined as the rating of $\mathbf{x}$ with respect to Criterion 2.3 and the set

$$\mathcal{I} := \left\{ F_{1_{[0,t]}} : \ 0 \leq t \leq 1 \right\}.$$

Stated formally, this means

$$\begin{aligned} D_N^\star(\mathbf{x}) &:= \sup_{\mathcal{I}} \Psi_{\mathbf{x}} \\ &= \sup_{0\leq t \leq 1} \left| F_{1_{[0,t]}}(\mathbf{x}) - E(F_{1_{[0,t]}}(\mathbf{X})) \right|. \end{aligned}$$

One might ask how a sequence's rating with respect to the particular set $\mathcal{I}$ can be used to bound its rating with respect to arbitrary functions of bounded variation. A hint is provided by a closer examination of the functions $F_{1_{[0,t]}}$. For one, we have

$$\begin{aligned} F_{1_{[0,t]}}(\mathbf{x}) &= \frac{1}{N}\sum_{n=0}^{N-1} 1_{[0,t]}(x_n) \\ &= \frac{1}{N}\# \left\{ x_n : \ x_n \leq t, \, 0 \leq n < N \right\}, \end{aligned}$$

and for the other

$$\begin{aligned} E(F_{1_{[0,t]}}(\mathbf{X})) &= \int 1_{[0,t]}(X_0)d\lambda \\ &= \lambda([0,t]) \\ &= t. \end{aligned}$$

So $F_{1_{[0,t]}}(\mathbf{x})$ is the empirical distribution function of the numbers x_n and $E(F_{1_{[0,t]}}(\mathbf{X}))$ is the equidistribution's distribution function, both evaluated at t. From this point of view, the star-discrepancy is the *distance* of these two distribution functions with respect to the supremum-norm. Introduced as such, the star-discrepancy is a special case of what is called Kolmogorov-Smirnov statistic in the literature (see Bury [9, Section 6.11 – 6.14] or Knuth [59, Section 3.3]).

For functions of the form $F_{1_{[0,t]}}$, the following inequality holds due to the definition of $D_N^\star(\mathbf{x})$:

$$\Psi_{\mathbf{X}}(F_{1_{[0,t]}}) \leq D_N^\star(\mathbf{x}).$$

This is generalized for more interesting functions than simple indicators by the concept of variation. The quantity $V(f)$, the variation of f on $[0,1]$, is defined as

$$V(f) := \sup\left\{\sum_{n=1}^{n} |f(a_n) - f(a_{n-1})| : \; n \geq 1, \, 0 \leq a_0 < a_1 < \ldots < a_n \leq 1\right\}.$$

Note that if f is non-decreasing or non-increasing, then $V(f) = |f(0) - f(1)|$. Hence, for $F_{1_{[0,t]}} \in \mathcal{I}$, we have $V(1_{[0,t]}) = t$ and $\sup\{V(1_{[0,t]}) : \; F_{1_{[0,t]}} \in \mathcal{I}\} = 1$.

(4.4.3) With this we are ready to use Koksma's inequality with the deterministic Criterion 2.3. For a fixed real number $c \geq 0$, let $\mathcal{F}_c$ be the set of all problems F_f for which the variation of f is at most c, i.e.

$$\mathcal{F}_c := \{F_f : \; V(f) \leq c\}.$$

For this set, the following equality[29] holds:

$$\sup_{\mathcal{F}_c} \Psi_{\mathbf{X}} = cD_N^\star(\mathbf{x}). \tag{2}$$

[29] The idea for this is from Niederreiter's [87, Theorem 2.12]. A consequence of his much stronger result is that, instead of all functions F_f with $V(f) \leq c$, only those might be used for which f is continuously differentiable infinitely often.

In the sense of Criterion 2.3, the sequence $\mathbf{x}$ is considered 'good' with respect to the set $\mathcal{F}_c$ if $cD_N^\star(\mathbf{x})$ is small. Just as with Koksma's inequality the contribution of $\mathbf{x}$ to the error bound was found to be independent of the contribution of the function f, we find here that the contribution of $\mathbf{x}$ to $\sup_{\mathcal{F}_c} \Psi_{\mathbf{x}}$ is independent of $\mathcal{F}_c$. This is particularly useful since it causes us to regard $\mathbf{x}$ as 'good' if $D_N^\star(\mathbf{x})$ is small, regardless of whatever value c might have.

Proof of (2): Avoiding trivialities[30], let $c > 0$ be fixed. Note that $\Psi_{\mathbf{x}}(F_{cf})$ and $V(cf)$ both are – by definition – linear in c:

$$\Psi_{\mathbf{x}}(F_{cf}) = c\Psi_{\mathbf{x}}(F_f)$$

and

$$V(cf) = cV(f).$$

Instead of considering the whole set $\mathcal{F}_c$, we will find that a quite small subset $\mathcal{I}_c$ is representative with respect to the supremum of $\Psi_{\mathbf{x}}$. Let

$$\mathcal{I}_c := \left\{F_{c1_{[0,t]}} :\ 0 \le t \le 1\right\}.$$

Note that $\mathcal{I}_1 = \mathcal{I}$. Since $\mathcal{F}_c$ is a superset of $\mathcal{I}_c$, we have of course

$$\sup_{\mathcal{F}_c} \Psi_{\mathbf{x}} \ge \sup_{\mathcal{I}_c} \Psi_{\mathbf{x}}.$$

On the other hand, there is[31]

$$\begin{aligned}
\sup_{\mathcal{F}_c} \Psi_{\mathbf{x}} &\le \sup\{D_N^\star(\mathbf{x})V(f) :\ F_f \in \mathcal{F}_c\} \\
&= cD_N^\star(\mathbf{x}) \\
&= c\sup_{\mathcal{I}} \Psi_{\mathbf{x}} \\
&= \sup_{\mathcal{I}} c\Psi_{\mathbf{x}} \\
&= \sup_{\mathcal{I}_c} \Psi_{\mathbf{x}}.
\end{aligned}$$

These two inequalities imply (2). □

[30] For $c = 0$, $\mathcal{F}_c$ contains only those F_f for which $V(f) = 0$, i.e. for which f is constant. For these F_f, the error $\Psi_{\mathbf{x}}(F_f)$ is always zero and (2) does always hold, independent of $\mathbf{x}$.

[31] In this series of (in-)equalities, we make use of Koksma's inequality, the definitions of $\mathcal{F}_c$ and $D_N^\star(\mathbf{x})$, the linearity of $\Psi_{\mathbf{x}}(F_{cf})$ in c, and the definition of $\mathcal{I}_c$.

Now suppose we search for a sequence $\mathbf{x}$ which should be 'good' with respect to $\mathcal{F}_c$ by applying Criterion 2.4 instead of Criterion 2.3. Will we prefer sequences with small star-discrepancy? Will the application of the weak criterion suggest the same sequences as the strong one?
There is no answer to this question in general. We might consider a statistical test T as highly relevant, which a sequence $\mathbf{x}$ passes, but which a sequence $\mathbf{y}$ with smaller star-discrepancy utterly fails. As described in Section 4.1, applying Criterion 2.4 in this case indicates that $\mathbf{x}$ should be preferred although its star-discrepancy is larger. But doing so, we would have used information – the relevance of the test T and the behavior of $\mathbf{x}$ and $\mathbf{y}$ in this test – which Criterion 2.3 per definition does not take into account. Let us try to apply Criterion 2.4 using no more information than we have needed to apply Criterion 2.3 in the first place.

Let $(\mathcal{F}_c, \mathcal{R}, \mu)$ be a probability space such that $\Psi_{\mathbf{x}}$ is a real-valued random quantity. Suppose *all* we know about this probability space is the value[32] of c; in particular, suppose $\mathcal{R}$ and μ are unknown to us. What can be derived about the expected approximation error $E_\mu(\Psi_{\mathbf{x}})$? We can of course use the deterministic error bound (2): since $\Psi_{\mathbf{x}} \leq \sup_{\mathcal{F}_c} \Psi_{\mathbf{x}}$ with certainty, we have

$$E_\mu(\Psi_{\mathbf{x}}) \leq c D_N^\star(\mathbf{x}). \qquad (3)$$

But how accurate is this upper bound? The measure μ is not known to us, and the expectation of $\Psi_{\mathbf{x}}$ with respect to this unknown measure can be quite far away from $cD_N^\star(\mathbf{x})$. The point is this:

> If only c is known but μ is not, then the upper bound in (3) is best possible.

By 'best possible', we understand that there are so many probability spaces consistent with the provided information (the value of c), i.e. there are so many σ-algebrae $\mathcal{R}$ on $\mathcal{F}_c$ and so many measures μ on these $\mathcal{R}$ that *all*, that can be said about $E_\mu(\Psi_{\mathbf{x}})$ is that it is at most $cD_N^\star(\mathbf{x})$. If nothing else is known, $E_\mu(\Psi_{\mathbf{x}})$ can be arbitrarily close to $cD_N^\star(\mathbf{x})$. Lacking additional information, we regard $\mathbf{x}$ as 'good' with respect to Criterion 2.4 if $cD_N^\star(\mathbf{x})$ is small. In particular, we regard a sequence with small star-discrepancy as 'good', regardless of the actual value of c.

Proof that (3) is best possible: Avoiding trivialities, let $c > 0$ be fixed and known to us. Let $\mathcal{R}$ be a consistent σ-algebra on $\mathcal{F}_c$, i.e. one

[32] This information is required by Criterion 2.3, too.

for which $\Psi_{\mathbf{x}}$ is measurable. Finally, let $\mathcal{M}$ be the set of all probability measures on $(\mathcal{F}_c, \mathcal{R})$ (note that $\mathcal{M}$ is the set of all probability measures consistent with the available information and with $\mathcal{R}$). For the set

$$\mathcal{E} := \{E_\mu(\Psi_{\mathbf{x}}) : \mu \in \mathcal{M}\},$$

we show

$$\sup \mathcal{E} = cD_N^\star(\mathbf{x}),$$

which means that (3) is best possible.

Proposition A.1 states that $\mathcal{E}$ is equal to the convex hull of the possible values $\Psi_{\mathbf{x}}$:

$$\mathcal{E} = \operatorname{conv} \Psi_{\mathbf{x}}(\mathcal{F}_c).$$

The one-dimensional convex set $\operatorname{conv} \Psi_{\mathbf{x}}(\mathcal{F}_c)$ is an interval[33] and, due to (2), its upper endpoint is $cD_N^\star(\mathbf{x})$. □

(4.4.4) Since Criterion 2.3 worked well for the set $\mathcal{F}_c$ and any finite $c \geq 0$, one might expect it to be applicable to the set

$$\mathcal{F}_\infty := \{F_f : V(f) < \infty\},$$

too. $\mathcal{F}_\infty$ is the union of all the sets $\mathcal{F}_c$ taken over all values $c \geq 0$ and therefore it is a superset of each specific $\mathcal{F}_c$; this and (2) yield

$$\forall c \geq 0 : \; cD_N^\star(\mathbf{x}) \leq \sup_{\mathcal{F}_\infty} \Psi_{\mathbf{x}}.$$

Hence $\sup_{\mathcal{F}_\infty} \Psi_{\mathbf{x}} = \infty$, independent of the star-discrepancy of $\mathbf{x}$. For this set $\mathcal{F}_\infty$, no 'good' sequence $\mathbf{x}$ can be found with Criterion 2.3 since $\sup_{\mathcal{F}_\infty} \Psi_{\mathbf{x}}$ equals ∞ for any sequence $\mathbf{x}$.

This is not the expected result. For any finite $c \geq 0$ and the corresponding set $\mathcal{F}_c$, Criterion 2.3 caused us to prefer a sequence with small star-discrepancy. Intuition suggests that for the union $\mathcal{F}_\infty$ of these sets, sequences with small star-discrepancy should be preferable, too.
Of course, if $V(f)$ is finite but arbitrarily large, (2) implies that $\Psi_{\mathbf{x}}(F_f)$ can also be arbitrarily large. The intuitive preference for sequences with small star-discrepancy seems to be based on the assumption that although $V(f)$ *can* be arbitrarily large, it is not *expected* to be. Let us express this assumption in the terminology of Criterion 2.4.

[33] See page 106 for a proof of this.

Let $(\mathcal{F}_\infty, \mathcal{R}, \mu)$ be a probability space such that $\Psi_{\mathbf{x}}$ and the mapping $V : F_f \mapsto V(f)$ are real-valued random variables. Suppose we know that $E_\mu(V) \leq c$, but $\mathcal{R}$ and μ are not known to us. Due to Koksma's inequality and $E_\mu(V) \leq c$, we have

$$E_\mu(\Psi_{\mathbf{x}}) \leq c D_N^\star(\mathbf{x}). \tag{4}$$

In analogy to the previous example, the following holds.

> In this state of information, the upper bound in (4) is best possible.

Hence, if no more information is available, then (4) is *all* that can be derived about $E_\mu(\Psi_{\mathbf{x}})$. We regard a sequence $\mathbf{x}$ as 'good' with respect to Criterion 2.4 and the given information if $cD_N^\star(\mathbf{x})$ is small; since the 'quality' of $\mathbf{x}$ does not depend on the actual value of c, we regard a sequence as 'good' with respect to the available information if $D_N^\star(\mathbf{x})$ is small, regardless of c.

Proof that (4) is best possible: Let $\mathcal{R}$ be any σ-algebra on $\mathcal{F}_\infty$ for which $\Psi_{\mathbf{x}}$ and V are measurable. As a auxiliary tool, we define the mapping

$$\begin{aligned} \Gamma : \mathcal{F}_\infty &\longrightarrow \mathbf{R}^2 \\ F_f &\longmapsto (\Psi_{\mathbf{x}}(F_f), V(f)), \end{aligned}$$

which is measurable since both $\Psi_{\mathbf{x}}$ and V are measurable.
What we have to show is the existence of probability measures μ on $(\mathcal{F}_\infty, \mathcal{R})$ which are consistent with the available information (i.e. $E_\mu(V) \leq c$) and for which $E_\mu(\Psi_{\mathbf{x}})$ gets arbitrarily close to $cD_N^\star(\mathbf{x})$.
Due to Proposition A.1, this is equivalent to showing that there are points (r, s) in the convex hull of $\Gamma(\mathcal{F}_\infty)$ with $s \leq c$ for which r gets arbitrarily close to $cD_N^\star(\mathbf{x})$.

The set $\mathcal{F}_c$ is contained in $\mathcal{F}_\infty$ and so are the corresponding convex hulls. In (4.4.3) we have shown that

$$\sup \{r : r \in \mathrm{conv}\{\Psi_{\mathbf{x}}(\mathcal{F}_c)\}\} = cD_N^\star(\mathbf{x}).$$

Since

$$\{(r, s) \in \mathrm{conv}\, \Gamma(\mathcal{F}_\infty) : \ s \leq c\} = \mathrm{conv}\, \Gamma(\mathcal{F}_c),$$

we have

$$\begin{aligned}\sup\{r:\ (r,s)\in \operatorname{conv}\Gamma(\mathcal{F}_\infty), s\le c\} &= \sup\{r:\ (r,s)\in \operatorname{conv}\Gamma(\mathcal{F}_c)\}\\ &= \sup\{r: r\in \operatorname{conv}\Psi_{\mathbf{x}}(\mathcal{F}_c)\}\\ &= cD_N^\star(\mathbf{x}).\end{aligned}$$

□

Chapter 5

Some generators

He deals the cards to find the answer
the sacred geometry of chance
the hidden law of a probable outcome.
The numbers lead a dance.
– Sting, Shape of My Heart

5.1 Preliminaries

(5.1.1) Apart from designing statistical tests, the best assistance mathematics can provide to a user searching for a 'good' random sequence is the analysis of random number generators. For these deterministic algorithms for generating random numbers, the presence or absence of certain defects can be *proven.* Understanding the nature of these defects and assessing their relevance for his simulation problem, the user can judge whether or not a given generator can be considered adequate ('good').

In the following, we give a precise definition of the generators presented in Section 4.2, and state and prove the existence or absence of the already mentioned defects. The reader who is not interested in the mathematical details involved and who is content to believe what we have said about the generators so far may skip this chapter.

(5.1.2) The generators we are about to consider and in fact virtually all generators in use today are *congruential* generators. They produce a sequence $(u_n)_{n=0}^{N-1}$ of integers which are spread evenly in some set $\{0, 1, \ldots, M-1\}$ and transform these integers to a sequence $(x_n)_{n=0}^{N-1}$ of random numbers in $[0, 1[$ by setting $x_n := u_n/M$. The reason for this is twofold. Performing all calculations in integer arithmetic except for the final scaling assures that no propagating round-off errors are introduced in the actual generation of the random numbers on a computer. In this way, a random number generator will produce the same sequence regardless of the floating point representation used by the computer it is run on[1]. The other reason for first computing an integer sequence is the mathematical structure of the set $\{0, \ldots, M-1\}$, which we will employ in the following and which is briefly sketched below; a more complete discussion of these algebraic and number theoretic topics is given by Lidl and Niederreiter in [75].

(5.1.3) For a positive integer M, let $\mathbf{Z}_M := \{0, 1, \ldots, M-1\}$ be the system of all residues modulo M. Furthermore, let $\mathbf{Z}_M^\star$ be the set of all numbers from $\mathbf{Z}_M$ which are coprime to M. To obtain some arithmetic structure on $\mathbf{Z}_M$, we use the modulo operation $a \bmod M$, which yields the residue of the integer division of a by M. The addition of two elements a and b of $\mathbf{Z}_M$ is defined as

$$(a+b) \bmod M$$

and their multiplication as

$$(a \cdot b) \bmod M.$$

We will omit the trailing mod and write $a+b$, $a \cdot b$, or simply ab when it is clear that we are operating in $\mathbf{Z}_M$.

With these conventions, we get the following:

$(\mathbf{Z}_M, +)$ and $(\mathbf{Z}_M^\star, \cdot)$ both are abelian groups.

Moreover, if $M = p$ is prime, then

$(\mathbf{Z}_p, +, \cdot)$ is a finite field[2].

[1] This holds provided the integer computations are correctly implemented, which is not always the case (see [1, Section 3.1.2] or [100]). Fortunately, it seems to have become unpopular nowadays to believe that a correct implementation of a random number generator is as good as an incorrect one since its output is supposed to be random anyway...

[2] For $a \in \mathbf{Z}_p^\star$, its inverse a^{-1} is the uniquely determined number such that $a \cdot a^{-1} = 1$ in $\mathbf{Z}_p$. The inverse can be obtained by the Euclidean algorithm: since a and p are coprime, there are integers k and l such that $ka + lp = 1$; thus $a^{-1} = k \bmod p$. Another way to get a^{-1} is Fermat's theorem, which states that $a^{p-1} = 1$ in $\mathbf{Z}_p$; therefore $a^{-1} = a^{p-2} \bmod p$.

Except for field-isomorphisms, there is a unique finite field with p elements called F_p; $\mathbf{Z}_p$ can therefore be identified with the general field F_p.
The integral domain of all polynomials in x over F_p is denoted by $F_p[x]$. Working with polynomials from $F_p[x]$ is very similar to working with polynomials over $\mathbf{R}$. The fundamental theorem of algebra[3] holds in $F_p[x]$, and two polynomials from $F_p[x]$ are equal if and only if their coefficients coincide. Since $(\mathbf{Z}_p, +, \cdot)$ is a field, there is – for $s \geq 1$ – an s-dimensional vector space over $\mathbf{Z}_p$ which we denote by $\mathbf{Z}_p^s$ or F_p^s.

5.2 LCG

Definition 5.1 *Let $a, b, u_0 \in \mathbf{Z}_M$. The* linear congruential generator*, LCG, for short, with parameters M, a, b, and u_0 defines a sequence $(u_n)_{n\geq 0}$ in $\mathbf{Z}_M$ by*

$$u_n := a \cdot u_{n-1} + b \qquad (n > 0)$$

and a sequence $(x_n)_{n\geq 0}$ of random numbers in $[0, 1[$ by

$$x_n := \frac{u_n}{M} \qquad (n \geq 0).$$

(5.2.1) The sequence $(u_n)_{n\geq 0}$ of an LCG is defined by a recursion of order one on the finite set $\mathbf{Z}_M$. Hence it will eventually repeat itself and so will the corresponding sequence of random numbers. Since this is usually considered a major defect, we adopt the convention of using only as many of the x_n as do not enter a cycle; formally, we use $(u_n)_{n=0}^{N-1}$ and $(x_n)_{n=0}^{N-1}$ with N defined as

$$N := \min\{j > 0 : \exists i < j : u_i = u_j\}.$$

Moreover, we will consider only LCGs whose sequences are purely periodic (i.e. $x_0 = x_N$); in this case N is called the *period* of the corresponding generator.
We denote the sequence $(u_n)_{n=0}^{N-1}$ by $\mathrm{lcg}(M, a, b, u_0)$ and the sequence $(x_n)_{n=0}^{N-1}$ by $\mathrm{LCG}(M, a, b, u_0)$.

In the following, we will consider the s-tuples $(x_n, \ldots, x_{n+s-1})$ formed from the sequence $(x_n)_{n=0}^{N-1}$ and the analogous sequence of s-tuples formed

[3] If $h \in F_p[x]$ is a non-zero polynomial of degree k, then h has at most k roots in F_p.

from $(u_n)_{n=0}^{N-1}$. To avoid technical difficulties, we always implicitly assume that the indices are reduced modulo N. This is, say,

$$(x_{N-1}, x_N, \ldots, x_{N+s-2}) = (x_{N-1}, x_0, \ldots, x_{s-2}).$$

In this way, we do not have to bother with a wrap-around in the s-tuples when n approaches N.

(5.2.2) Since we are operating on the M-element set $\mathbf{Z}_M$, the period of an LCG is at most M. Necessary and sufficient conditions on the parameters a, b, and u_0 to achieve the maximal possible period length are given in Knuth [59, Section 3.2., Theorem A, B, and C] or Ripley [93, Theorem 2.1, 2.2, and 2.3].
For our present purpose, we restrict ourselves to the following three standard types[4] of maximal possible period LCGs which are either mathematically interesting or computationally efficient.

1. M a power of 2, $a = 5 \bmod 8$, b odd.
   In this case, we get a period of $N = M$.

2. $M = p$ prime, a a primitive root[5] modulo p, $b = 0$, $u_0 \neq 0$.
   In this case, we get a period of $N = p - 1$.

3. $M \geq 4$ a power of 2, $a = 5 \bmod 8$, $b = 0$, u_0 odd.
   In this case, we get a period of $N = M/4$.

Note that for each type, the set $\{u_n :\ 0 \leq n < N\}$ of the integers produced by an LCG is suspiciously regularly distributed: it is equal to $\mathbf{Z}_M$ for Type[6] 1 and equal to $\mathbf{Z}_M \setminus \{0\}$ for Type[7] 2; for Type 3, it is equal to all numbers of the form $2k+1$ in $\mathbf{Z}_M$, where $k \geq 0$ is either an even or an odd integer

[4] These standard types of LCGs are obtained from Ripley [92] or from Niederreiter [87, p.169].

[5] $a \in \mathbf{Z}_p$ is a primitive root modulo p if the set of powers $\{a^1, a^2, \ldots, a^{p-1}\}$ is equal to $\mathbf{Z}_p^\star$, or, in other words, if a generates $\mathbf{Z}_p^\star$.

[6] The recursion operates in the M-element set $\mathbf{Z}_M$ where it achieves the maximal possible period of M; since it produces exactly M different values, each value in $\mathbf{Z}_M$ is produced exactly once.

[7] In this case, the recursion operates in the $(M-1)$-element set $\mathbf{Z}_M \setminus \{0\}$ and has a period of $M - 1$.

depending on whether the number l, for which $u_0 = 2l+1$, is even or odd[8]. The regular pattern of the u_n in $\mathbf{Z}_M$ is transposed to a regular pattern of the x_n in $[0,1[$.

(5.2.3) To prove the existence of a lattice structure in the LCG's output, let $s \geq 2$ be a fixed integer and let

$$\mathcal{P} := \{\mathbf{x}_n = (x_n, x_{n+1}, \ldots, x_{n+s-1}) : \ 0 \leq n < N\}$$

be the set of s-dimensional points produced by the LCG. To describe the lattice-like structure of $\mathcal{P}$, we need the following formal definition of an s-dimensional lattice, which is given according to Gruber and Lekkerkerker [46, Section 1.3, Definition 1].

Definition 5.2 *Let* $\mathbf{a}_0, \ldots, \mathbf{a}_{s-1} \in \mathbf{R}^s$ *be linearly independent. The set*

$$\Lambda := \left\{ \sum_{i=0}^{s-1} k_i \mathbf{a}_i \ : \ k_i \in \mathbf{Z} \right\}$$

is called the s-dimensional lattice *with basis* $\{\mathbf{a}_0, \ldots, \mathbf{a}_{s-1}\}$.

A lattice shifted by a vector $\mathbf{x}$ is denoted by $\mathbf{x} + \Lambda$.

The lattice structure of the set $\mathcal{P}$ produced by an LCG is described by the following proposition which, as its proof, is given according to Ripley [92] (see also [87, Theorem 7.6]).

Proposition 5.1 *For an LCG of Type 1 and 3, there exists a lattice* Λ *such that*

$$\mathcal{P} \ = \ [0,1[^s \cap (\mathbf{x}_0 + \Lambda).$$

For an LCG of Type 2, there exists a lattice Λ *such that*

$$\mathcal{P} \ =]0,1[^s \cap \Lambda.$$

[8] A simple inductive argument shows that, if $u_0 = 2(2 \cdot k) + 1$, the LCG's recursion operates in the $(M/4)$-element set $\{4 \cdot i + 1 : \ 0 \leq i < M/4\}$ where it achieves a period of $M/4$. Conversely, if $u_0 = 2(2 \cdot k + 1) + 1 = 4 \cdot k + 3$, then the recursion operates in the $(M/4)$-element set $\{4 \cdot i + 3 : \ 0 \leq i < M/4\}$ where it again achieves a period of $M/4$.

Proof of Proposition 5.1: We prove Proposition 5.1 only for LCGs of Type 1; the proof for the remaining types uses the same basic idea and differs only in some details. The idea is based on the following trick to represent the j-th component of the integer vector $\mathbf{u}_n = (u_n, \ldots, u_{n+s-1})$. For any $\mathrm{lcg}(M, a, b, u_0)$, we have

$$u_{n+j} = u_j + a^j(u_n - u_0) \qquad (0 \leq j < s)$$

in $\mathbf{Z}_M$ as is easily proved by induction[9].

Let $(u_n)_{n=0}^{M-1} = \mathrm{lcg}(M, a, b, u_0)$ be of Type 1. Since the period of this generator is M, we may assume[10] $u_0 = 0$. For a fixed index n, the above representation of u_{n+j} yields

$$u_{n+j} = u_j + a^j u_n \qquad (0 \leq j < s)$$

in $\mathbf{Z}_M$. Since addition and multiplication in $\mathbf{Z}_M$ are defined via reduction modulo M, there are uniquely determined integers k_j for which we have

$$u_{n+j} = u_j + a^j u_n + k_j M \qquad (0 \leq j < s)$$

in $\mathbf{R}$. For the corresponding x_n, this means

$$x_{n+j} = x_j + \frac{a^j}{M} u_n + k_j \qquad (0 \leq j < s). \tag{1}$$

To extend this equality to the vectors $\mathbf{x}_n$, we define the s linearly independent vectors

$$\begin{aligned} \mathbf{b}_0 &:= \frac{1}{M}(1, a, a^2, \ldots, a^{s-1}), \\ \mathbf{b}_j &:= \mathbf{e}_j, \text{ the } j\text{-vector of the standard ordered basis of } \mathbf{R}^s \ (1 \leq j < s). \end{aligned}$$

[9] Let the index n be fixed. The equality is trivial for $j = 0$. Supposing it holds for some $j \geq 0$, we get

$$\begin{aligned} u_{n+j+1} &= a u_{n+j} + b \\ &= a(u_j + a^j(u_n - u_0)) + b \\ &= a u_{j+1} + a^{j+1}(u_n - u_0) \end{aligned}$$

in $\mathbf{Z}_M$.

[10] The sequences $\mathrm{lcg}(M, a, b, 0) = (u_n)_{n=0}^{M-1}$ and $\mathrm{lcg}(M, a, b, u_0) = (v_n)_{n=0}^{M-1}$ of Type 1 differ only by a cyclic permutation; in particular, both sequences yield the same set $\mathcal{P}$.

Note that we use 0-based indices. Hence the standard ordered basis of $\mathbf{R}^s$ is $\{\mathbf{e}_0, \ldots, \mathbf{e}_{s-1}\}$, All the components of $\mathbf{e}_i$ are 0 except for the $(i+1)$-st, which is 1.

With this, we can write (1) in vector form:

$$\mathbf{x}_n = \mathbf{x}_0 + u_n \mathbf{b}_0 + \sum_{j=1}^{s-1} k_j \mathbf{b}_j.$$

Hence every point $\mathbf{x}_n$ lies on the shifted lattice $\mathbf{x}_0 + \Lambda$, where Λ is spanned by the lattice basis $\{\mathbf{b}_0, \ldots, \mathbf{b}_{s-1}\}$[11]. In other words,

$$\mathcal{P} \subseteq [0,1[^s \cap (\mathbf{x}_0 + \Lambda).$$

Conversely, if we take an arbitrary point $\mathbf{p}$ from $[0,1[^s \cap (\mathbf{x}_0 + \Lambda)$, it has the form

$$\mathbf{p} = \mathbf{x}_0 + l_0 \mathbf{b}_0 + \sum_{j=1}^{s-1} l_j \mathbf{b}_j \qquad (l_j \in \mathbf{Z}).$$

Since $\mathbf{p}$ lies in $[0,1[^s$, its first coordinate is between 0 and 1: $0 \le x_0 + l_0 \frac{1}{M} < 1$. Recalling that we assumed $x_0 = 0$ and multiplying this inequality by M, we get

$$0 \le l_0 < M,$$

i.e. $l_0 \in \mathbf{Z}_M$. Since the integers u_n run through the whole set $\mathbf{Z}_M$, there is $l_0 = u_n$ for some n. Thus

$$\mathbf{p} = \mathbf{x}_0 + u_n \mathbf{b}_0 + \sum_{j=1}^{s-1} l_j \mathbf{b}_j.$$

This representation of $\mathbf{p}$ is quite similar to the above representation of $\mathbf{x}_n$; it remains to show that $l_i = k_j$ for $1 \le j < s$.

As we have noted before, the integers k_j are uniquely determined by the reduction modulo M, i.e. by demanding that $u_{n+j} \in \mathbf{Z}_M$ or, equivalently, that $x_{n+j} \in [0,1[$. In other words: the integers $k_1, \ldots, k_{s-1}$ are the only ones for which

$$\mathbf{x}_0 + u_n \mathbf{b}_0 + \sum_{j=1}^{s-1} k_j \mathbf{b}_j \in [0,1[^s.$$

[11] This lattice basis is the same for LCGs of Type 1 and 2. The lattice basis for a Type 3 LCG is similar, just with M replaced by $M/4$ in the definition of $\mathbf{b}_0$.

Thus $l_i = k_i$ for $1 \leq i < s$, $\mathbf{p} = \mathbf{x}_n$, and

$$[0,1[^s \cap (\mathbf{x}_0 + \Lambda) \subseteq \mathcal{P}.$$

With this, the two sets are equal. □

(5.2.4) Although we have proven the existence of a (shifted) lattice formed by an LCG's s-dimensional points with Proposition 5.1 and although we obtained a basis for the lattice in the proof, this particular basis does not seem to be well suited for assessing the lattice's quality. Considering the lattices formed by the LCGs in Figure 3 on page 49 and looking at the corresponding basis vectors as obtained from the proof, we have to admit that these lattice bases are not those we would have chosen by intuition. For the generator in Figure 3a, the lattice basis found in the proof is

$$\begin{aligned} \mathbf{b}_0 &= (\frac{1}{1024}, \frac{1021}{1024}), \\ \mathbf{b}_1 &= (0,1), \end{aligned}$$

whereas the corresponding lattice basis for the generator in Figure 3b is

$$\begin{aligned} \mathbf{b}_0 &= (\frac{1}{1024}, \frac{997}{1024}), \\ \mathbf{b}_1 &= (0,1). \end{aligned}$$

In the contrary to the resulting lattices, these bases do not seem to be much different. The basis suggested by the visual perception of a 2- or 3-dimensional lattice is one whose vectors have the shortest possible length.

This feature is provided by a Minkowski-reduced lattice basis (MRLB, for short) which is defined by the following – non-constructive – procedure: given an s-dimensional lattice Λ, a MRLB $\{\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}\}$ is obtained by first choosing a basis vector $\mathbf{m}_0$ as short as possible and, if $\mathbf{m}_0, \ldots, \mathbf{m}_{k-1}$ $(0 < k < s)$ have been chosen, choosing a basis vector $\mathbf{m}_k$ as short as possible. Note that as a consequence of this procedure, the vectors of a MRLB are ordered according to their length: $||\mathbf{m}_0|| \leq ||\mathbf{m}_1|| \leq \ldots \leq ||\mathbf{m}_{s-1}||$. Although an MRLB of a lattice does always exist, it is, in general, not unique. More precisely, for $s \geq 7$, one can find a lattice Λ and two MRLBs $\{\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}\}$ and $\{\mathbf{n}_0, \ldots, \mathbf{n}_{s-1}\}$ of Λ with $||\mathbf{m}_k|| \neq ||\mathbf{n}_k||$ for some k. For $s < 7$, however, this is not possible.

Since Minkowski-reduced lattice bases are used only as an auxiliary notion in the next paragraph of this section, we will define and discuss it and the

above claims more formally later, in Section B.2 in the appendix.
Algorithms to compute a MRLB for a given LCG and dimensions $s = 2$, $3 \leq s \leq 6$, and $s \geq 7$, respectively, are given by Afflerbach in [1, Chapter 3, Algorithms MB.2, MB.3-6, and MB.n][12].

(5.2.5) The notion of a Minkowski-reduced lattice basis gives us a precise formulation of the intuitive idea of choosing a basis of successively shortest possible vectors. With this, we can define the figures of merit for assessing the quality of a lattice, which we informally introduced in (4.2.4).

The Beyer-quotient q_s was pictured there as 'the relation q_s of the shortest to the longest edge of the unit cell of the lattice'. Since the vectors of a MRLB span what is intuitively called a unit cell, we can – at least for $s \leq 6$ – define according to Beyer [5]:

> Let Λ be the s-dimensional lattice of an LCG as in Proposition 5.1. For a MRLB $\{\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}\}$ of Λ, the quantity
>
> $$q_s := \frac{||\mathbf{m}_0||}{||\mathbf{m}_{s-1}||}$$
>
> is called the corresponding *Beyer-quotient.*

Because the lengths of the vectors of a MRLB are uniquely determined only for $s \leq 6$, the same may apply to the Beyer-quotient; see Section B.2 for a discussion of this.
It is clear that $0 < q_s \leq 1$. Values of q_s close to 1 are considered 'good' since the corresponding unit cell is more cube-like.

The second figure of merit presented in (4.2.4), the spectral test, was introduced as 'the maximal distance $1/\nu_s$ of parallel hyperplanes which cover the LCG's lattice in $\mathbf{R}^s$'. For an LCG$(M, a, b, u_0) = (x_n)_{n=0}^{N-1}$ and $s \leq N$, we have the following[13]:

[12] Although Afflerbach's last algorithm, MB.n, works for arbitrary dimensions, it should be used only with greatest care and suspicion because it possibly produces wrong results: given any basis of an s-dimensional lattice Λ as input, the algorithm transforms it into supposedly equivalent bases until it ends up with a MRLB. Two sets of s linearly independent vectors span the same lattice Λ if and only if they are transformed to each other by a unimodular matrix (i.e. an integral matrix with determinant ± 1; see [46, Section 1.3, Theorem 7]). Due to an oral communication with Niederreiter [88], *non-unimodular* matrices can occur in the progress of algorithm MB.n, and so its result can be a basis spanning a lattice Λ' different from Λ.

[13] For $s > N$, the $N \leq s - 1$ points produced by the LCG are all contained in just one hyperplane in $\mathbf{R}^s$ and should therefore not be used anyway.

> Let Λ be the s-dimensional lattice of an LCG as in Proposition 5.1. Then Λ defines the so called dual lattice $\Lambda^{\star}$[14]. The value of the spectral test is
>
> $$\frac{1}{\nu_s} = \frac{1}{||\mathbf{m}||},$$
>
> where $\mathbf{m}$ is the shortest nonzero vector in $\Lambda^{\star}$.

Thus computing the spectral test amounts to finding the shortest vector in $\Lambda^{\star}$, which can be found as described in Dieter [25], Knuth [59, Section 3.3, Algorithm S], or Ripley [93, p.37] (a FORTRAN code for dimensions $s \leq 8$ is given in [93, Section B.3]).
We postpone the proof of the above statement to Section B.3 in the appendix. The reason is that the complete discussion of the spectral test is lengthy and not as much concerned with LCGs than with lattices in general. We avoid the detour of handling the spectral test here.

(5.2.6) Having proven the existence of a lattice structure in the output of an LCG, it is fairly easy to describe the specific long-range correlations, the critical distances, in its sequence. The following[15] is a consequence of understanding what Eichenauer-Hermann and Grothe showed in [33].

Proposition 5.2 *For an LCG $(x_n)_{n=0}^{N-1}$ of Type 1 to 3, the points*

$$(x_n, x_{n+s}) \qquad (0 \leq n < N)$$

have a lattice structure in $\mathbf{R}^2$.

Proof: Let $(x_n)_{n=0}^{M-1} = \mathrm{LCG}(M, a, b, 0)$ be of Type 1 (We prove this proposition only for this type. The proof for the remaining types is analogous). Since the LCG is purely periodic with period M and the proposition is trivial for $s = 0$, we can assume $0 < s < M$. Let $\mathbf{x}_n = (x_n, \ldots, x_{n+s})$ be the $(s+1)$-dimensional points obtained from the

[14] Λ does not only define $\Lambda^{\star}$. Given a lattice basis of Λ, a lattice basis of $\Lambda^{\star}$ can be computed; see (B.1.3)

[15] There is a slight generalization of this, whose statement and proof unfortunately require too many technical details from lattice theory to be stated here. The interested reader may consult Section B.4 or, most likely, the whole Appendix B.

LCG and let $\mathcal{P} = \{\mathbf{x}_n : 0 \leq n < M\}$ be the set of all these points. From Proposition 5.1, we know that

$$\mathcal{P} = [0, 1[^{s+1} \cap (\mathbf{x}_0 + \Lambda),$$

where the lattice Λ is spanned by the vectors

$$\begin{aligned} \mathbf{b}_0 &:= \frac{1}{M}(1, a, a^2, \ldots, a^s), \\ \mathbf{b}_j &:= \mathbf{e_j} \qquad (0 < j \leq s). \end{aligned}$$

To obtain the tuples (x_n, x_{n+s}), we use the $2 \times (s+1)$-matrix

$$T := \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$$

which defines a function from $\mathbf{R}^{s+1}$ to $\mathbf{R}^2$, mapping $\mathbf{x} \in \mathbf{R}^{s+1}$ to $T\mathbf{x}$. With this, we have

$$T\mathcal{P} = \{(x_n, x_{n+s}) : 0 \leq n < M\}.$$

The proof is complete if we can show that

$$\Lambda' := T\Lambda \qquad \text{is a lattice in } \mathbf{R}^2 \tag{2}$$

and that

$$T\mathcal{P} = [0, 1[^2 \cap (T\mathbf{x}_0 + \Lambda'). \tag{3}$$

To prove (2), observe the following. Λ is the set of all integer linear combinations of the $\mathbf{b}_i$ $(0 \leq i \leq s)$. T is linear, so $\Lambda' = T\Lambda$ is the set of all integer linear combinations of the $T\mathbf{b}_i$. Moreover, for $0 < i < s$, we have $T\mathbf{b}_i = (0, 0)$. Hence Λ' is the set of all integer linear combinations of $T\mathbf{b}_0 = 1/M(1, a^s)$ and $T\mathbf{b}_s = (0, 1)$. Because these two vectors are linearly independent, Λ' is a lattice in $\mathbf{R}^2$.

To prove (3), we show that the sets on each side of the equation are included in each other. For one, we have

$$\begin{aligned} T\mathcal{P} &= T\left([0, 1[^{s+1} \cap (\mathbf{x}_0 + \Lambda)\right) \\ &\subseteq T[0, 1[^{s+1} \cap T(\mathbf{x}_0 + \Lambda) \\ &= [0, 1[^2 \cap (T\mathbf{x}_0 + \Lambda'). \end{aligned}$$

The second inclusion needs some extra consideration. First note that $\mathbf{Z}^{s+1}$ is a subset[16] of Λ. Any point $\mathbf{z} \in \mathbf{R}^{s+1}$ can be written as the sum of two other points

$$\mathbf{z} = \lfloor \mathbf{z} \rfloor + \{\mathbf{z}\},$$

where the coordinates of $\lfloor \mathbf{z} \rfloor$ are all integers and the coordinates of $\{\mathbf{z}\}$ are all between 0 inclusive and 1 exclusive. From the definition of a lattice, we see that for any two points $\mathbf{p}, \mathbf{q} \in \Lambda$, their sum $\mathbf{p} + \mathbf{q}$ is as well contained in Λ. For any lattice point $\mathbf{p}$, we have $\lfloor \mathbf{x}_0 + \mathbf{p} \rfloor \in \mathbf{Z}^{s+1}$ and this yields

$$\begin{aligned} \{\mathbf{x}_0 + \mathbf{p}\} &= \mathbf{x}_0 + (\mathbf{p} - \lfloor \mathbf{x}_0 + \mathbf{p} \rfloor) \\ &\in \mathbf{x}_0 + \Lambda. \end{aligned}$$

Now we are ready for the second inclusion of (3). For

$$T\mathbf{x}_0 + \mathbf{x}' \in [0,1[^2 \cap (T\mathbf{x}_0 + \Lambda'),$$

there is a lattice point $\mathbf{p} \in \Lambda$ with

$$T(\mathbf{x}_0 + \mathbf{p}) = T\mathbf{x}_0 + \mathbf{x}'.$$

Choosing $\mathbf{x}$ such that

$$\mathbf{x}_0 + \mathbf{x} := \{\mathbf{x}_0 + \mathbf{p}\}$$

implies $\mathbf{x}_0 + \mathbf{x} \in [0,1[^{s+1} \cap (\mathbf{x}_0 + \Lambda)$. Since $T\mathbf{x}_0 + \mathbf{x}' \in [0,1[^2$, the first and the last coordinate of $\mathbf{x}_0 + \mathbf{p}$ are in $[0,1[$. This and the definition of $\mathbf{x}_0 + \mathbf{x} = \{\mathbf{x}_0 + \mathbf{p}\}$ yield

$$T(\mathbf{x}_0 + \mathbf{x}) = T\mathbf{x}_0 + \mathbf{x}',$$

which completes the proof. □

[16] The lattice $\mathbf{Z}^{s+1}$ is spanned by the standard ordered basis $\{\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_s\}$ of $\mathbf{R}^{s+1}$, so it is contained in Λ if the $\mathbf{e}_i$ are. For $i = 1, \ldots, s$, the $\mathbf{e}_i$ are of course in Λ because they are part of a lattice basis. For $i = 0$, we have

$$\mathbf{e}_0 = M\mathbf{b}_0 - \sum_{i=1}^{s} a^i \mathbf{b}_i$$

which is thus contained in Λ.

5.3 ICG

(5.3.1) The inversive congruential generator (ICG) is a recursive congruential generator using a recursion of order one, just like the LCG. But unlike linear generators, the ICG's recursion is based on a highly nonlinear function. For a (large) prime p and $a \in \mathbf{Z}_p$, let[17]

$$\overline{a} := \begin{cases} a^{-1} & : \quad a \in \mathbf{Z}_p^\star, \\ 0 & : \quad a = 0 \end{cases}$$

in $\mathbf{Z}_p$.

Definition 5.3 *Let p be a (large) prime and $a, b, u_0 \in \mathbf{Z}_p$. The* inversive congruential generator*, ICG, for short, with parameters p, a, b, and u_0 defines a sequence $(u_n)_{n \geq 0}$ in $\mathbf{Z}_p$ by*

$$u_n := a \cdot \overline{u_{n-1}} + b \qquad (n > 0)$$

and a sequence $(x_n)_{n \geq 0}$ of random numbers in $[0, 1[$ *by*

$$x_n := \frac{u_n}{p} \qquad (n \geq 0).$$

(5.3.2) An ICG is of course periodic for the same reason an as LCG. The same considerations as in (5.2.1) apply to the ICG. In particular, we only consider ICG which are purely periodic with some period length N, and for these ICG, we agree to use only the first N numbers $(x_n)_{n=0}^{N-1}$. And as before, it seems desirable to choose the ICG's parameters as to maximize its period length.
We denote the sequence $(u_n)_{n=0}^{p-1}$ by $\mathrm{icg}(p, a, b, u_0)$ and the sequence $(x_n)_{n=0}^{p-1}$ by $\mathrm{ICG}(p, a, b, u_0)$.

A necessary and sufficient condition under which $\mathrm{ICG}(p, a, b, u_0)$ has the maximal period length $N = p$ is given by Flahive and Niederreiter in [40]. This necessary and sufficient condition, and even a just sufficient one given in [31] or [87, Theorem 8.4], are such that their statement and their proof require some background in the theory of finite fields that we do not want

[17] Note that we are not computing in $\mathbf{R}$ but in the finite field $\mathbf{Z}_p$ now. As pointed out in (5.1.3), for any $a \in \mathbf{Z}_p^\star$, there is a unique element $a^{-1} \in \mathbf{Z}_p^\star$ such that $aa^{-1} = 1$ in $\mathbf{Z}_p$.

to introduce here. In the following, whenever we refer to a generator as an ICG, we simply assume it has maximal period p. Moreover, since we only present results for the totality $\{u_0, \ldots, u_{p-1}\} = \mathbf{Z}_p$ of an ICG's output, we additionally assume $u_0 = 0$. Thus, when talking about an ICG, we always assume it has the form

$$\mathrm{ICG}(p, a, b, 0) = (x_n)_{n=0}^{p-1}.$$

A description of algorithms to compute the parameters for maximal period ICGs is given by Hellekalek in [51]. For $p \leq 2^{31} - 1$ being a Mersenne prime, tables of parameters for maximal period ICGs are presented by Hellekalek, Mayer, and Weingartner in [52].

(5.3.3) To study the structure of the s-dimensional points $\mathbf{x}_n = (x_n, \ldots, x_{n+s-1})$ of an ICG, we switch to the corresponding integer s-tuples

$$\mathbf{u}_n = (u_n, \ldots, u_{n+s-1}),$$

which we view as points not in $\mathbf{R}^s$ but in the finite vector space $\mathbf{Z}_p^s$. We will find this to be technically convenient. As in (5.2.1), we assume that the indices of the u_n and x_n are reduced modulo the generator's period p.

Switching from $\mathbf{x}_n$ to $\mathbf{u}_n$ does not really matter since they differ only by a constant scaling factor p; any structure of the $\mathbf{x}_n$ is inherited by the $\mathbf{u}_n$, and vice versa. The change from $\mathbf{R}^s$ to $\mathbf{Z}_p^s$ however, from an infinite vector space to a finite one with different arithmetical structure, needs some explanation. $\mathbf{Z}_p^s$ is contained in $\mathbf{R}^s$ because we have

$$\mathbf{Z}_p^s = [0, p[^s \cap \mathbf{Z}^s.$$

A hyperplane $H_{\mathbf{n},c}$ in $\mathbf{R}^s$, defined by a nonzero vector $\mathbf{n} \in \mathbf{R}^s$ and a constant $c \in \mathbf{R}$, is the set of those points $\mathbf{x} \in \mathbf{R}^s$ for which[18] $<\mathbf{n}, \mathbf{x}> = c$. A hyperplane in, say, $\mathbf{Z}_p^2$ is a discrete set of integer points. If we picture these as points in $\mathbf{R}^2$, we find that they all lie on one straight line which is folded when it leaves the square $[0, p[^2$. More generally, a hyperplane in $\mathbf{Z}_p^s$

[18] For two points $\mathbf{x} = (x_0, \ldots, x_{s-1})$ and $\mathbf{y} = (y_0, \ldots, y_{s-1})$ in $\mathbf{R}^s$, we denote their inner product by $<\mathbf{x}, \mathbf{y}>$:

$$<\mathbf{x}, \mathbf{y}> := \sum_{i=0}^{s-1} x_i y_i.$$

corresponds[19] to a finite number of parallel hyperplanes in $\mathbf{R}^s$. Conversely, any hyperplane in $\mathbf{R}^s$ which contains at least s affinely independent points from $[0,p[^s \cap \mathbf{Z}^s$ corresponds[20] to a single hyperplane in $\mathbf{Z}_p^s$.

We are about to show that the s-dimensional points of an ICG do literally *avoid* the hyperplanes in $\mathbf{Z}_p^s$. Due to the above correspondence of hyperplanes in $\mathbf{Z}_p^s$ an $\mathbf{R}^s$, hyperplanes in $\mathbf{R}^s$ are also avoided. In this sense, switching from $\mathbf{R}^s$ to $\mathbf{Z}_p^s$ is meaningful. The technical convenience gained by changing to $\mathbf{Z}_p^s$ will become apparent in the forthcoming proofs.

(5.3.4) To prove that the $\mathbf{u}_n$ avoid the hyperplanes in $\mathbf{Z}_p^s$, we use a technique due to Niederreiter [87, Theorem 8.6]. We show that for most of these points, the condition

> $\mathbf{u}_n$ is contained in a given hyperplane

can be translated to

> n is root of a corresponding nonzero polynomial of degree s over $\mathbf{Z}_p$.

Since such a polynomial has at most s roots in $\mathbf{Z}_p$, it will follow that at most s of these points are contained in any given hyperplane.

The primary device to translate between the above conditions for a full period $\mathrm{icg}(p,a,b,0) = (u_n)_{n=0}^{p-1}$ is the mapping

$$\begin{array}{rcl} P : \mathbf{Z}_p & \longrightarrow & \mathbf{Z}_p \\ x & \longmapsto & a\overline{x} + b \end{array}$$

[19] For $\mathbf{n} \in \mathbf{Z}_p^s \setminus \{\mathbf{0}\}$ and $c \in \mathbf{Z}_p$, the hyperplane

$$\{\mathbf{x} \in \mathbf{Z}_p^s : \ <\mathbf{n},\mathbf{x}> = c \text{ (in } \mathbf{Z}_p)\}$$

in $\mathbf{Z}_p^s$ corresponds to those hyperplanes $H_{\mathbf{n},c+kp}$ $(k \in \mathbf{Z})$ in $\mathbf{R}^s$ which intersect $[0,p[^s \cap \mathbf{Z}^s$.

[20] Let $H_{\mathbf{n},c}$ be a hyperplane in $\mathbf{R}^s$ containing s affinely independent points $\mathbf{p}_0,\ldots,\mathbf{p}_{s-1}$ from $[0,p[^s \cap \mathbf{Z}_p^s$. The $\mathbf{p}_i$ are contained in the vector space $\mathbf{Z}_p^s$, and it is easy to see that they are affinely independent in $\mathbf{Z}_p^s$. Hence there is a uniquely determined hyperplane $H_{\mathbf{n}',c'}$ in $\mathbf{Z}_p^s$ which contains all the $\mathbf{p}_i$. As we have seen above, this hyperplane in $\mathbf{Z}_p^s$ corresponds to some of the hyperplanes $H_{\mathbf{n}',c'+kp}$ $(k \in \mathbf{Z})$ in $\mathbf{R}^s$. Since we assumed that the $\mathbf{p}_i$ are covered by *one* hyperplane $H_{\mathbf{n},c}$ in $\mathbf{R}^s$, we can conclude that for one integer k', we have $H_{\mathbf{n},c} = H_{\mathbf{n}',c+k'p}$.

and its iterates

$$P^j(x) := \left\{ \begin{array}{rcl} x & : & j = 0, \\ P(P^{j-1}(x)) & : & j > 0. \end{array} \right.$$

For these, we have[21]

Lemma 5.1 *Let* $(u_n)_{n=0}^{p-1} = icg(p, a, b, 0)$ *be a full period ICG. If* $1 \leq j < p$ *and*

$$x \neq -a\overline{u_i} \qquad (0 \leq i < j)$$

in $\mathbf{Z}_p$*, then*[22]

$$P^j(x) = u_j \frac{x + a\overline{u_j}}{x + a\overline{u_{j-1}}}.$$

Proof: We use induction on j. Throughout the proof, we always compute in $\mathbf{Z}_p$. Let $j = 1$ and $x \neq a\overline{u_0}$ (which means $x \neq 0$ since we chose $u_0 = 0$).

With this, the definition of P and the basic recurrence of the ICG yield

$$\begin{array}{rcl} P^1(x) & = & b + a\overline{x} \\ & = & \dfrac{bx + a}{x} \\ & = & b\dfrac{x + a\overline{b}}{x} \\ & = & u_1\dfrac{x + a\overline{u_1}}{x + a\overline{u_0}}. \end{array}$$

Now suppose the induction hypothesis holds for any integer between 1 and j inclusive with $1 < j + 1 < p$. Let $x \in \mathbf{Z}_p$ such that

$$x \neq -a\overline{u_i} \qquad (0 \leq i < j + 1).$$

Before we can apply the induction hypothesis to $P^{j+1}(x) = P^j(P(x))$, we have to verify its applicability for $P(x)$. By the choice of x, we have

$$a\overline{x} + b \neq -u_i + b \qquad (0 \leq i < j + 1).$$

[21] See [87, Equation (8.6)].

[22] Instead of ab^{-1}, we also write $\frac{a}{b}$.

The left side of this equation is equal to $P(x)$. The definition of $u_i = a\overline{u_{i-1}} + b$ yields that the right side is $-u_i + b = -a\overline{u_{i-1}}$. This means

$$P(x) \neq -a\overline{u_i} \qquad (0 \leq i < j),$$

and the induction hypothesis is applicable for $P(x)$:

$$\begin{aligned} P^{j+1}(x) &= P^j(P(x)) \\ &= u_j \frac{a\overline{x} + b + a\overline{u_j}}{a\overline{x} + b + a\overline{u_{j-1}}} \\ &= u_j \frac{a\overline{x} + u_{j+1}}{a\overline{x} + u_j} \\ &= u_j \frac{a + u_{j+1}x}{a + u_j x} \\ &= \frac{a + u_{j+1}x}{a\overline{u_j} + x} \\ &= u_{j+1} \frac{x + a\overline{u_{j+1}}}{x + a\overline{u_j}}. \end{aligned}$$

Hence the induction hypothesis holds for $j + 1$, too. □

(5.3.5) With Lemma 5.1, it is fairly easy to prove that ICGs avoid the planes, as was originally shown by Eichenauer-Hermann in [32] (the proof given here stems from Niederreiter [87, p.184]).

Proposition 5.3 *Let $s \geq 2$ and let* $\mathrm{ICG}(p, a, b, u_0)$ *be a full period ICG. Then any hyperplane in $\mathbf{Z}_p^s$ contains at most s of the points*

$$\mathbf{u}_n = (u_n, \ldots, u_{n+s-1}) \qquad (0 \leq n < p)$$

for which the first $s - 1$ coordinates are nonzero.

The statement is trivial for $s \geq p$ even without the condition of nonzero coordinates, because the ICG can produce only $p \leq s$ distinct points. Otherwise, $s - 1$ of those $\mathbf{u}_0, \ldots, \mathbf{u}_{p-1}$ are excluded from the proposition for which $\mathbf{x}_n = \mathbf{u}_n/p$ lies on one of the 'left-sided' faces of $[0,1[^s$. This is not much of a restriction if s is small. If s is large and, in particular, if s is close to p, all except just $p - s$ of the $\mathbf{x}_n$ lie on the 'left-sided' faces of $[0,1[^s$; in this case, you probably would not want to use the given ICG anyway and prefer a generator with longer period instead.

Proof: Throughout the proof, we always compute in $\mathbf{Z}_p$. The proposition trivially holds for $s \geq p$ as noted above; therefore let $s < p$. Without loss of generality, we assume $u_0 = 0$. The ICG$(p, a, b, 0)$ has the full period p, so the sequence $(u_n)_{n=0}^{p-1}$ visits every element of $\mathbf{Z}_p$ exactly once. Hence

$$\begin{aligned} &\{(u_n, u_{n+1}, \ldots, u_{n+s-1}) : \ 0 \leq n < p\} \\ = &\left\{(P^0(x), P^1(x), \ldots, P^{s-1}(x)) : \ x \in \mathbf{Z}_p\right\}. \end{aligned}$$

Let us see how the condition 'the first $s-1$ coordinates of $\mathbf{u}_n$ are nonzero' can be formulated for the second set above. A point in the second set has a nonzero first coordinate if $P^0(x) = x \neq 0$; by the choice of $u_0 = 0$, this means $x \neq -a\overline{u_0}$. If we have $x \neq -a\overline{u_0}$, then Lemma 5.1 yields that $P^1(x) \neq 0$ if $x \neq -a\overline{u_1}$. If we have $x \neq -a\overline{u_0}$ and $x \neq -a\overline{u_1}$, then Lemma 5.1 yields that $P^2(x) \neq 0$ if $x \neq -a\overline{u_2}$. Proceeding this way, we get that a point $(P^0(x), \ldots, P^{s-1}(x))$ in the second set has its first $s-1$ coordinates nonzero if and only if $x \neq -a\overline{u_i}$ for $0 \leq i < s-1$.

We have to show that the set of points

$$\mathcal{P} := \{(u_n, \ldots, u_{n+s-1}) : \ 0 \leq n < p, \ u_n, \ldots, u_{n+s-2} \neq 0\}$$

intersects any given hyperplane in $\mathbf{Z}_p^s$ in at most s points. With the above considerations, we have

$$\mathcal{P} = \left\{(P^0(x), \ldots, P^{s-1}(x)) : \ x \in \mathbf{Z}_p \setminus \{-a\overline{u_0}, \ldots, -a\overline{u_{s-2}}\}\right\}.$$

Now let $\mathbf{b} = (b_0, \ldots, b_{s-1}) \in \mathbf{Z}_p^s \setminus \{\mathbf{0}\}$ and $c \in \mathbf{Z}_p$. A point $(P^0(x), \ldots, P^{s-1}(x)) \in \mathcal{P}$ lies on the hyperplane defined by $\mathbf{b}$ and c if and only if

$$\sum_{i=0}^{s-1} b_i P^i(x) - c = 0.$$

Lemma 5.1 is applicable to the coordinates of any point in $\mathcal{P}$. With this, the above hyperplane-equation translates to

$$b_0 x + \sum_{i=1}^{s-1} b_i u_i \frac{x + a\overline{u_i}}{x + a\overline{u_{i-1}}} - c = 0.$$

Clearing denominators, this is equivalent to $h(x) = 0$, where h is a polynomial over $\mathbf{Z}_p$:

$$h(x) = (b_0 x - c) \prod_{i=1}^{s-1} (x + a\overline{u_{i-1}}) + \sum_{i=1}^{s-1} b_i u_i (x + a\overline{u_i}) \prod_{\substack{j = 1 \\ j \neq i}}^{s-1} (x + a\overline{u_{j-1}}).$$

We will show that h has at most s roots in $\mathbf{Z}_p$, which essentially completes the proof: in the above representation of $\mathcal{P}$, every point is uniquely identified by a particular $x \in \mathbf{Z}_p \setminus \{-a\overline{u_0}, \ldots, -a\overline{u_{s-2}}\}$. If h has at most s roots in $\mathbf{Z}_p$, it has of course at most s roots in $\mathbf{Z}_p \setminus \{-a\overline{u_0}, \ldots, -a\overline{u_{s-2}}\}$, so at most s points of $\mathcal{P}$ lie on the hyperplane defined by $\mathbf{b}$ and c.

The degree $\deg(h)$ of h is at most s, so we are finished if h is not the zero-polynomial. If h were the zero-polynomial, the coefficient b_0 of x^s would be zero. Then, for $1 \le k \le s-1$, we would have

$$\begin{aligned} h(-a\overline{u_{k-1}}) &= b_k u_k(-a\overline{u_{k-1}} + a\overline{u_k}) \prod_{\substack{j=1 \\ j \ne k}}^{s-1} (-a\overline{u_{k-1}} + a\overline{u_{j-1}}) \\ &= 0. \end{aligned}$$

From $u_0 = 0$ and $1 \le k \le s-1 < p$, we conclude that $u_k \ne 0$. For $i \ne j$ in $\mathbf{Z}_p$, we have $u_i \ne u_j$. Therefore $-a\overline{u_{k-1}} + a\overline{u_k} \ne 0$ and

$$-a\overline{u_{k-1}} + a\overline{u_{j-1}} \ne 0 \qquad (j = 1, \ldots, s-1,\ j \ne k).$$

With this, all terms of $h(-a\overline{u_{k-1}}) = 0$ except b_k are nonzero. Thus, for $1 \le k \le s-1$, we have $b_k = 0$. We get that the vector $\mathbf{b} = (b_0, \ldots, b_{s-1})$ is equal to the zero vector $\mathbf{0}$, which contradicts the fact that we have chosen $\mathbf{b} \in \mathbf{Z}_p^s \setminus \{\mathbf{0}\}$ in the first place. So h is not the zero-polynomial. □

(5.3.6) The avoidance of hyperplanes by the s-dimensional points of an ICG does not strictly imply the absence of long-range correlations. We show at least the absence of critical distances for reasonably small shifts s.

Proposition 5.4 *Let* $\mathrm{ICG}(p, a, b, u_0)$ *be a full period ICG. Then any hyperplane in* $\mathbf{Z}_p^2$ *contains at most two of the points*

$$(u_n, u_{n+s}) \qquad (0 \le n < p)$$

for which $u_n, u_{n+1}, \ldots, u_{n+s-1}$ *are all nonzero.*

Proof: We will proceed as in the proof of Proposition 5.3. Without loss of generality, we assume $u_0 = 0$. Since the sequence is purely periodic with period p, we can assume $s < p$. Moreover, to avoid trivialities, let $1 < s < p$.

The proposition states that any hyperplane defined by $\mathbf{b} = (b_0, b_1) \in \mathbf{Z}_p^2 \setminus \{\mathbf{0}\}$ and $c \in \mathbf{Z}_p$ contains at most two points from

$$\begin{aligned} \mathcal{P} &:= \{(u_n, u_{n+s}) : \, 0 \leq n < p, \, 0 \notin \{u_n, u_{n+1}, \ldots, u_{n+s-1}\}\} \\ &= \left\{(P^0(x), P^s(x)) : \, x \in \mathbf{Z}_p \setminus \{-a\overline{u_0}, \ldots, -a\overline{u_{s-1}}\}\right\}. \end{aligned}$$

A point $(P^0(x), P^s(x))$ from $\mathcal{P}$ lies on the hyperplane defined by $\mathbf{b}$ and c if

$$b_0 P^0(x) + b_1 P^s(x) - c = 0.$$

Lemma 5.1 holds for the coordinates of every point from $\mathcal{P}$, so the above condition translates to

$$b_0 x + b_1 u_s \frac{x + a\overline{u_s}}{x + a\overline{u_{s-1}}} - c = 0.$$

Clearing denominators, we get that x is a root of the polynomial

$$h(x) = (b_0 x - c)(x + a\overline{u_{s-1}}) + b_1 u_s (x + a\overline{u_s})$$

over $\mathbf{Z}_p$. Since $\deg(h) \leq 2$, h has at most 2 roots in $\mathbf{Z}_p$ and therefore at most 2 roots in $\mathbf{Z}_p \setminus \{-a\overline{u_0}, \ldots, -a\overline{u_{s-1}}\}$ – if it is not the zero-polynomial.

Suppose h were the zero-polynomial. Then the coefficient b_0 of x^2 would be zero. Moreover, we would have

$$\begin{aligned} h(-a\overline{u_{s-1}}) &= b_1 u_s(-a\overline{u_{s-1}} + a\overline{u_s}) \\ &= 0. \end{aligned}$$

As before, this would yield that b_1 is zero, too. We would get that $\mathbf{b}$, which was chosen from $\mathbf{Z}_p^s \setminus \{\mathbf{0}\}$, is equal to $\mathbf{0}$. This is a contradiction, so h is not the zero-polynomial. □

5.4 EICG

(5.4.1) The last generator we will study in detail is the explicit inversive congruential generator. It uses the same inversion function as the ICG does, but it is a *nonrecursive* generator.

Definition 5.4 *Let p be a (large) prime and let $a, b, n_0 \in \mathbf{Z}_p$. The* explicit inversive congruential generator*, EICG, for short, with parameters p, a, b, and n_0 defines a sequence $(u_n)_{n \geq 0}$ in $\mathbf{Z}_p$ by*

$$u_n := \overline{a \cdot (n_0 + n) + b} \qquad (n \geq 0)$$

and a sequence $(x_n)_{n \geq 0}$ of random numbers in $[0, 1[$ *by*

$$x_n := \frac{u_n}{p} \qquad (n \geq 0).$$

(5.4.2) Choosing parameters p, a, b, and n_0 to obtain the maximal possible period length is particularly easy. If only p is prime and $a \neq 0$ in $\mathbf{Z}_p$, any choice of b and n_0 gives an EICG with period p producing $\{u_0, \ldots, u_{p-1}\} = \mathbf{Z}_p$. The reason for this is that due to the group structure of $\mathbf{Z}_p$, the function $f(n) := \overline{a \cdot (n_0 + n) + b}$ (being composed of bijective functions on $\mathbf{Z}_p$) is itself a bijection on $\mathbf{Z}_p$. Hence, for $0 \leq n < p$, the resulting values $f(n) = u_n$ are all distinct. For $n \geq p$, we have $n \bmod p = n'$ for some $n' \in \mathbf{Z}_p$, and hence $u_n = u_{n'}$. In the following, we only consider maximal period EICGs. For $a \in \mathbf{Z}_p^\star$ and $b, n_0 \in \mathbf{Z}_p$, we denote the sequence $(u_n)_{n=0}^{p-1}$ by $\mathrm{eicg}(p, a, b, n_0)$ and the sequence $(x_n)_{n=0}^{p-1}$ by $\mathrm{EICG}(p, a, b, n_0)$.

(5.4.3) Since virtually any choice of parameters defines a full period EICG, there seems to exist quite a lot of different EICGs for a given prime modulus p. We will see that this is not exactly true. First of all, for $a \neq 0$, the EICGs $\mathrm{EICG}(p, a, b, 0)$ and $\mathrm{EICG}(p, a, b, n_0)$ differ only by a cyclic permutation. Therefore, considering the totality $\{x_0, \ldots, x_{p-1}\}$ produced by $\mathrm{EICG}(p, a, b, n_0)$, we can always assume $n_0 = 0$.

Next, we show that one of the EICG's parameters is redundant[23].

Lemma 5.2 *Let $a \in \mathbf{Z}_p^\star$ be fixed. For any $b \in \mathbf{Z}_p$, we have*

$$\mathrm{eicg}(p, a, b, 0) = \mathrm{eicg}(p, a, 0, \overline{a}b).$$

Conversely, for any $n_0 \in \mathbf{Z}_p$, we have

$$\mathrm{eicg}(p, a, 0, n_0) = \mathrm{eicg}(p, a, n_0 a, 0).$$

[23] The idea for this stems from Karl Entacher's observation that the 2-dimensional scatter-plots of two EICGs with the same p, a, and n_0 but different values of b look exactly alike. The same idea is implicitly contained in Niederreiter's [85, p.5].

For any choice of $b \neq 0$ and $n_0 \neq 0$, the resulting $\mathrm{eicg}(p, a, b, n_0)$ differs from $\mathrm{eicg}(p, a, 0, 0)$ by just a cyclic permutation, and any cyclic permutation of $\mathrm{eicg}(p, a, 0, 0)$ is equal to $\mathrm{eicg}(p, a, b, 0)$ for some b. The reason we introduced the EICG including the redundant parameter in the first place is technical and will soon become apparent.

Proof: All calculations in the proof are carried out in $\mathbf{Z}_p$. Let $a \in \mathbf{Z}_p^\star$, $b \in \mathbf{Z}_p$, let $(u_n)_{n=0}^{p-1} = \mathrm{eicg}(p, a, b, 0)$, and $(v_n)_{n=0}^{p-1} = \mathrm{eicg}(p, a, 0, \overline{a}b)$. With this, we have

$$\begin{aligned} u_0 &= \overline{a \cdot (0) + b} \\ &= \overline{b} \\ &= \overline{a(\overline{a}b) + 0} \\ &= v_0. \end{aligned}$$

Note that u_n can be computed recursively as $u_n = \overline{\overline{u_{n-1}} + a}$ and the same holds for v_n, too. So both the u_n and the v_n depend on their predecessors by a recursion of order 1. We get $u_1 = v_1$ and, inductively, $(u_n)_{n=0}^{p-1} = (v_n)_{n=0}^{p-1}$. This proves the first part of the lemma.
Proving the second part is equally elementary. □

Finally, there is an obvious relation between the sequences $\mathrm{eicg}(p, a, 0, 0)$ and $\mathrm{eicg}(p, 1, 0, 0)$.

Lemma 5.3 *Let $a \in \mathbf{Z}_p^\star$. The sequence* $\mathrm{eicg}(p, a, 0, 0)$ *is obtained by selecting every a-th element from* $\mathrm{eicg}(p, 1, 0, 0)$.

Remark: As we did for the LCG and ICG, we implicitly assume that the index n of the numbers x_n and u_n produced by an EICG are reduced modulo p. Hence selecting every a-th element from $\mathrm{eicg}(p, 1, 0, 0) = (u_n)_{n=0}^{p-1}$ means selecting $u_0, u_{a \bmod p}, u_{2a \bmod p}, \ldots, u_{(p-1)a \bmod p}$.

Proof: Let $(u_n)_{n=0}^{p-1} = \mathrm{eicg}(p, a, 0, 0)$ and $(v_n)_{n=0}^{p-1} = \mathrm{eicg}(p, 1, 0, 0)$. We have

$$\begin{aligned} u_n &= \overline{an} \\ &= \overline{1 \cdot (an)} \\ &= v_{an}. \end{aligned}$$

in $\mathbf{Z}_p$. □

With these observations, for any $a \in \mathbf{Z}_p^\star$ and $b, n_0 \in \mathbf{Z}_p$, the sequence $\mathrm{eicg}(p, a, b, n_0)$ is obtained from $\mathrm{eicg}(p, 1, 0, 0)$ as follows:

- select every a-th element from $\mathrm{eicg}(p, 1, 0, 0)$, and
- to the selected elements, apply a cyclic permutation whose 'shift' is determined by b and n_0.

(5.4.4) Lemma 5.2 and 5.3 show that all maximal period EICGs for a given prime modulus p are closely related; more precisely, they show that every $\mathrm{EICG}(p, a, b, n_0)$ is obtained from $\mathrm{EICG}(p, 1, 0, 0)$ by selecting and then rotating a subsequence of equal stride. In this sense, for a given prime p, the corresponding EICGs are *linearly* related[24] to each other. This observation gives us a new way to look at the s-tuples $\mathbf{u}_n = (u_n, u_{n+1}, \ldots, u_{n+s-1})$ obtained from an $\mathrm{eicg}(p, a, 0, 0) = (u_n)_{n=0}^{p-1}$.

> The i-th coordinate u_{n+i} of $\mathbf{u}_n$ is just the n-th element of the sequence $\mathrm{eicg}(p, a, ia, 0) = (u_{n+i})_{n=0}^{p-1}$.

The result we are about to state concerns a more general form of s-tuples. We consider the s EICGs

$$(u_n^{(i)})_{n=0}^{p-1} = \mathrm{eicg}(p, a_i, b_i, 0) \qquad (0 \leq i < s)$$

with $a_i \in \mathbf{Z}_p^{\star}$ to construct the s-tuples

$$\mathbf{u}_n := (u_n^{(0)}, \ldots, u_n^{(s-1)}) \qquad (0 \leq n < p).$$

The following result for these general s-tuples is due to Niederreiter [85].

Proposition 5.5 *If $\overline{a_0}b_0, \ldots, \overline{a_{s-1}}b_{s-1}$ are mutually different in $\mathbf{Z}_p$, then any hyperplane in $\mathbf{Z}_p^s$ contains at most s of the $\mathbf{u}_n$ whose coordinates are nonzero.*
If the hyperplane contains $\mathbf{0}$, then in contains at most $s-1$ of these points.

This is the strongest form of nonlinearity we have encountered so far: take the sequence $\mathrm{eicg}(p, 1, 0, 0) = (u_n)_{n=0}^{p-1}$ and form s new sequences $(u_n^{(i)})_{n=0}^{p-1}$, each by selecting and then rotating an equal-strided subsequence of $(u_n)_{n=0}^{p-1}$. If only the s new sequences are all formed in a *different way*[25], the resulting s-tuples $\mathbf{u}_n$ avoid the hyperplanes in $\mathbf{Z}_p^s$ (except for at most s points which

[24] Refer back to Definition 5.4 for the cause of this linear relationship.

[25] This is what the condition on $\overline{a_0}b_0, \ldots, \overline{a_{s-1}}b_{s-1}$ in Proposition 5.5 translates to.

are excluded from the proposition). The same considerations as in (5.3.5) apply to these excluded points.

Proof: We will proceed as we did in the proof of Proposition 5.3. Throughout this proof, all calculations are performed in $\mathbf{Z}_p$. Let the $\overline{a_0}b_0, \ldots, \overline{a_{s-1}}b_{s-1}$ be mutually different.

A hyperplane $H_{\mathbf{b},c}$ in $\mathbf{Z}_p^s$ is uniquely identified by a vector $\mathbf{b} = (b_0, \ldots, b_{s-1}) \in \mathbf{Z}_p^s \setminus \{\mathbf{0}\}$ and a scalar $c \in \mathbf{Z}_p$. Due to the definition of the $u_n^{(i)}$, the coordinates of $\mathbf{u}_n$ are nonzero if and only if

$$n \notin \{-\overline{a_0}b_0, \ldots, -\overline{a_{s-1}}b_{s-1}\}.$$

For such n, we have $\mathbf{u}_n \in H_{\mathbf{b},c}$ if and only if

$$\begin{aligned} 0 &= c - \sum_{i=0}^{s-1} b_i u_n^{(i)} \\ &= c - \sum_{i=0}^{s-1} \frac{b_i}{a_i n + b_i}. \end{aligned}$$

Clearing denominators, we get that n is a root of the polynomial

$$h(x) = c \prod_{i=0}^{s-1} (a_i x + b_i) - \sum_{i=0}^{s-1} b_i \prod_{\substack{j=0 \\ j \neq i}}^{s-1} (a_j x + b_j).$$

If $c \neq 0$, then h is a nonzero polynomial[26] of degree s over $\mathbf{Z}_p$. Since such h has at most s roots in $\mathbf{Z}_p$, the hyperplane $H_{\mathbf{b},c}$ contains at most s of the $\mathbf{u}_n$ with $n \notin \{-\overline{a_0}b_0, \ldots, -\overline{a_{s-1}}b_{s-1}\}$.

If $c = 0$, i.e. if $\mathbf{0} \in H_{\mathbf{b},c}$, we get

$$h(x) = -\sum_{i=0}^{s-1} b_i \prod_{\substack{j=0 \\ j \neq i}}^{s-1} (a_j x + b_j),$$

[26] To avoid trivial sequences, we assumed all the a_i are nonzero in (5.4.2). Therefore the coefficient of x^s in the above equation is nonzero if c is.

whose degree is at most $s-1$. It remains to show that h is not the zero-polynomial. $\mathbf{b}$ is not the zero-vector, so one of its coordinates is nonzero. For $b_k \neq 0$, we have

$$\begin{aligned} h(-\overline{a_k}b_k) &= -b_k \prod_{\substack{j=0\\ j\neq k}}^{s-1} (-a_j\overline{a_k}b_k + b_j) \\ &= b_k \prod_{\substack{j=0\\ j\neq k}}^{s-1} a_j(\overline{a_k}b_k - \overline{a_j}b_j). \end{aligned}$$

b_k is nonzero by the choice of k and the a_j are all nonzero by assumption; finally, the terms $(\overline{a_k}b_k - \overline{a_j}b_j)$ were all assumed to be nonzero, so h is not the zero-polynomial. □

(5.4.5) With Proposition 5.5, it is fairly obvious that the points $(x_n, \ldots, x_{n+s-1})$ produced by the EICG$(p, a, 0, 0) = (x_n)_{n=0}^{p-1}$ have no lattice structure. Applying the proposition to the special case

$$\begin{aligned} a_i &= a \qquad (0 \leq i < s), \\ b_i &= i \cdot a \qquad (0 \leq i < s) \end{aligned}$$

yields that the points $\mathbf{x}_n = (x_n, \ldots, x_{n+s-1})$ avoid the hyperplanes in $\mathbf{R}^s$ (with s exceptions). The restriction to EICG$(p, a, 0, 0)$ does not really matter. The sequence EICG$(p, a, 0, 0)$ differs from EICG(p, a, b, n_0) just by a cyclic permutation and so do the corresponding s-tuples.

(5.4.6) In a similar vein, Proposition 5.5 is used to prove the absence of critical distances in the points (x_n, x_{n+s}) formed from EICG$(p, a, 0, 0) = (x_n)_{n=0}^{p-1}$. For any [27] $0 < s < p$, the special case

$$\begin{aligned} a_0 &= a, & a_1 &= a, \\ b_0 &= 0, & b_1 &= s \cdot a \end{aligned}$$

yields that the points (x_n, x_{n+s}) avoid the lines in $\mathbf{R}^2$ (with just two points being excluded from the proposition).

[27] As before, the case $s = 0$ is trivial and the case $s \geq p$ can be reduced to $s' < p$ with $s' = s \bmod p$.

With this, the EICG does not exhibit the special kind of of long-range correlation inherent to the LCG. Although a similar result was shown for the ICG in Proposition 5.4, the present result for the EICG is stronger. For an ICG, all except s of the points (x_n, x_{n+s}) avoid the lines in $\mathbf{R}^2$; for an EICG, we have the same for all but just two of these points.

5.5 Other generators

(5.5.1) So far, we have presented three generators in detail – the LCG being one of the oldest and *the* most widely used, and the ICG and EICG being new generators. These three types were chosen because the LCG's defects are thoroughly explored and because the inversive generators can be proven to lack just these defects[28]. A number of other random number generators in use today had to be excluded from this text[29]. Many of them have both theoretically and practically desirable as well as undesirable properties which, though sometimes different in the details, are similar to those we encountered so far. We refer the reader to the surveys [3], [30], [65], [69], [86], [94], or [95]. A more detailed discussion of specific random number generators is given by Afflerbach in [1, Chapter 3], Knuth in [59, Section 3.2.1 and 3.2.2], Niederreiter in [87, Chapter 4, 7 – 10], Ripley in [93, Section 2.2, 2.3, and 2.5], and Weingartner in [113].

[28] Regrettably, we know of no exploration of the defects inherent to inversive generators (which do of course exist; see Chapter 2 and 3). We wonder what they are and for which sorts of simulation they might be considered relevant (in the sense of Section 4.2).

[29] The main reason for this is the fact that we have limited our study to those generators for which the most qualitative results were available.

Appendix A

Expectation and convexity

A.1 The idea

(A.1.1) Most introductory textbooks on probability note that the expectation of certain random variables (as, say, events) can be pictured as the barycenter or center of gravity obtained by distributing mass among the possible values proportional to their probabilities.

The center of gravity of a finite number of points is usually defined as follows[1].

Definition A.1 *Let* $\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}$ *be points in* $\mathbf{R}^s$. *If a total mass of* $\lambda > 0$ *is distributed among them such that* $\mathbf{x}_i$ *is assigned the mass* $\lambda_i \geq 0$, *then*

$$\mathbf{x} := \frac{1}{\lambda} \sum_{i=0}^{n-1} \lambda_i \mathbf{x}_i$$

is the corresponding center of gravity *of the points* $\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}$.

It is clear that the center of gravity $\mathbf{x}$ does not change when we replace λ by $\lambda/\lambda = 1$ and λ_i by λ_i/λ; hence we can assume without loss of generality that $\lambda = \sum_{i=0}^{n-1} \lambda_i = 1$. Doing so, the distribution of mass becomes a probability distribution and the center of gravity $\mathbf{x}$ becomes the corresponding

[1] For our purpose, it suffices to use this special case of the more general center of gravity as defined by Leichtweiß in [74, Definition 3.3].

expectation. There is yet another name for this $\mathbf{x}$: in the theory of convex sets, it is called a convex combination of the points $\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}$.

For a finite number of points, we have found three concepts from quite different areas which all denote the same thing: the physical concept of 'center of gravity', the probabilistic concept of 'expectation', and the geometric concept of 'convex combination'. In this chapter, we investigate the extent of this mutual equivalence.

(A.1.2) The set of all possible convex combinations which can be formed from a set $\mathbf{A}$ of points is called the convex hull of $\mathbf{A}$.

Definition A.2 *For* $\mathbf{A} \subseteq \mathbf{R}^s$*, the set*

$$\operatorname{conv} \mathbf{A} := \left\{ \sum_{i=0}^{n-1} \lambda_i \mathbf{x}_i : \ n \geq 1, \, \mathbf{x}_i \in \mathbf{A}, \, \lambda_i \geq 0, \, \sum_{i=0}^{n-1} \lambda_i = 1 \right\}$$

is called the convex hull *of* $\mathbf{A}$.

Suppose $\mathbf{A}$ contains only a finite number of points. Then there are two additional interpretations of $\operatorname{conv} \mathbf{A}$ besides the geometric one we used to define it. $\operatorname{conv} \mathbf{A}$ is the set of all possible centers of gravity obtained from all conceivable distributions of mass among the points in $\mathbf{A}$, and $\operatorname{conv} \mathbf{A}$ is the set of all expectations a random variable with values in $\mathbf{A}$ can assume.

With the identification of 'center of gravity' and 'expectation' in mind and recalling the fact that even more general probability measures can be viewed as 'distribution of mass', one might suspect this result holds also for more general sets:

> Let $\mathbf{A} \subseteq \mathbf{R}^s$ and $\mathbf{x} \in \mathbf{R}^s$. There is a probability distribution on $\mathbf{A}$ such that $\mathbf{x}$ is the corresponding expectation if and only if $\mathbf{x} \in \operatorname{conv} \mathbf{A}$.

Before we set out to prove it, we develop a more formal representation of this statement in the next section.

A.2 The formal statement

(A.2.1) Let $(\Omega, \mathcal{A})$ be a measurable space and let

$$\begin{array}{rcl} \mathbf{X} : \Omega & \longrightarrow & \mathbf{R}^s \\ \omega & \longmapsto & (X_0(\omega), \ldots, X_{s-1}(\omega)) \end{array}$$

be measurable. For a probability measure μ on $(\Omega, \mathcal{A})$, let

$$E_\mu(\mathbf{X}) := (E_\mu(X_0), \ldots, E_\mu(X_{s-1}))$$

be the expectation of $\mathbf{X}$ with respect to μ. Finally, let

$$\begin{array}{rcl} \mathcal{M} & := & \{\text{probability measures } \mu \text{ on } (\Omega, \mathcal{A}) \text{ such that } E_\mu(\mathbf{X}) \in \mathbf{R}^s\}, \\ \mathcal{E} & := & \{E_\mu(\mathbf{X}) : \ \mu \in \mathcal{M}\}, \\ \mathcal{C} & := & \operatorname{conv} \mathbf{X}(\Omega). \end{array}$$

With these conventions, we can express our claim from (A.1.2) as

Proposition A.1

$$\mathcal{E} = \mathcal{C}.$$

Remark: Since the set $\mathcal{C}$ does not depend on the σ-algebra $\mathcal{A}$ on Ω, we conclude that the set $\mathcal{E}$ of possible finite expectations of $\mathbf{X}$ is independent of $\mathcal{A}$.

(A.2.2) The idea for Proposition A.1 is derived from De Finetti's 'Theory of Probability' [15]: in Section 3.4 of his book, he notes that $\mathcal{E} = \overline{\mathcal{C}}$, i.e. that $\mathcal{E}$ is the closure of $\mathcal{C}$. This is due to the fact that De Finetti uses a concept of probability measure which must be finite- but not necessarily σ-additive. The most widely accepted notion of probability measure due to Kolmogorov [61] which we use in this text *must* be σ-additive. Due to the requirement of σ-additivity, we get that $\mathcal{E}$ equals $\mathcal{C}$ instead of $\overline{\mathcal{C}}$.

To see that the difference between finite-additive measures and σ-additive measures is responsible for getting $\mathcal{E} = \overline{\mathcal{C}}$ and $\mathcal{E} = \mathcal{C}$, respectively, consider the following example. Let $\mathbf{X}$ be a real-valued random variable and let $]0,1[$ be the set of its possible values[2]. For finite-additive measures,

[2] This is $\mathbf{X} : \Omega \mapsto]0,1[$ is measurable with respect to some σ-algebra on Ω and the Borel σ-algebra on $]0,1[$, and $\mathbf{X}(\Omega) =]0,1[$.

we have $\mathcal{E} = [0,1]$ due to [15, Section 3.4] and, for σ-additive measures, we claim that $\mathcal{E} =]0,1[$.

Every number $\mathbf{y} \in]0,1[$ is the expectation of a point measure concentrated on $\mathbf{y}$. Since each such point measure is σ-additive, the relation $\mathcal{E} \subseteq]0,1[$ holds for σ-additive probability measures.

Now suppose there is a σ-additive probability measure μ with $E_\mu(\mathbf{X}) = 0$ (This is the only case which needs consideration. Since $\mathbf{X}$ is positive, the expectation of $\mathbf{X}$ cannot be negative, and expectations ≥ 1 are handled analogously). For any positive integer n, we have

$$\begin{aligned} 0 &= E_\mu(\mathbf{X}|\mathbf{X} < 1/n)P_\mu(\mathbf{X} < 1/n) + \\ &\quad E_\mu(\mathbf{X}|\mathbf{X} \geq 1/n)P_\mu(\mathbf{X} \geq 1/n), \end{aligned}$$

which implies[3] that

$$P_\mu(\mathbf{X} \geq 1/n) = \mu(\{\mathbf{X} \geq 1/n\}) = 0.$$

We get a sequence of measurable sets $\{\mathbf{X} \geq 1/n\}$ with

$$\bigcup_{n \geq 1} \{\mathbf{X} \geq 1/n\} = \{\mathbf{X} > 0\}.$$

Each of the events $\{\mathbf{X} \geq 1/n\}$ has measure 0, but the limit $\{\mathbf{X} > 0\}$ has measure 1. Thus μ is not continuous from below. On the other hand, we have assumed that μ is σ-additive, which implies continuity from below. Since this is a contradiction, we conclude that $E_\mu(\mathbf{X}) = 0$ is impossible.

The point of all this is the following: a finite-additive measure is σ-additive if and only if it is continuous from below. So the difference between the finite-additive and σ-additive notions of 'measure' concerning $\mathcal{E}$ is that – in this example – we get $\mathcal{E} = [0,1]$ in the finite-additive and $\mathcal{E} =]0,1[$ in the σ-additive case.

(A.2.3) We expected a short search through the literature would certainly reveal a proof of Proposition A.1 or some similar statement. We were astonished to find none of these. The best we can come up with are two rather unsatisfying references. The first is Leichtweiß [74, Satz 3.6], where a special case of Proposition A.1 is shown. Second, Stoelinga apparently[4]

[3] $\mathbf{X}$ is nonnegative and so are both of the conditional expectations $E_\mu(\mathbf{X}|\ldots)$. Since the probabilities $P_\mu(\ldots)$ are nonnegative too, both expressions in the sum above must be zero. On the other hand, we have $1/n \leq E_\mu(\mathbf{X}|\mathbf{X} \geq 1/n)$, which implies $P_\mu(\mathbf{X} \geq 1/n) = 0$.

[4] According to [6, Section 2.6]. We would like to thank Johann Linhart for providing us with these references.

proved a special case from a geometric point of view in his Ph.D. thesis [108]. However, we were unable to locate his work in time; moreover, locating his work would not have been sufficient either since it is in Dutch.

A.3 Some utilities

(A.3.1) Before we prove Proposition A.1, we recall some notions and lemmata from the fields convex geometry and measure theory.

For two points $\mathbf{a} = (a_0, \ldots, a_{s-1})$ and $\mathbf{b} = (b_0, \ldots, b_{s-1})$ in $\mathbf{R}^s$, we denote their inner product $\sum_{i=0}^{s-1} a_i b_i$ by $<\mathbf{a}, \mathbf{b}>$.
By $H = H_{\mathbf{n},\alpha}$, we denote the hyperplane defined by the vector $\mathbf{n} \in \mathbf{R}^s \setminus \{\mathbf{0}\}$ and the constant $\alpha \in \mathbf{R}$:

$$H_{\mathbf{n},\alpha} := \{\mathbf{y} \in \mathbf{R}^s : \ <\mathbf{n}, \mathbf{y}> = \alpha\}.$$

By H^+ and H^-, we denote the closed half-spaces defined by the hyperplane H; this is

$$H^+_{\mathbf{n},\alpha} := \{\mathbf{y} \in \mathbf{R}^s : \ <\mathbf{n}, \mathbf{y}> \geq \alpha\}$$

and $H^-_{\mathbf{n},\alpha}$ is defined analogously.
Let $\mathbf{x} \in \mathbf{R}^s$ and $\mathbf{A} \subseteq \mathbf{R}^s$. A hyperplane H is said to separate $\mathbf{x}$ and $\mathbf{A}$ if

$$\mathbf{A} \subseteq H^-$$

and

$$x \in H^+,$$

or vice versa. We say that $H_{\mathbf{n},\alpha}$ strongly separates $\mathbf{x}$ and $\mathbf{A}$ if there is an $\epsilon > 0$ such that both $H_{\mathbf{n},\alpha-\epsilon}$ and $H_{\mathbf{n},\alpha+\epsilon}$ separate $\mathbf{x}$ and $\mathbf{A}$. Finally, we say that $\mathbf{x}$ and $\mathbf{A}$ are (strongly) separable if there is a hyperplane which (strongly) separates $\mathbf{x}$ and $\mathbf{A}$. Note that if $\mathbf{x}$ and $\mathbf{A}$ are separable, then

$$\exists \mathbf{n} \in \mathbf{R}^s \setminus \{\mathbf{0}\} \ \forall \mathbf{a} \in \mathbf{A} : \ <\mathbf{n}, \mathbf{a}> \ \leq \ <\mathbf{n}, \mathbf{x}> \ .$$

If $\mathbf{x}$ and $\mathbf{A}$ are strongly separable, then the same statement as above with just the '$\leq$' replaced by '$<$' holds.

Definition A.3 $\mathbf{C} \subseteq \mathbf{R}^s$ *is* convex *if*

$$\forall \mathbf{x}, \mathbf{y} \in \mathbf{C} \ \forall 0 \leq \alpha \leq 1 : \ \alpha \mathbf{x} + (1 - \alpha)\mathbf{y} \in \mathbf{C}.$$

Lemma A.1 *Let* $\mathbf{A} \subseteq \mathbf{R}^s$ *be convex and let* $\mathbf{x} \in \mathbf{R}^s$*.*

1. *If* $\mathbf{x} \notin \overline{\mathbf{A}}$*, then* $\mathbf{x}$ *and* $\mathbf{A}$ *are strongly separable.*
2. *If* $\mathbf{x} \in \partial\mathbf{A}$*, then* $\mathbf{x}$ *and* $\mathbf{A}$ *are separable*[5]*.*

This lemma is a standard result from convex geometry stated in a form suitable for our purpose. Part 1 is a consequence of Schneider [102, Theorem 1.1.9 and 1.3.4] and Part 2 is a consequence of [102, Theorem 1.3.2].

Lemma A.2 *Let* $\mathbf{A} \subseteq \mathbf{R}^s$*. Then*

$$\operatorname{conv} \mathbf{A} = \bigcap_{\substack{\mathbf{C} \subseteq \mathbf{R}^s \text{ is convex and} \\ \mathbf{A} \subseteq \mathbf{C}}} \mathbf{C}$$

and

$$\operatorname{conv} \mathbf{A} \textit{ is convex.}$$

For a proof of this, see [102, Theorem 1.1.2].

Lemma A.3 *Let* $T : \mathbf{R}^s \to \mathbf{R}^t$ *be linear and* $\mathbf{A} \subseteq \mathbf{R}^s$*. Then*

$$\operatorname{conv} T(\mathbf{A}) = T(\operatorname{conv} \mathbf{A}).$$

Proof: First, we show that $\operatorname{conv} T(\mathbf{A}) \subseteq T(\operatorname{conv} \mathbf{A})$. We have $\mathbf{A} \subseteq \operatorname{conv} \mathbf{A}$ and therefore

$$T(\mathbf{A}) \subseteq T(\operatorname{conv} \mathbf{A}).$$

Observe that $T(\operatorname{conv} \mathbf{A})$ is convex since T is linear[6]. Thus $T(\operatorname{conv} \mathbf{A})$ is a convex superset of $T(\mathbf{A})$, and Lemma A.2 implies

$$\operatorname{conv} T(\mathbf{A}) \subseteq T(\operatorname{conv} \mathbf{A}).$$

[5] By $\partial\mathbf{A}$, we denote the boundary of A, i.e. $\partial\mathbf{A} := \overline{\mathbf{A}} \setminus \underline{\mathbf{A}}$.

[6] Let $\mathbf{y}_1, \mathbf{y}_2 \in T(\operatorname{conv} \mathbf{A})$ and $0 \leq \alpha \leq 1$. Then there are $\mathbf{x}_1, \mathbf{x}_2 \in \operatorname{conv} \mathbf{A}$ such that $\mathbf{y}_1 = T(\mathbf{x}_1)$ and $\mathbf{y}_2 = T(\mathbf{x}_2)$. Since T is linear and $\operatorname{conv} \mathbf{A}$ is convex, we get

$$\begin{aligned} \alpha\mathbf{y}_1 + (1-\alpha)\mathbf{y}_2 &= T(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \\ &\in T(\operatorname{conv} \mathbf{A}). \end{aligned}$$

Next, we show $T(\operatorname{conv} \mathbf{A}) \subseteq \operatorname{conv} T(\mathbf{A})$. Let $\mathbf{y} \in T(\operatorname{conv} \mathbf{A})$. Due to the definition of $\operatorname{conv} \mathbf{A}$, there is some finite number of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbf{A}$ and there are nonnegative numbers $\lambda_1, \ldots, \lambda_n$ whith sum up to 1 such that

$$\begin{aligned} \mathbf{y} &= T(\sum_{i=0}^{n-1} \lambda_i \mathbf{x}_i) \\ &= \sum_{i=0}^{n-1} \lambda_i T(\mathbf{x}_i). \end{aligned}$$

The points $T(\mathbf{x}_i)$ all lie in $T(\mathbf{A})$. Hence $\mathbf{y}$ is a convex combination of points in $T(\mathbf{A})$, which implies $\mathbf{y} \in \operatorname{conv} T(\mathbf{A})$. □

Lemma A.4 *The convex hull of a $\mathbf{A} \subseteq \mathbf{R}^s$ is independent of the choice of the basis of $\mathbf{R}^s$.*

Proof: Suppose we choose two bases of $\mathbf{R}^s$. Let T be the corresponding coordinate transformation (which is of course linear and bijective). Let $\mathbf{A}$ be the representation of a subset of $\mathbf{R}^s$ with respect to the first basis and $\mathbf{A}'$ be the representation of the same set with respect to the second basis; this is

$$T(\mathbf{A}) = \mathbf{A}'.$$

With Lemma A.3, we get

$$T(\operatorname{conv} \mathbf{A}) = \operatorname{conv} \mathbf{A}',$$

and – since T^{-1} is linear, too –

$$T^{-1}(\operatorname{conv} \mathbf{A}') = \operatorname{conv} \mathbf{A}.$$

□

Lemma A.5 *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let*

$$h : \Omega \longrightarrow [0, \infty]$$

be measurable. Then

$$\int h d\mu \geq 0$$

and

$$\int h d\mu = 0 \Longleftrightarrow h = 0 \text{ almost everywhere.}$$

This lemma is a standard result of measure theory and is easily derived from, say, [11, Corollary 2.3.19 and Proposition 2.3.4].

Lemma A.6 *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let f, g be real-valued μ-integrable functions on Ω. Then*

$$\begin{aligned} f < g \quad &\Longrightarrow \quad \int f d\mu < \int g d\mu, \text{ and} \\ f \leq g, \int f d\mu = \int g d\mu \quad &\Longrightarrow \quad f = g \text{ almost everywhere.} \end{aligned}$$

Proof: Let $(\Omega, \mathcal{A}, \mu)$, f, and g be as demanded in the lemma and let $h := g - f$. Observe that h is real-valued, nonnegative, and μ-integrable.

For the proof of the first implication of Lemma A.6, let $f < g$, which implies $h > 0$. We can apply the second part of Lemma A.5 to derive $\int h d\mu \neq 0$. Due to the first part of Lemma A.5, the integral of h is nonnegative. We have

$$\int h d\mu > 0,$$

which is equivalent to $\int f d\mu < \int g d\mu$.

For the proof of the second implication of Lemma A.6, let $f \leq g$ and $\int f d\mu = \int g d\mu$. This means $h \geq 0$ and $\int h d\mu = 0$. Using the second part of Lemma A.5, we get $h = 0$ almost everywhere or, equivalently, $f = g$ almost everywhere. □

A.4 Proof of the formal statement

Proof of Proposition A.1: We conduct the proof in three steps:

$$\begin{aligned} \mathcal{C} &\subseteq \mathcal{E}, && (1) \\ \mathcal{E} &\subseteq \overline{\mathcal{C}}, && (2) \\ \mathcal{E} \cap \partial\mathcal{C} &\subseteq \mathcal{C}. && (3) \end{aligned}$$

With these three inclusions, it is easy[7] to derive $\mathcal{E} = \mathcal{C}$. Observe that (2) can be written in the form

$$\mathcal{E} \subseteq \mathcal{C} \cup \partial\mathcal{C}.$$

[7] We would like to thank Stefan Wegenkittl for pointing out this simplification of our original argument.

Intersecting both sides of this with $\mathcal{E}$ yields

$$\mathcal{E} \subseteq (\mathcal{C} \cap \mathcal{E}) \cup (\partial\mathcal{C} \cap \mathcal{E}).$$

With this, $\mathcal{E}$ is contained in the union of two sets, each of which is in turn contained in $\mathcal{C}$: $\mathcal{C} \cap \mathcal{E}$ is contained in $\mathcal{C}$ by definition and $\partial\mathcal{C} \cap \mathcal{E}$ is because of (3). Therefore $\mathcal{E}$ itself is contained in $\mathcal{C}$. On the other hand, (1) states that $\mathcal{C}$ is contained in $\mathcal{E}$. Hence the two sets are equal.

Proof of (1): We show that $\mathcal{E}$ is a convex superset of $\mathbf{X}(\Omega)$. This and Lemma A.2 imply (1).

Let $\mathbf{x}_0 = \mathbf{X}(\omega_0) \in \mathbf{X}(\Omega)$. We define a measure μ as the point measure concentrated at $\mathbf{x}_0$:

$$\begin{aligned} \mu : \mathcal{A} &\longrightarrow [0,1] \\ A &\longmapsto 1_A(\omega_0). \end{aligned}$$

It is easy to see that μ is a probability measure on $(\Omega, \mathcal{A})$ (it is a nonnegative, σ-additive functional on $\mathcal{A}$ with $\mu(\Omega) = 1$). For this measure, we have

$$\mathbf{X} = \mathbf{x}_0 \qquad \mu\text{-almost everywhere},$$

and therefore

$$E_\mu(\mathbf{X}) = \mathbf{x}_0.$$

Thus $\mu \in \mathcal{M}$ and $\mathbf{x}_0 \in \mathcal{E}$. This proves that $\mathcal{E}$ is a superset of $\mathbf{X}(\Omega)$.

Let $\mu_1, \mu_2 \in \mathcal{M}$ with $E_{\mu_1}(\mathbf{X}) = \mathbf{x}_1$ and $E_{\mu_2}(\mathbf{X}) = \mathbf{x}_2$, and let $0 \leq \alpha \leq 1$. To prove that $\mathcal{E}$ is convex, we have to show that $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$ is in $\mathcal{E}$. For this, we define a measure μ as

$$\begin{aligned} \mu : \mathcal{A} &\longrightarrow [0,1] \\ A &\longmapsto \alpha\mu_1(A) + (1-\alpha)\mu_2(A). \end{aligned}$$

It is easy to see[8] that μ is a well-defined probability measure on $(\Omega, \mathcal{A})$. For a step-function f, we have

$$\alpha \int f d\mu_1 + (1-\alpha) \int f d\mu_2 = \int f d\mu.$$

[8] For $A \in \mathcal{A}$, the numbers $\mu_1(A)$ and $\mu_2(A)$ are both in $[0,1]$. Since $\mu(A)$ is a convex combination of these numbers, it is in $[0,1]$ as well. The σ-additivity of μ_1 and μ_2 is simply inherited by μ. Finally, since $\mu(\Omega) = \alpha 1 + (1-\alpha)1 = 1$, μ is indeed a probability measure on $(\Omega, \mathcal{A})$.

Let f be an function which is both μ_1- and μ_2-integrable. Then f can be represented as the point-wise limit of a sequence $(f_n)_{n\geq 1}$ of step-functions. Moreover, the, say, μ_1-integral of f is the limit of the μ_1-integrals of the step functions f_n. This yields

$$\begin{aligned}
\alpha \int f d\mu_1 + (1-\alpha) \int f d\mu_2 &= \alpha \lim_{n\to\infty} \int f_n d\mu_1 + (1-\alpha) \lim_{n\to\infty} \int f_n d\mu_2 \\
&= \lim_{n\to\infty} \left(\alpha \int f_n d\mu_1 + (1-\alpha) \int f_n d\mu_2 \right) \\
&= \lim_{n\to\infty} \int f_n d\mu \\
&= \int f d\mu.
\end{aligned}$$

$\mathbf{X}$ is both μ_1- and μ_2-integrable, and so

$$\begin{aligned}
E_\mu(\mathbf{X}) &= \alpha E_{\mu_1}(\mathbf{X}) + (1-\alpha) E_{\mu_2}(\mathbf{X}) \\
&= \alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2.
\end{aligned}$$

Since $E_\mu(\mathbf{X}) \in \mathbf{R}^s$, we have $\mu \in \mathcal{M}$ and finally $E_\mu(\mathbf{X}) \in \mathcal{E}$.

Proof of (2): We assume $\mathcal{E}\backslash\overline{\mathcal{C}} \neq \emptyset$ and show this gives a contradiction. Let $\mathbf{x} = E_\mu(\mathbf{X}) \in \mathcal{E} \setminus \overline{\mathcal{C}}$. Since $\mathbf{x}$ is not contained in the closure of the convex set $\mathcal{C}$, we can apply Lemma A.1, Part 1 to conclude that $\mathcal{C}$ and $\mathbf{x}$ are strongly separable. There is an $\mathbf{n} \in \mathbf{R}^s \setminus \{\mathbf{0}\}$ with

$$\forall \mathbf{c} \in \mathcal{C} : \ <\!\mathbf{n}, \mathbf{c}\!> \ < \ <\!\mathbf{n}, \mathbf{x}\!> \ .$$

The possible values $\mathbf{X}(\Omega)$ of $\mathbf{X}$ are a subset of $\mathcal{C}$, so

$$<\!\mathbf{n}, \mathbf{X}\!> \ < \ <\!\mathbf{n}, \mathbf{x}\!> \ .$$

The mapping $<\!\mathbf{n}, \mathbf{X}\!> = \sum_{i=0}^{s-1} n_i X_i$ is measurable. Because $\mu \in \mathcal{M}$, i.e. all the X_i are μ-integrable, so is $<\!\mathbf{n}, \mathbf{X}\!>$. We have two integrable functions, $<\!\mathbf{n}, \mathbf{X}\!>$ and the constant function $<\!\mathbf{n}, \mathbf{x}\!>$, where one is always smaller than the other; Lemma A.6 implies

$$E_\mu(<\!\mathbf{n}, \mathbf{X}\!>) < \ <\!\mathbf{n}, \mathbf{x}\!> \ .$$

Conversely, using the fact that E_μ is a linear functional, we get

$$\begin{aligned}
E_\mu(<\!\mathbf{n}, \mathbf{X}\!>) &= \ <\!\mathbf{n}, E_\mu(\mathbf{X})\!> \\
&= \ <\!\mathbf{n}, \mathbf{x}\!> \ .
\end{aligned}$$

$E_\mu(<\mathbf{n},\mathbf{X}>)$ is simultaneously both smaller than and equal to $<\mathbf{n},\mathbf{x}>$, which is a contradiction. Hence $\mathcal{E} \setminus \overline{\mathcal{C}} = \emptyset$ and $\mathcal{E} \subseteq \overline{\mathcal{C}}$.

Proof of (3): For this final part, we use induction on the dimension s. For both parts of the proof (the cases $s = 1$ and $s-1 \to s$), we use essentially the same argument as presented in (A.2.2) to point out that the difference between finite-additive and σ-additive measures is, when $\mathcal{E}$ is concerned, just the difference between $\overline{\mathcal{C}}$ and $\mathcal{C}$. Instead of directly using the σ-additivity or the continuity from below as in (A.2.2), we use Lemma A.6 here. The proof can be conducted using continuity from below, but it is easier with the lemma.

Let $s = 1$. Recall the definition of an interval: an interval[9] $I \subseteq \mathbf{R}$ is a set which with two points contains all the points in between as well. Since this is the one-dimensional case of the general Definition A.3 of a convex set, a subset of $\mathbf{R}$ is convex if and only if it is an interval.
If $\mathcal{E} \cap \partial\mathcal{C}$ is empty, then (3) trivially holds. Now suppose it is not empty and choose $\mathbf{x} = E_\mu(\mathbf{X}) \in \mathcal{E} \cap \partial\mathcal{C}$. We have to show that $\mathbf{x} \in \mathcal{C}$.
Because $\mathbf{x} \in \partial\mathcal{C}$ and $\mathcal{C}$ is convex, $\mathbf{x}$ is an end-point of the interval $\mathcal{C}$. This interval contains all the values $\mathbf{X}(\Omega)$, so we can conclude that either $\mathbf{X} \leq \mathbf{x}$ or $\mathbf{X} \geq \mathbf{x}$. We apply the second part of Lemma A.6 to conclude

$$\mathbf{X} = \mathbf{x} \qquad \mu\text{-almost everywhere.}$$

Thus there is at least one $\omega \in \Omega$ with $\mathbf{X}(\omega) = \mathbf{x}$, which means

$$\mathbf{x} \in \mathbf{X}(\Omega) \subseteq \mathcal{C}.$$

For the step $s - 1 \to s$, let $s > 1$ and assume (3) holds for all dimensions smaller than s. As before, only the case $\mathcal{E} \cap \partial\mathcal{C} \neq \emptyset$ is of interest.
Let $\mathbf{x} = E_\mu(\mathbf{X}) \in \mathcal{E} \cap \partial\mathcal{C}$. Due to Lemma A.1, Part 2, we can separate $\mathbf{x}$ and $\mathcal{C}$ by a hyperplane $H_{\mathbf{n},\alpha}$. Since $\mathbf{X}(\Omega)$ is a subset of $\mathcal{C}$, we have, say,

$$<\mathbf{n},\mathbf{X}> \leq <\mathbf{n},\mathbf{x}> ,$$

and – just as in the case $s = 1$ – we conclude

$$<\mathbf{n},\mathbf{X}> = <\mathbf{n},\mathbf{x}> \qquad \mu\text{-almost everywhere.}$$

[9] In the sense that intervals are subsets of $\mathbf{R}$ of the form $[a, b]$, $]a, b[$, $]a, b]$, or $[a, b[$ for $a, b \in \mathbf{R} \cup \{-\infty, \infty\}$ and $a \leq b$.

This means that the s-dimensional random quantity $\mathbf{X}$ is almost certainly contained in the hyperplane $H_{\mathbf{n},\alpha}$. We will use this to apply the induction hypothesis to an $(s-1)$-dimensional space, the hyperplane. Before doing so, we have to get rid of those possible points of $\mathbf{X}$ which do not fall in the hyperplane.

There is at least one $\omega_0 \in \Omega$ with $<\mathbf{n}, \mathbf{X}(\omega_0)> = <\mathbf{n}, \mathbf{x}>$. We change $\mathbf{X}$ on a set of measure zero, defining

$$\mathbf{X}'(\omega) := \begin{cases} \mathbf{X}(\omega) & \text{if } <\mathbf{n}, \mathbf{X}(\omega)> = <\mathbf{n}, \mathbf{x}> , \\ \mathbf{X}(\omega_0) & \text{otherwise.} \end{cases}$$

The random quantity $\mathbf{X}'$ is again μ-integrable,

$$\begin{aligned} E_\mu(\mathbf{X}') &= \mathbf{x}, \qquad \text{and} \\ <\mathbf{n}, \mathbf{X}'> &= <\mathbf{n}, \mathbf{x}> . \end{aligned}$$

Now we choose vectors $\mathbf{b}_0, \ldots, \mathbf{b}_{s-2}$ which span the hyperplane $H_{\mathbf{n},\alpha}$. The set $\mathbf{B} := \{\mathbf{b}_0, \ldots, \mathbf{b}_{s-2}, \mathbf{n}\}$ is a basis of $\mathbf{R}^s$. Let T be the coordinate transformation from the standard ordered basis $\{\mathbf{e}_0, \ldots, \mathbf{e}_{s-1}\}$ of $\mathbf{R}^s$ to $\mathbf{B}$, which can be can be uniquely identified with an $s \times s$-matrix[10]. For any point $\mathbf{z} \in \mathbf{R}^s$, we denote its coordinates with respect to $\mathbf{B}$ by $T(\mathbf{z}) = (T_0(\mathbf{z}), \ldots, T_{s-1}(\mathbf{z}))$. Note that the representation of $T(H_{\mathbf{n},\alpha})$ of our hyperplane with respect to the new basis is very simple. $T(H_{\mathbf{n},\alpha})$ is just the set of points whose last coordinate is equal to $T_{s-1}(\mathbf{x})$. We get

$$T_{s-1}(\mathbf{X}') = T_{s-1}(\mathbf{x}).$$

So the random quantity $T(\mathbf{X}')$ is completely contained in the hyperplane $T(H_{\mathbf{n},\alpha})$ and its last coordinate function is constant.

Applying the induction hypothesis to the $(s-1)$-dimensional random quantity $(T_0(\mathbf{X}'), \ldots, T_{s-2}(\mathbf{X}'))$ yields[11]

$$(T_0(\mathbf{x}), \ldots, T_{s-2}(\mathbf{x})) \in \operatorname{conv}\left\{(T_0(\mathbf{X}'(\omega)), \ldots, T_{s-2}(\mathbf{X}'(\omega))) :\ \omega \in \Omega\right\}.$$

Projecting $\mathbf{R}^{s-1}$ (where we applied the induction hypothesis) onto the hyperplane $T(H_{\mathbf{n},\alpha})$[12] and using Lemma A.3, we get

$$T(\mathbf{x}) \in \operatorname{conv} T(\mathbf{X}'(\Omega)).$$

[10] Formally, we have $T = S^{-1}$, where the i-th column-vector of S is $\mathbf{b}_i$ $(0 \leq i \leq s-2)$ and its $(s-1)$-th column-vector is $\mathbf{n}$.

[11] $T(\mathbf{X}') = (T_0(\mathbf{X}'), \ldots, T_{s-1}(\mathbf{X}'))$ is measurable because $\mathbf{X}'$ is measurable and T is linear (each $T_i(\mathbf{X}')$ is a linear combination of the coordinate functions X'_j $(0 \leq j < s)$ of $\mathbf{X}'$). Thus $(T_0(\mathbf{X}'), \ldots, T_{s-2}(\mathbf{X}'))$ is measurable and therefore a random quantity.

[12] The point $(y_0, \ldots, y_{s-2}) \in \mathbf{R}^{s-1}$ is projected to $(y_0, \ldots, y_{s-2}, T_{s-1}(\mathbf{x})) \in T(H_{\mathbf{n},\alpha})$.

By Lemma A.4, this is equivalent to

$$\mathbf{x} \in \mathrm{conv}\, \mathbf{X}'(\Omega).$$

By definition of $\mathbf{X}'$, $\mathbf{X}'(\Omega)$ is contained in $\mathbf{X}(\Omega)$ and so are the corresponding convex hulls. Hence

$$\mathbf{x} \;\in\; \mathrm{conv}\, \mathbf{X}(\Omega) \;=\; \mathcal{C}.$$

□

Appendix B

Lattices

B.1 Basic lattice properties

The definition of a lattice was already given in Definition 5.2. In this section, we present some elementary properties of lattices which we make use of later.

(B.1.1) First, we give an alternative characterization of a lattice in Proposition B.1 and B.2, which are stated according to Gruber and Lekkerkerker [46, Section 1.3, Theorem 1 and 2].

Proposition B.1 *An s-dimensional lattice is a discrete subgroup of* $\mathbf{R}^s$*.*

Proof: Let Λ be a lattice with basis $\{\mathbf{b}_0, \ldots, \mathbf{b}_{s-1}\}$. Due to the definition of Λ, we have $\mathbf{0} \in \Lambda$ and, for $\mathbf{x}, \mathbf{y} \in \Lambda$, $\mathbf{x} \pm \mathbf{y}$ is as well in Λ. Hence Λ is an additive subgroup of $\mathbf{R}^s$.

To show that Λ is discrete, we consider the 'cube'

$$\mathbf{W} := \left\{ \sum_{i=0}^{s-1} t_i \mathbf{b}_i : \ |t_i| < 1 \right\}.$$

Then $\Lambda \cap \mathbf{W} = \{\mathbf{0}\}$[1]. Similarly, for any $\mathbf{x} \in \Lambda$, we have $\Lambda \cap (\mathbf{x} + \mathbf{W}) = \{\mathbf{x}\}$.

[1] If this set contained a lattice point $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{x}$ would have two different representations with respect to the lattice basis, one with all integer coordinates and one with at least one noninteger coordinate. Since the vectors of a lattice basis span $\mathbf{R}^s$, this cannot be.

The 'cube' $\mathbf{W}$ contains an open sphere[2] $K_\epsilon(\mathbf{0}) := \{\mathbf{z} : \ ||\mathbf{z}|| < \epsilon\}$ for some radius $\epsilon > 0$. With this, two lattice points have a distance of at least ϵ. Hence Λ is a discrete set. □

Lemma B.1 *Let* $\mathbf{A}$ *be a nonempty subset of a discrete subgroup* $\mathbf{L}$ *of* $\mathbf{R}^s$*. Then* $\mathbf{A}$ *contains a shortest vector:*

$$\exists \mathbf{a} \in \mathbf{A} : \ ||\mathbf{a}|| = \inf \{||\mathbf{b}|| : \ \mathbf{b} \in \mathbf{A}\} .$$

Proof: Let $\mathbf{A}$ be a nonempty subset of the discrete subgroup $\mathbf{L} \subset \mathbf{R}^s$. Since $\mathbf{A}$ is nonempty and since the norm is nonnegative and finite, the quantity

$$\alpha := \inf \{||\mathbf{b}|| : \ \mathbf{b} \in \mathbf{A}\}$$

is well-defined and finite.

Let us assume that there is no $\mathbf{a}$ in $\mathbf{A}$ with $||\mathbf{a}|| = \alpha$. Then there exists a sequence $(\mathbf{a}_n)_{n\geq 0}$ of points in $\mathbf{A}$ with decreasing norm[3]. Because $\mathbf{L}$ is discrete, there is a constant $\epsilon > 0$ such that any two points in $\mathbf{L}$ have a distance of at least ϵ. In particular, any open sphere

$$K_\epsilon(\mathbf{a}_n)$$

of radius ϵ centered at $\mathbf{a}_n$ intersects $\mathbf{L}$ only at the point at its center: $K_\epsilon(\mathbf{a}_n) \cap \mathbf{L} = \{\mathbf{a}_n\}$. In this way, we obtain a sequence of infinitely many disjoint spheres $(K_\epsilon(\mathbf{a}_n))_{n\geq 0}$ of the same, positive volume. For any n, we have $||\mathbf{a}_n|| < ||\mathbf{a}_0||$, so all these disjoint spheres are contained in one big sphere

$$K_{||\mathbf{a}_0||+\epsilon}(\mathbf{0})$$

centered at the origin. Because $||\mathbf{a}_0|| + \epsilon$ is finite, so is the volume of the big sphere. On the other hand, it contains infinitely many disjoint spheres, each of which has the same, positive volume. This implies the volume of the big sphere to be infinite and yields a contradiction. □

Proposition B.2 *Let* Λ *be a discrete subgroup of* $\mathbf{R}^s$ *which is not contained in an* $(s-1)$*-dimensional subspace of* $\mathbf{R}^s$*. Then* Λ *is a lattice.*

[2] Otherwise, $\mathbf{W}$ would not be an open set which it certainly is due to the linear independence of the s spanning vectors.

[3] Choose an arbitrary $\mathbf{a}_0$ from the nonempty set $\mathbf{A}$. Since we have assumed $\alpha < ||\mathbf{a}_0||$, there is an $\mathbf{a}_1 \in \mathbf{A}$ with $\alpha < ||\mathbf{a}_1|| < ||\mathbf{a}_0||$, and so on ...

Proof: Let Λ fulfill the conditions of Proposition B.2. We inductively choose linearly independent points $\mathbf{a}_0, \ldots, \mathbf{a}_{s-1}$, and then show that the lattice spanned by these points is equal to Λ.

Λ is a group, so we have $\mathbf{0} \in \Lambda$. Because Λ is not contained in an $(s-1)$-dimensional subspace of $\mathbf{R}^s$, $\Lambda \setminus \{\mathbf{0}\}$ is not empty. We choose $\mathbf{a}_0 \in \Lambda \setminus \{\mathbf{0}\}$ such that $||\mathbf{a}_0||$ is minimal. Since Λ is discrete, Lemma B.1 assures that $\mathbf{a}_0$ is well-defined.

If $\mathbf{a}_0, \ldots, \mathbf{a}_{k-1}$ $(k < s)$ have been chosen, we choose $\mathbf{a}_k$ as follows. Λ is not contained in the k-dimensional subspace $[\mathbf{a}_0, \ldots, \mathbf{a}_{k-1}]$ spanned by the $\mathbf{a}_0, \ldots, \mathbf{a}_{k-1}$, so we can choose some

$$\mathbf{b} \in \Lambda \setminus [\mathbf{a}_0, \ldots, \mathbf{a}_{k-1}].$$

Let $\mathbf{P}_k$ be the parallelepiped spanned by the $\mathbf{a}_i$ chosen so far and $\mathbf{b}$:

$$\mathbf{P}_k := \left\{ \sum_{i=0}^{k-1} \lambda_i \mathbf{a}_i + \lambda \mathbf{b} : \; 0 \leq \lambda_i \leq 1, \, 0 \leq \lambda \leq 1 \right\}.$$

Since $\mathbf{b} \in \mathbf{P}_k$, we have

$$\mathbf{P}_k \setminus [\mathbf{a}_0, \ldots, \mathbf{a}_{k-1}] \neq \emptyset.$$

Because $\mathbf{P}_k$ is bounded and Λ is discrete, the set

$$\mathbf{P}_k \cap \Lambda$$

is finite[4]. Let $\mathbf{a}_k$ be that point in $\mathbf{P}_k \cap \Lambda$ which has minimal *positive* distance from $[\mathbf{a}_0, \ldots, \mathbf{a}_{k-1}]$.

Suppose the $\mathbf{a}_0, \ldots, \mathbf{a}_{s-1}$ are chosen according to the procedure above. Since

$$\mathbf{a}_0 \neq \mathbf{0}, \text{ and } \mathbf{a}_k \notin [\mathbf{a}_0, \ldots, \mathbf{a}_{k-1}] \qquad (0 < k < s),$$

the $\mathbf{a}_i$ are linearly independent.
Λ is a group containing the $\mathbf{a}_i$, so it contains any point of the form $\sum_{i=0}^{s-1} k_i \mathbf{a}_i$ $(k_i \in \mathbf{Z})$ as well. Thus Λ contains the lattice spanned by $\{\mathbf{a}_0, \ldots, \mathbf{a}_{s-1}\}$. It

[4] Assume $\mathbf{P}_k \cap \Lambda$ is infinite. For $\delta = \sup\{||\mathbf{z}|| : \; \mathbf{z} \in \mathbf{P}_k\}$, which is of course finite, the closed sphere $\overline{K_\delta(\mathbf{0})}$ is a superset of $\mathbf{P}_k$. $\mathbf{P}_k$ contains ininitely many points of the discrete set Λ, each of which can be separated from the rest by a sphere of some fixed radius $\epsilon > 0$. This yields the contradiction that $\overline{K_\delta(\mathbf{0})}$ contains infinitely many disjoint spheres of the same, positive volume.

remains to show that there are no other points in Λ.
Due to the linear independence of the $\mathbf{a}_i$, any $\mathbf{a} \in \Lambda$ can be written as

$$\mathbf{a} = \sum_{i=0}^{s-1} t_i \mathbf{a}_i \qquad (t_i \in \mathbf{R}).$$

We have to show that the t_i are all integers or, equivalently, that the values $u_i := t_i - \lfloor t_i \rfloor$ are all zero.
Because the group Λ contains the lattice spanned by the $\mathbf{a}_i$, we have

$$\sum_{i=0}^{s-1} u_i \mathbf{a}_i \in \Lambda.$$

Recall that we have chosen $\mathbf{a}_{s-1}$ as that point in $\mathbf{P}_{s-1} \cap \Lambda$ which has minimal positive distance from $[\mathbf{a}_0, \ldots, \mathbf{a}_{s-2}]$. Together with $0 \leq u_{s-1} < 1$, this implies that u_{s-1} is zero: assume $u_{s-1} > 0$; the choice of $\mathbf{a}_{s-1} = \sum_{i=0}^{s-2} \lambda_i \mathbf{a}_i + \lambda \mathbf{b} \in \mathbf{P}_{s-1}$ yields

$$\sum_{i=0}^{s-1} u_i \mathbf{a}_i = \sum_{i=0}^{s-2} (u_i + u_{s-1}\lambda_i)\mathbf{a}_i + u_{s-1}\lambda \mathbf{b}.$$

Since $\mathbf{a}_i \in \Lambda$, we get[5]

$$\sum_{i=0}^{s-2} \{u_i + u_{s-1}\lambda_i\}\mathbf{a}_i + u_{s-1}\lambda \mathbf{b} \ \in \ \mathbf{P}_{s-1} \cap \Lambda.$$

Since $0 < u_{s-1}\lambda < \lambda$, this point would have *smaller* positive distance from $[\mathbf{a}_0, \ldots, \mathbf{a}_{s-2}]$ than $\mathbf{a}_{s-1}$. By the choice of $\mathbf{a}_{s-1}$, this can not be.
Similarly, the choice of $\mathbf{a}_{s-2}$ and $0 \leq u_{s-2} < 1$ imply that u_{s-2} is zero, etc. .... We get that $u_{s-1}, \ldots, u_1$ are all zero. Finally, u_0 is zero since otherwise $||\mathbf{a}_0||$ would not be minimal. □

(B.1.2) From Definition 5.2, it is clear that there is more than one basis spanning any given lattice. Any lattice basis is transformed to an equivalent basis by, say, reversing some of its vectors. To characterize which lattice points can be part of a lattice basis, we need the notion of a primitive set, which we define according to [46, Section 1.3, Definition 2].

[5] Note that for $\mu \in \mathbf{R}$, we denote the fractional part of μ by $\{\mu\}$; i.e. $\mu = \lfloor \mu \rfloor + \{\mu\}$, where $\lfloor \mu \rfloor$ is an integer and $0 \leq \{\mu\} < 1$.

Definition B.1 *A system of $k \leq s$ linearly independent points $\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}$ in an s-dimensional lattice Λ is called* primitive *if*

$$\Lambda \cap [\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}] = \left\{ \sum_{i=0}^{k-1} c_i \mathbf{b}_i : \; c_i \in \mathbf{Z} \right\}.$$

Note that any lattice basis and any subset of a lattice basis is primitive; moreover, any primitive set of s vectors is a lattice basis. Finally, consider the case of a set of just one nonzero vector $\{\mathbf{b}\}$; this set is primitive if the line-segment joining $\mathbf{0}$ and $\mathbf{b}$ contains no other lattice point. For the sake of simplicity, we say that $\mathbf{b}$ is primitive if $\{\mathbf{b}\}$ is primitive.

The reason for defining the notion of a primitive set of lattice points is the following statement, which is a slight modification of [46, Section 1.3, Theorem 5][6].

Proposition B.3 *A set of points of a lattice Λ is primitive if and only if it can be completed to a basis of Λ.*

Proof: If the $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}\}$ can be completed to a basis of Λ, they are of course primitive since every subset of a basis is primitive; this proves the 'only if'.

For the 'if'-part of Proposition B.3, let Λ be an s-dimensional lattice and $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}\}$ be primitive; to avoid trivialities, let $k < s$. It is sufficient to show the existence of a $\mathbf{b}_k \in \Lambda$ such that $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{b}_k\}$ is primitive; by repeated enlargement of the primitive set, it will eventually contain s vectors and therefore be a basis.
Since $k < s$ and Λ is a lattice, we can choose some $\mathbf{c} \in \Lambda \setminus [\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}]$. Let $\mathbf{P}$ be the parallelepiped spanned by the $\mathbf{b}_i$ and $\mathbf{c}$:

$$\mathbf{P} := \left\{ \sum_{i=0}^{k-1} \lambda_i \mathbf{b}_i + \lambda \mathbf{c} : \; 0 \leq \lambda_i \leq 1, \, 0 \leq \lambda \leq 1 \right\}.$$

Since $\mathbf{c} \in \mathbf{P}$, we have

$$\mathbf{P} \setminus [\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}] \neq \emptyset.$$

By a similar argument as used in the proof of Lemma B.1, we see that

$$\mathbf{P} \cap \Lambda$$

[6] To avoid introducing intermediate results, the proof is composed of fragments of the proofs of Theorem 2 and Theorem 5 in [46, Section 1.3]

is finite. As the new point $\mathbf{b}_k$, we choose that point in $\mathbf{P} \cap \Lambda$ which has minimal *positive* distance from $[\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}]$. It remains to show that $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{b}_k\}$ is primitive.
The $\mathbf{b}_0, \ldots, \mathbf{b}_k$ are linearly independent because $\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}$ are linearly independent and $\mathbf{b}_k \notin [\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}]$.
Let

$$\mathbf{p} \in \Lambda \cap [\mathbf{b}_0, \ldots, \mathbf{b}_k].$$

It can be written as

$$\mathbf{p} = \sum_{i=0}^{k} t_i \mathbf{b}_i$$

for some reals t_i which we will show are integers. To be precise, we show that the values $u_i := t_i - \lfloor t_i \rfloor$ $(0 \leq i \leq k)$ are all zero. Since $\mathbf{p}$ is an element of the additive group Λ and the $\lfloor t_i \rfloor$ are integers, the point

$$\mathbf{q} := \sum_{i=0}^{k} u_i \mathbf{b}_i$$

is in Λ. As a linear combination of the $\mathbf{b}_i$, $\mathbf{q}$ is also contained in $[\mathbf{b}_0, \ldots, \mathbf{b}_k]$. Finally, since $0 \leq u_i < 1$, $\mathbf{q}$ is contained in $\mathbf{P}$, too. By the choice of $\mathbf{b}_k$, the scalar u_k must be zero[7]. Hence

$$\mathbf{q} \in \Lambda \cap [\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}].$$

Since $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}\}$ is primitive, the remaining u_i $(0 \leq i < k)$ are integers and therefore are equal to zero. □

We found that a single point $\mathbf{n} \in \Lambda \setminus \{\mathbf{0}\}$ is primitive if and only if the line-segment joining $\mathbf{n}$ and $\mathbf{0}$ contains no other lattice point. For a basis $\{\mathbf{b}_0, \ldots, \mathbf{b}_{s-1}\}$ of Λ, let $\mathbf{n} = \sum_{i=0}^{s-1} l_i \mathbf{b}_i$. Note that the line-segment joining $\mathbf{n}$ and $\mathbf{0}$ contains no other lattice point if and only if the integers $l_0, \ldots, l_{s-1}$ are coprime. In the following, we generalize this observation.

Proposition B.4 *Let $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}\}$ be a primitive subset of the s-dimensional lattice Λ, $k < s$, and let $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{c}_k, \ldots, \mathbf{c}_{s-1}\}$ be the corresponding lattice basis. Let $\mathbf{n}$ be some lattice point with*

$$\mathbf{n} = \sum_{i=0}^{k-1} l_i \mathbf{b}_i + \sum_{i=k}^{s-1} l_i \mathbf{c}_i.$$

[7] Assume $u_k > 0$. By the same argument as used in the corresponding part of the proof of Proposition B.2, we conclude that the point $\mathbf{q} \in \mathbf{P}$ has a smaller but positive distance from $[\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}]$ than $\mathbf{b}_k$. This gives a contradiction with the choice of $\mathbf{b}_k$.

Then

$$\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{n}\} \qquad \textit{is primitive}$$

if and only if[8]

$$\gcd(l_k, \ldots, l_{s-1}) = 1.$$

Proof: First, let $d := \gcd(l_k, \ldots, l_{s-1})$ and assume that $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{n}\}$ is primitive.
If d were equal to 0, then all the $l_k, \ldots, l_{s-1}$ would be zero and $\mathbf{n}$ would be a linear combination of the $\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}$. The resulting linearly dependent vectors $\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{n}$ would not form a primitive set.
If d were greater than 1, the lattice point

$$\sum_{i=k}^{s-1} \frac{l_i}{d}\mathbf{c}_i \;=\; \frac{1}{d}\mathbf{n} - \sum_{i=0}^{k-1} \frac{l_i}{d}\mathbf{b}_i$$

would be in $[\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{n}]$, but its uniquely determined representation as a linear combination of the $\mathbf{b}_i$ $(0 \leq i < k)$ and $\mathbf{n}$ would require at least a noninteger scalar for $\mathbf{n}$. In this case, too, $\{\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{n}\}$ would not be primitive. Therefore d is equal to 1.

Conversely, assume that $d = \gcd(l_k, \ldots, l_{s-1})$ is equal to 1. For an arbitrary lattice point

$$\begin{aligned}\mathbf{v} &= \sum_{i=0}^{k-1} w_i \mathbf{b}_i + w\mathbf{n} \\ &\in \Lambda \cap [\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}, \mathbf{n}],\end{aligned}$$

[8] Since the gcd is taken over scalars of which some or all can be zero, we briefly recall the definition of the greatest common divisor as we use it here. For $a, c \in \mathbf{Z}$, we say 'c divides a' or $c|a$ if there is a nonzero integer b for which $bc = a$:

$$c|a \;\Leftrightarrow\; \exists b \in \mathbf{Z} \setminus \{0\} : \; bc = a.$$

The greatest common divisor of two integers a, b is defined as

$$\gcd(a, b) := \begin{cases} |a| + |b| & \text{if } a \text{ or } b \text{ is zero,} \\ \max\{c \in \mathbf{Z} : \; c|a \text{ and } c|b\} & \text{otherwise.} \end{cases}$$

This means $\gcd(0, a) = \gcd(a, 0) = |a|$ and $\gcd(0, 0) = 0$, which is very important for our usage of gcd to make sense.
The greatest common divisor of more than two integers $a_0, \ldots, a_{n-1}$ is defined recursively by

$$\gcd(a_0, \ldots, a_{n-1}) := \gcd(\gcd(a_0, \ldots, a_{n-2}), a_{n-1}).$$

we have to show that all the w_i and w are integers. Recalling the representation of $\mathbf{n}$, we get

$$\mathbf{v} = \sum_{i=0}^{k-1}(w_i + wl_i)\mathbf{b}_i + \sum_{i=k}^{s-1} wl_i\mathbf{c}_i.$$

In this representation of $\mathbf{v} \in \Lambda$ as a linear combination of vectors of a lattice basis, all the involved scalars are integers. In particular,

$$wl_i \in \mathbf{Z} \qquad (k \leq i < s).$$

Since $d = 1$, at least one of the l_i $(k \leq i < s)$ is nonzero. Hence w is rational. If w is zero, then the above representation of $\mathbf{v}$ yields that the remaining w_i $(0 \leq i < k)$ are integers.
If w is nonzero, it can be written as $w = p/q$ for coprime integers p and $q \geq 1$. From the wl_i $(k \leq i < s)$ being integers, it follows that q divides every nonzero l_i, which implies $q \leq d$. Since $1 \leq q$ and $d = 1$, we have $q = 1$, and thus $w \in \mathbf{Z}$. From this and the above representation of $\mathbf{v}$, it follows that the remaining w_i $(0 \leq i < k)$ are integers, too. □

(B.1.3) A lattice Λ defines a set $\Lambda^\star$ of those points $\mathbf{n}$ for which $<\mathbf{n}, \mathbf{p}>$ is an integer for every $\mathbf{p} \in \Lambda$. This property identifies the so-called polar lattice $\Lambda^\star$, which we define according to [46, Section 1.3, Definition 4].

Definition B.2 *For a lattice Λ in $\mathbf{R}^s$, the set*

$$\Lambda^\star := \{\mathbf{n} \in \mathbf{R}^s : \ \forall \mathbf{p} \in \Lambda : \ <\mathbf{n}, \mathbf{p}> \in \mathbf{Z}\}$$

is called the corresponding polar lattice.

The notation already implies that $\Lambda^\star$ is a lattice. In fact, we have the following result, which we state according to [46, Section 1.3, Theorem 8].

Proposition B.5 *The polar lattice $\Lambda^\star$ is a lattice.*

Proof: Let $\{\mathbf{b}_0, \ldots, \mathbf{b}_{s-1}\}$ be a lattice basis of Λ. Gram-Schmidt orthogonalization yields vectors $\mathbf{b}_i^\star$ $(0 \leq i < s)$ such that

$$<\mathbf{b}_i^\star, \mathbf{b}_j> = \delta_{i,j} \qquad (0 \leq i, j < s).$$

We show that the lattice Λ^+ spanned by $\{\mathbf{b}_0^\star, \ldots, \mathbf{b}_{s-1}^\star\}$ is equal to $\Lambda^\star$.

Let $\mathbf{n} \in \Lambda^\star$. Since the $\mathbf{b}_i^\star$ span $\mathbf{R}^s$, we can represent $\mathbf{n}$ as

$$\mathbf{n} = \sum_{i=0}^{s-1} t_i \mathbf{b}_i^\star \qquad (t_i \in \mathbf{R}).$$

For the lattice points $\mathbf{b}_i \in \Lambda$, the definition of $\Lambda^\star$ yields that

$$<\mathbf{n}, \mathbf{b}_i> \in \mathbf{Z} \qquad (0 \leq i < s).$$

Since all the $<\mathbf{n}, \mathbf{b}_i> = t_i$ are integers, we have $\mathbf{n} \in \Lambda^+$.

Now let $\mathbf{n} \in \Lambda^+$. $\mathbf{n}$ is an integer linear combination of the $\mathbf{b}_i^\star$, so

$$<\mathbf{n}, \mathbf{b}_i> \in \mathbf{Z} \qquad (0 \leq i < s).$$

A point $\mathbf{p} \in \Lambda$ is an integer linear combination of the $\mathbf{b}_i$, so $<\mathbf{n}, \mathbf{p}>$ is an integer. This holds for any lattice point of Λ, so $\mathbf{n} \in \Lambda^\star$. □

B.2 Minkowski-reduced lattice bases

(B.2.1) The basic problem in lattice basis reduction is described by Gruber [45, p.751] as this: given a lattice Λ, "determine (by means of a suitable algorithm) a basis (the 'reduced' basis) having 'nice' geometric or arithmetic properties." In this section, we define and discuss a type of reduced basis whose 'nice geometric properties' include that it coincides with the basis of shortest possible vectors in 2- or 3-dimensional lattices.

Definition B.3 *A basis* $\{\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}\}$ *of the s-dimensional lattice* Λ *is a* Minkowski-reduced lattice basis *(MRLB, for short) if*

$$\begin{aligned} ||\mathbf{m}_0|| &= \min\{||\mathbf{m}|| : \{\mathbf{m}\} \textit{ is primitive}\}, \textit{ and} & (1)\\ ||\mathbf{m}_k|| &= \min\{||\mathbf{m}|| : \{\mathbf{m}_0, \ldots, \mathbf{m}_{k-1}, \mathbf{m}\} \textit{ is primitive}\} \quad (0 < k < s) & (2)\end{aligned}$$

Note that this definition requires that the vectors of a MRLB $\{\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}\}$ of Λ are ordered by their length: $||\mathbf{m}_0|| \leq \ldots \leq ||\mathbf{m}_{s-1}||$.

The notion of a MRLB was introduced before in (5.2.4) as a basis obtained by successively choosing shortest possible basis vectors. With Proposition B.3, any set of basis vectors is necessarily primitive, and vice versa.

Therefore, the informal introduction of a MRLB in (5.2.4) is consistent with Definition B.3.

Definition B.3, which is taken from [46, p.149], seems to be quite different from the more technical definition of a MRLB as used in the literature on random numbers (see [1, Section 3.2.3, Definition 7], for example). However, Proposition B.4 shows that our Definition B.3 and that in, say, [1] are equivalent.

To verify the existence of a MRLB, let Λ be the lattice spanned by $\{\mathbf{b}_0, \ldots, \mathbf{b}_{s-1}\}$. Just as in (5.2.4), we inductively choose lattice points $\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}$ which are then shown to satisfy Definition B.3.
Let $\mathbf{M}_0 := \{\mathbf{m} : \ \{\mathbf{m}\} \text{ is primitive}\}$. Since $\mathbf{M}_0$ is non-empty ($\mathbf{b}_0 \in \mathbf{M}_0$), Lemma B.1 assures that we can choose $\mathbf{m}_0$ as one of the shortest vectors in $\mathbf{M}_0$. Note that $\{\mathbf{m}_0\}$ is primitive.
If $\{\mathbf{m}_0, \ldots, \mathbf{m}_{k-1}\}$ has been chosen and is primitive ($0 < k < s$), let $\mathbf{M}_k := \{\mathbf{m} : \ \{\mathbf{m}_0, \ldots, \mathbf{m}_{k-1}, \mathbf{m}\} \text{ is primitive}\}$. By the choice of the $\mathbf{m}_i$ ($0 \leq i < k$), there is a basis $\{\mathbf{m}_0, \ldots, \mathbf{m}_{k-1}, \mathbf{n}_k, \ldots, \mathbf{n}_{s-1}\}$ of Λ. Since $\mathbf{n}_k \in \mathbf{M}_k \neq \emptyset$, we can choose $\mathbf{m}_k$ as one of the shortest vectors in $\mathbf{M}_k$. Again, note that $\{\mathbf{m}_0, \ldots, \mathbf{m}_k\}$ is primitive.
Proceeding this way, we end up with $\{\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}\}$ being primitive, i.e. being a basis of Λ. By the definition of $\mathbf{M}_k$ and the choice of $\mathbf{m}_k$ ($0 \leq k < s$), it follows that this basis satisfies Definition B.3.

(B.2.2) Unfortunately, a MRLB is not uniquely determined by Definition B.3. Suppose that $\{\mathbf{m}_0, \ldots, \mathbf{m}_{s-1}\}$ is a MRLB of Λ. Then, obviously, $\{\pm\mathbf{m}_0, \ldots, \pm\mathbf{m}_{s-1}\}$ is a MRLB of Λ, too. Moreover, there may be other MRLBs for a given lattice.
For $0 < k < s$, both $\mathbf{m}_k$ and its length depend on $\mathbf{m}_0, \ldots, \mathbf{m}_{k-1}$; $||\mathbf{m}_k||$ is a *local minimum* depending on the k vectors chosen before. Therefore, even the existence of two MRLBs $\{\mathbf{m}_k : \ 0 \leq k < s\}$ and $\{\mathbf{n}_k : \ 0 \leq k < s\}$ of the same lattice Λ with

$$(||\mathbf{m}_0||, \ldots, ||\mathbf{m}_{s-1}||) \neq (||\mathbf{n}_0||, \ldots, ||\mathbf{n}_{s-1}||) \tag{3}$$

is conceivable. According to Gruber and Lekkerkerker [46, p.152], this is impossible for dimensions $s \leq 6$; for $s \geq 7$, however, Ryshkov [98, 99] was able to prove that (3) can in fact occur by giving a 7-dimensional example.

(B.2.3) The possibility of (3) in dimensions $s \geq 7$ raises the question of whether the Beyer-quotient q_s (see 5.2.5) is well-defined in general. If the lengths of a MRLB's vectors are ambiguous in higher dimensions, so may be the Beyer-quotient which depends on just these lengths. Scanning the available literature we were profoundly surprised to find the problem of the Beyer-quotient's uniqueness nowhere addressed let alone answered[9].

One notable exception is Afflerbach, who states that the parallelepipeds spanned by two MRLBs of the same lattice are always congruent ("Jedoch sind die von zwei Minkowski-reduzierten Basen desselben Gitters aufgespannten Parallelepipede stets kongruent."). This claim, given in [1, p.42] without further reference or proof, is clearly invalidated by Ryshkov's counterexample.
Let Λ be the seven-dimensional lattice and let $\{\mathbf{m}_0, \ldots, \mathbf{m}_6\}$ and $\{\mathbf{n}_0, \ldots, \mathbf{n}_6\}$ be the two MRLBs of Λ given by Ryshkov in [99] such that (3) holds. Furthermore, let $\mathbf{A} := \{\sum_{i=0}^{6} t_i \mathbf{m}_i : \ 0 \leq t_i \leq 1\}$ be the parallelepiped spanned by the $\mathbf{m}_i$ and $\mathbf{B}$ that spanned by the $\mathbf{n}_i$. If $\mathbf{A}$ were congruent with $\mathbf{B}$, it could be rotated and translated such that the edges of $\mathbf{A}$ coincide with the edges of $\mathbf{B}$.
The numbers $(||\mathbf{m}_0||, \ldots, ||\mathbf{m}_6||)$ are just the lengths of those edges of $\mathbf{A}$ which start at the origin, in ascending order. More generally, $(||\mathbf{m}_0||, \ldots, ||\mathbf{m}_6||)$ are the lengths of the edges of $\mathbf{A}$ starting at any fixed of its vertices[10]. Similarly, $(||\mathbf{n}_0||, \ldots, ||\mathbf{n}_6||)$ are the lengths of the edges starting at any fixed vertex of $\mathbf{B}$. From (3), it follows that $\mathbf{A}$ and $\mathbf{B}$ are incongruent.

B.3 Covering lattices with parallel hyperplanes

(B.3.1) The goal of this section is to derive a method for computing the spectral test as introduced in (4.2.4): the maximal distance or spacing $1/\nu_s$ of parallel hyperplanes which cover a lattice Λ.

(B.3.2) A family $\mathcal{H} = \mathcal{H}_{\mathbf{n},C}$ of parallel hyperplanes in $\mathbf{R}^s$ is uniquely defined by a nonzero vector $\mathbf{n} \in \mathbf{R}^s$ and a set of scalars $C \subseteq \mathbf{R}$:

$$\mathcal{H}_{\mathbf{n},C} = \{H_{\mathbf{n},c} : \ c \in C\}\,.$$

[9] Actually, we found the Beyer-quotient applied in [1], [26], [49], [69], [87], and [92], but, except for the first, none of these even address the problem. [1] and [69] actually compute Beyer-quotients in dimensions well above seven.

[10] It is easy to see this for 2- or 3-dimensional parallelepipeds. For the general case, use induction on the dimension s.

We say that $\mathcal{H}_{\mathbf{n},C}$ is a cover of the s-dimensional lattice Λ if

$$\Lambda \subset \bigcup_{c \in C} H_{\mathbf{n},c}$$

and

$$C \text{ is the smallest set with this property.}$$

The second condition serves to avoid covers containing 'useless' hyperplanes, i.e. hyperplanes which contain no lattice point at all.

Let $\mathcal{H}_{\mathbf{n},C}$ be a cover of Λ. The distance of two hyperplanes H and H' (or, more generally, the distance of any two nonempty sets in $\mathbf{R}^s$) is defined as

$$d(H, H') := \inf \{ ||\mathbf{x} - \mathbf{y}|| : \ (\mathbf{x}, \mathbf{y}) \in H \times H' \} .$$

The shortest distance of neighbouring hyperplanes in the cover, its *spacing*, is

$$d(\mathcal{H}_{\mathbf{n},C}) := \inf \{ d(H_{\mathbf{n},c}, H_{\mathbf{n},c'}) : \ c, c' \in C, \, c \neq c' \} .$$

With these conventions, the value of the spectral test for a lattice Λ is

$$\frac{1}{\nu_s} = \sup \{ d(\mathcal{H}_{\mathbf{n},C}) : \ \mathcal{H}_{\mathbf{n},C} \text{ is a cover of } \Lambda \} .$$

(B.3.3) Observe that, in a cover of a lattice, the distance of a hyperplane H to its nearest neighbouring hyperplane H' is independent of H in the sense that the distance to the nearest neighbour is the same for any hyperplane. Although this is quite obvious by intuition, we prove it as

Lemma B.2 *Let $\mathcal{H}_{\mathbf{n},C}$ be a cover of the lattice Λ. Then C is a group.*

Proof: Let $c, c' \in C$. $H_{\mathbf{n},c}$ contains at least one lattice point $\mathbf{x}$ and $H_{\mathbf{n},c'}$ contains at least one lattice point $\mathbf{x}'$. The lattice Λ is a group, so the point $\mathbf{y} := \mathbf{x} \pm \mathbf{x}'$ is a lattice point, too. $\mathbf{y}$ is contained in $H_{\mathbf{n},c \pm c'}$ and, therefore $c \pm c' \in C$. □

The group structure of C yields

$$d(H_{\mathbf{n},c}, H_{\mathbf{n},c'}) = d(H_{\mathbf{n},0}, H_{\mathbf{n},c-c'}),$$

and hence

$$d(\mathcal{H}_{\mathbf{n},C}) = \inf \{ d(H_{\mathbf{n},0}, H_{\mathbf{n},c}) : \ c \in C \setminus \{0\} \} .$$

With this, we can move a step towards the actual computation of $1/\nu_s$ in[11]

Lemma B.3 *Let $\mathcal{H}_{\mathbf{n},C}$ be a cover of Λ. Then*

$$d(\mathcal{H}_{\mathbf{n},C}) = \frac{1}{||\mathbf{n}||} \inf \{|c| : \; c \in C \setminus \{0\}\} .$$

Note that we cannot replace the above infimum over $C \setminus \{0\}$ by a minimum. For $\Lambda := \mathbf{Z} \times \mathbf{Z}$, $\mathbf{n} := (\sqrt{2}, 1)$, and $C := \{a + b\sqrt{2} : \; a, b \in \mathbf{Z}\}$, the set $\mathcal{H}_{\mathbf{n},C}$ is a cover of Λ. However, C is dense[12] in $\mathbf{R}$, and therefore the spacing $d(\mathcal{H}_{\mathbf{n},C})$ is equal to 0.

Proof: Let $c \in C \setminus \{0\}$. Due to the group structure of C, we may assume $c > 0$. For any $\mathbf{x} \in H_{\mathbf{n},0}$ and any $\mathbf{y} \in H_{\mathbf{n},c}$, there is $||\mathbf{x} - \mathbf{y}|| = ||(\mathbf{x} - \mathbf{x}) - (\mathbf{y} - \mathbf{x})||$ and $\mathbf{y} - \mathbf{x}$ is contained in $H_{\mathbf{n},c}$. So

$$d(H_{\mathbf{n},0}, H_{\mathbf{n},c}) = \inf \{||\mathbf{y}|| : \; \mathbf{y} \in H_{\mathbf{n},c}\} .$$

Any $\mathbf{y} \in H_{\mathbf{n},c}$ can be represented as the sum of two vectors $\mathbf{y} = \mathbf{m} + \mathbf{y}'$, where

$$\mathbf{m} := \frac{c}{||\mathbf{n}||^2}\mathbf{n}$$

is contained in $H_{\mathbf{n},c}$ and $\mathbf{y}' \in H_{\mathbf{n},0}$. Note that, given $\mathbf{m}$, $\mathbf{y}'$ is uniquely determined by $\mathbf{y}$. We get

$$\begin{aligned} ||\mathbf{y}||^2 &= <\mathbf{y}, \mathbf{y}> \\ &= <\mathbf{m} + \mathbf{y}', \mathbf{m} + \mathbf{y}'> \\ &= ||\mathbf{m}||^2 + ||\mathbf{y}'||^2 + 2 \; <\mathbf{m}, \mathbf{y}'> \\ &= ||\mathbf{m}||^2 + ||\mathbf{y}'||^2 + 2\frac{c}{||\mathbf{n}||^2} \; <\mathbf{n}, \mathbf{y}'> \; . \end{aligned}$$

Since $\mathbf{y}' \in H_{\mathbf{n},0}$, the last term is equal to zero. Hence

$$||\mathbf{y}||^2 \geq ||\mathbf{m}||^2.$$

This observation and the triangle inequality yield

$$||\mathbf{m}|| \leq ||\mathbf{y}|| \leq ||\mathbf{m}|| + ||\mathbf{y}'||.$$

[11] Assuming $C \subseteq \mathbf{Z}$, this lemma is similar to [39, Equation (13)].

[12] This is an immediate consequence of Kronecker's Approximation Theorem [63, 64]; see also Hlawka [55, p.3].

Note that $\mathbf{y}' \in H_{\mathbf{n},0}$ is not only uniquely determined by $\mathbf{y}$, but any such $\mathbf{y}'$ uniquely determines a point $\mathbf{y} = \mathbf{m} + \mathbf{y}'$ in $H_{\mathbf{n},c}$. The infimum of $||\mathbf{y}||$ taken over all points in $H_{\mathbf{n},c}$ is therefore equal to $||\mathbf{m}||$. Hence

$$\begin{aligned} d(H_{\mathbf{n},0}, H_{\mathbf{n},c}) &= ||\mathbf{m}|| \\ &= \frac{c}{||\mathbf{n}||}, \end{aligned}$$

which means that $d(\mathcal{H}_{\mathbf{n},C})$ is equal to the infimum of all values of the form $\frac{c}{||\mathbf{n}||}$ for positive $c \in C$. □

The value of the spectral test for a lattice Λ can thus be written as

$$\frac{1}{\nu_s} = \sup \left\{ \frac{1}{||\mathbf{n}||} \inf \{|c| : c \in C \setminus \{0\}\} : \mathcal{H}_{\mathbf{n},C} \text{ is a cover of } \Lambda \right\}.$$

(B.3.4) In this formula for $1/\nu_s$, we consider all covers of a lattice. We now show that it is sufficient to restrict to covers of a certain kind which are uniquely defined by the lattice itself.

Proposition B.6 *Let $\mathcal{H}_{\mathbf{n},C}$ be a cover of Λ. Then*

$$d(\mathcal{H}_{\mathbf{n},C}) > 0$$

if and only if

$$\lambda\mathbf{n} \in \Lambda^\star \text{ is primitive}$$

for some $\lambda > 0$.

Proof: Let $\mathcal{H}_{\mathbf{n},C}$ be a cover of Λ. Observe that, for any $\lambda > 0$, the definitions of a cover and of its spacing yield that

$$\mathcal{H}_{\mathbf{n},C} = \mathcal{H}_{\lambda\mathbf{n},\lambda C}.$$

For the 'if'-part, let $c_0 := d(\mathcal{H}_{\mathbf{n},C}) > 0$. Since C is a group and c_0 is its smallest positive element (Lemma B.2 and B.3), we have $C = c_0\mathbf{Z}$. This and the above observation yield

$$\begin{aligned} \mathcal{H}_{\mathbf{n},C} &= \mathcal{H}_{\mathbf{n},c_0\mathbf{Z}} \\ &= \mathcal{H}_{1/c_0\mathbf{n},\mathbf{Z}}. \end{aligned}$$

Let $\lambda := 1/c_0$ (which is, of course, positive). Since $\mathbf{Z} = \{ <\lambda\mathbf{n}, \mathbf{p}> : \mathbf{p} \in \Lambda\}$, the vector $\lambda\mathbf{n}$ is in the dual lattice $\Lambda^\star$.
Let $\mu\lambda\mathbf{n} \in \Lambda^\star \cap [\lambda\mathbf{n}]$. To prove that $\lambda\mathbf{n}$ is primitive, we have to show that μ is an integer. Since $\mu\lambda\mathbf{n} \in \Lambda^\star$, $<\mu\lambda\mathbf{n}, \mathbf{p}>$ is an integer for any $\mathbf{p} \in \Lambda$. Because c_0 is the smallest positive element in the discrete group $C = c_0\mathbf{Z}$, there is a $\mathbf{p}_0 \in \Lambda$ such that $<\mathbf{n}, \mathbf{p}_0> = c_0$. The choice of λ and $\mathbf{p}_0$ yield

$$\begin{aligned} <\mu\lambda\mathbf{n}, \mathbf{p}_0> &= \mu\frac{1}{c_0} <\mathbf{n}, \mathbf{p}_0> \\ &= \mu, \end{aligned}$$

so μ is an integer.

For the 'only if'-part, let $\lambda\mathbf{n} \in \Lambda^\star$ be primitive. Then $\mathcal{H}_{\mathbf{n},C} = \mathcal{H}_{\lambda\mathbf{n},\lambda C}$ is a cover of Λ and λC is a subgroup of $\mathbf{Z}$. Since $\{0\} \neq \lambda C$ (otherwise, the whole lattice Λ would be contained in just one hyperplane $H_{\lambda\mathbf{n},0}$, which is impossible), it contains a smallest nonzero element. With Proposition B.3, we get

$$\begin{aligned} d(\mathcal{H}_{\mathbf{n},C}) &= d(\mathcal{H}_{\lambda\mathbf{n},\lambda C}) \\ &= \frac{1}{||\lambda\mathbf{n}||}\inf\{|\lambda c| : \lambda c \in \lambda C \setminus \{0\}\} \\ &> 0. \end{aligned}$$

□

There is, in fact, more information in the above proof than stated in Proposition B.6. In the 'if'- part, we assumed that $d(\mathcal{H}_{\mathbf{n},C}) > 0$ and concluded that

$$\mathcal{H}_{\mathbf{n},C} = \mathcal{H}_{\lambda\mathbf{n},\mathbf{Z}}$$

for $\lambda\mathbf{n} \in \Lambda^\star$ being primitive. Note that the spacing of this cover is $d(\mathcal{H}_{\lambda\mathbf{n},\mathbf{Z}}) = 1/||\lambda\mathbf{n}||$. In accordance with Knuth [59, Section 3.3.4] and Ripley [93, Theorem 2.13], we can conclude:

> The value of the spectral test for a lattice Λ is
>
> $$\frac{1}{\nu_s} = \frac{1}{||\mathbf{m}||},$$
>
> where $\mathbf{m}$ is the shortest nonzero vector in $\Lambda^\star$.

B.4 Linear transformations of lattices

(B.4.1) Throughout this section, let

$$\Lambda = \left\{ \sum_{i=0}^{s-1} k_i \mathbf{b}_i \,:\; k_i \in \mathbf{Z} \right\}$$

be the s-dimensional lattice spanned by the $\mathbf{b}_i$ and, for $s \geq t$, let

$$T : \mathbf{R}^s \longrightarrow \mathbf{R}^t$$

be linear and surjective. This mapping is uniquely determined by a $t \times s$-matrix which we also call T. The image of Λ under T is

$$T\Lambda = \left\{ \sum_{i=0}^{s-1} k_i T \mathbf{b}_i \,:\; k_i \in \mathbf{Z} \right\}.$$

In this section, we give a sufficient condition on T for $T\Lambda$ to be a lattice.

(B.4.2) The set $T\Lambda$ does already look quite similar to a lattice: it is made up of all integer linear combinations of the vectors $T\mathbf{b}_i$. The only problem is that these s vectors are not necessarily linearly independent in $\mathbf{R}^t$ and some of them may even be equal to $\mathbf{0}$.
For another reason why $T\Lambda$ does not need to be a lattice, recall our observation from (B.3.3): the set $C := \{a + b\sqrt{2} :\; a, b \in \mathbf{Z}\}$ is dense in $\mathbf{R}$. Setting $\Lambda := \mathbf{Z} \times \mathbf{Z}$ and $T = (1, \sqrt{2})$ yields that $T\Lambda = C$ is dense in $\mathbf{R}^1$ and therefore cannot be a lattice.

(B.4.3) In some cases however, one of them being of special interest when studying congruential random number generators, $T\Lambda$ is a lattice.

Proposition B.7 *If every row-vector of T is in $\Lambda^\star$ (up to a constant scaling factor), then $T\Lambda$ is a lattice.*

Proof: Let every row-vector $\mathbf{n}_i$ $(0 \leq i < t)$ of T be in $\Lambda^\star$ (up to a constant scaling factor). Without loss of generality, we assume that the scaling factor is 1, so $\mathbf{n}_i \in \Lambda^\star$. We show that $T\Lambda$ is a discrete subgroup of $\mathbf{R}^t$ which is not contained in a $(t-1)$-dimensional subspace of $\mathbf{R}^t$. With this and Proposition B.2, $T\Lambda$ is a lattice.

From the above representation of $T\Lambda$, it is clear that with any two points $\mathbf{x}$ and $\mathbf{y}$, it contains $\mathbf{x} \pm \mathbf{y}$ as well. Hence $T\Lambda$ is a group.

Λ is a lattice and therefore not contained in an $(s-1)$-dimensional subspace of $\mathbf{R}^s$. Since T is surjective, $T\mathbf{R}^s$ is equal to $\mathbf{R}^t$. Using the linearity of T, we get that $T\Lambda$ is not contained in a $(t-1)$-dimensional subspace of $\mathbf{R}^t$.

For a basis $\{\mathbf{b}_0, \ldots, \mathbf{b}_{s-1}\}$ of Λ, we have

$$T\mathbf{b}_i = \begin{pmatrix} <\mathbf{n}_0, \mathbf{b}_i> \\ \vdots \\ <\mathbf{n}_{t-1}, \mathbf{b}_i> \end{pmatrix} \qquad (0 \leq i < s).$$

All the $\mathbf{n}_j$ are in $\Lambda^\star$, so all the $T\mathbf{b}_i$ are in $\mathbf{Z}^t$. The set $T\Lambda$ is made up of all integer combinations of the $T\mathbf{b}_i$, so $T\Lambda \subseteq \mathbf{Z}^t$. Finally, since $\mathbf{Z}^t$ is discrete, so is $T\Lambda$. □

(B.4.4) We have to stress that the given condition for $T\Lambda$ being a lattice, although sufficient, is far from being necessary. For example, consider the lattice $\Lambda := \mathbf{Z} \times \mathbf{Z}$ an the bijective linear transformation defined by

$$T := \begin{pmatrix} 1 & \sqrt{2} \\ 1 & -\sqrt{2} \end{pmatrix}.$$

Since $\sqrt{2}$ is irrational, the row-vectors of T cannot be scaled to lie in $\Lambda^\star \backslash \{\mathbf{0}\}$. However, we will see that $T\Lambda$ is a lattice.

Using the same argument as in the proof above yields that $T\Lambda$ is a subgroup of $\mathbf{R}^2$ which is not contained in a one-dimensional subspace of $\mathbf{R}^2$. It remains to show that $T\Lambda$ is discrete.
Observe that for any $k \in \mathbf{Z}$, we have

$$\begin{aligned} T\{(k,l) : \ l \in \mathbf{Z}\} &= \{(k + l\sqrt{2}, k - l\sqrt{2}) : \ l \in \mathbf{Z}\} \\ &= (k,k) + \{l(\sqrt{2}, -\sqrt{2}) : \ l \in \mathbf{Z}\} \\ &=: k(1,1) + L. \end{aligned}$$

It is clear that L is a discrete set. The euclidean distance of each point to its nearest neighbours in L is equal to 2. Moreover, note that L is contained in a hyperplane through the origin. We have

$$T\Lambda \ = \ \bigcup_{k \in \mathbf{Z}} T\{(k,l) : \ l \in \mathbf{Z}\}$$

$$= \bigcup_{k \in \mathbf{Z}} k(1,1) + L.$$

Each individual set $k(1,1) + L$ is discrete. For $k \neq k'$, the set $k'(1,1) + L$ is just $k(1,1) + L$ shifted by the nonzero vector $(k' - k)(1,1)$. Since these two sets are disjoint and their distance is positive, their union is discrete, too. By induction, it follows that the union of all the sets $k(1,1) + L$ for $k \in \mathbf{Z}$, i.e. $T\Lambda$, is discrete.

(B.4.5) The reason we have presented Proposition B.7 at all is its use for studying critical distances in congruential generators, i.e. generators $(x_n)_{n=0}^{N-1}$ whose numbers have the form $x_n = u_n/M$ for $u_n, M \in \mathbf{Z}$ (virtually all random number generators well suited for computer implementation are of this type; see (5.1.2)). We will see that the existence of a lattice structure in the s-dimensional points $\mathbf{x}_n = (x_n, \ldots, x_{n+s-1})$ leads to a lattice structure in the 2-dimensional points $\mathbf{x}'_n = (x_n, x_{n+s-1})$.

As noted in (5.2.6), the $\mathbf{x}'_n$ can be written as $\mathbf{x}'_n = T\mathbf{x}_n$, where T is the $2 \times s$-matrix

$$T := \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

For this T, we have the following result[13].

Proposition B.8 *Let Λ be a lattice in $\mathbf{R}^s$ which contains $\mathbf{Z}^s$; furthermore, let $\Lambda' := T\Lambda$. Then*

$$\Lambda' \textit{ is a lattice} \tag{1}$$

and, for any $\mathbf{x} \in \mathbf{R}^s$,

$$T((\mathbf{x} + \Lambda) \cap [0,1[^s) = (T\mathbf{x} + \Lambda') \cap [0,1[^2. \tag{2}$$

The meaning of the condition $\mathbf{Z}^s \subseteq \Lambda$ is explained in

Lemma B.4 *Let Λ be a lattice containing $\mathbf{Z}^s$. Then there exists a positive integer M such that*

$$\Lambda \subseteq \frac{1}{M}\mathbf{Z}^s.$$

[13] Which is not as obvious as it seems; see (B.4.4).

Proof: First, we show that Λ is made up of rational points, i.e. $\Lambda \subseteq \mathbf{Q}^s$, and then we choose M accordingly.

Assume there is a lattice point $\mathbf{p} \in \Lambda \setminus \mathbf{Q}^s$. Since at least one coordinate of $\mathbf{p}$ is irrational, the points[14]

$$\{k\mathbf{p}\} \qquad (k \in \mathbf{Z})$$

are all distinct. Since the s coordinates of each $\{k\mathbf{p}\}$ are between 0 and 1, we have $||\{k\mathbf{p}\}|| \leq \sqrt{s}$. This means that Λ contains an infinite number of points $\{k\mathbf{p}\}$ whose norm is at most $\sqrt{s}$. But Λ is discrete, so this cannot be. Λ is made up of rational points.

Next, observe that any $\mathbf{p} \in \Lambda$ can be uniquely represented as the sum of two lattice points

$$\mathbf{p} = \lfloor \mathbf{p} \rfloor + \{\mathbf{p}\},$$

where $\{\mathbf{p}\} \in [0,1[^s$. There is just a finite number of lattice points in $[0,1[^s$, all of which are rational. We can choose an integer $M > 0$ such that all the $\{\mathbf{p}\}$ are in $\frac{1}{M}\mathbf{Z}^s$. Since the integer points $\lfloor \mathbf{p} \rfloor$ are in $\frac{1}{M}\mathbf{Z}^s$ too, the proof is complete. □

Proof of Proposition B.8: It suffices to show (1). The proof of (2) is completely equivalent to the proof of the corresponding part of Proposition 5.2.

Let Λ be an s-dimensional lattice and let T be defined as above. With Lemma B.4, there is an integer $M > 0$ such that $\Lambda \subseteq 1/M\mathbf{Z}^s$. Multiplying each row-vector $\mathbf{n}_i$ $(i = 0, 1)$ of T by M, we get $M\mathbf{n}_0, M\mathbf{n}_1 \in \Lambda^\star$. Due to Proposition B.7, the points (x_n, x_{n+s-1}) form a lattice in $\mathbf{R}^2$. □

[14] See (5.2.6) for the meaning of the notation $\{\mathbf{x}\}$ and $\lfloor \mathbf{x} \rfloor$ for s-dimensional points $\mathbf{x}$.

Bibliography

[1] L. Afflerbach. Die Gütebewertung von Pseudo-Zufallszahlen-Generatoren aufgrund theoretischer Analysen und algorithmischer Berechnungen. *Grazer Mathematische Berichte*, **309**, 1990.

[2] J. Ahrens, U. Dieter, and A. Grube. Pseudo-random numbers: a new proposal for the choice of multiplicators. *Computing*, **6**:121–138, 1970.

[3] S.L. Anderson. Random number generators on vector supercomputers and other advanced architectures. *SIAM Rev.*, **32**:221–251, 1990.

[4] I. Balásházy and W. Hofmann. *Particle deposition in airway bifurcations for inspiratory flow. Part II: calculations of particle trajectories and deposition patterns.* Hungarian Academy of Sciences, Central Research Institute for Physics, Budapest, 1992.

[5] W.A. Beyer, R.B. Roof, and D. Williamson. The lattice structure of multiplicative congruential pseudo-random vectors. *Math. Comp.*, **25**:345–363, 1971.

[6] T. Bonnesen and W. Fenchel. *Theorie der konvexen Körper.* Chelsea Publishing Company, New York, 1971. Reprint; first published by Springer, Berlin, 1934. In German.

[7] K.O. Bowman and M.T. Robinson. Studies of random number generators for parallel processing. In M.T. Heath, editor, *Proc. Second Conference on Hypercube Multiprocessors*, pages 445–453, Philadelphia, 1987. SIAM.

[8] P. Bratley, B.L. Fox, and L.E. Schrage. *A Guide to Simulation.* Springer, New York, 2nd edition, 1987.

[9] K.V. Bury. *Statistical Models in Applied Science.* Robert E. Krieger Publishing Company, Inc., Malabar, Florida, reprint edition, 1986.

[10] R.P. Chambers. Random number generation. *IEEE Spectrum*, **4**:48–56, 1967.

[11] D.L. Cohn. *Measure Theory.* Birkhäuser, Boston, 1980.

[12] A. Compagner. Definitions of randomness. *Am. J. Phys.*, **59**:700–705, 1991.

[13] R.R. Coveyou. Serial correlation in the generation of pseudo-random numbers. *J. Assoc. Comput. Mach.*, **7**:72–74, 1960.

[14] R.R. Coveyou and R.D. MacPherson. Fourier analysis of uniform random number generators. *J. Assoc. Comp. Mach.*, **14**:100–119, 1967.

[15] B. De Finetti. *Theory of Probability*, volume 1. John Wiley, New York, 1979.

[16] A. De Matteis, J. Eichenauer-Hermann, and H. Grothe. Computation of critical distances within multiplicative congruential pseudorandom number sequences. *J. Comp. Appl. Math.*, **39**:49–55, 1992.

[17] A. De Matteis and B. Faleschini. Some arithmetical properties in connection with pseudo-random numbers. *Boll. Unione Mat. Ital.*, **18**:171–184, 1963.

[18] A. De Matteis and S. Pagnutti. Parallelization of random number generators and long-range correlations. *Numer. Math.*, **53**:595–608, 1988.

[19] A. De Matteis and S. Pagnutti. A class of parallel random number generators. *Parallel Comput.*, **13**:193–198, 1990.

[20] A. De Matteis and S. Pagnutti. Long-range correlations in linear and non-linear random number generators. *Parallel Comput.*, **14**:207–210, 1990.

[21] A. De Matteis and S. Pagnutti. Long-range correlation analysis of the Wichmann-Hill random number generator. *Statistics and Computing*, **3**:67–70, 1993.

[22] I. Deák. Uniform random number generators for parallel computers. *Parallel Comput.*, **15**:155–164, 1990.

[23] L.P. Devroye. *Non-Uniform Random Variate Generation.* Springer, New York, 1986.

[24] U. Dieter. Autokorrelation multiplikativ erzeugter Pseudo-Zufallszahlen. *Operations Research Verfahren*, **6**:69–85, 1968.

[25] U. Dieter. How to calculate shortest vectors in a lattice. *Math. Comp.*, **29**:827–833, 1975.

[26] U. Dieter. Erzeugung von gleichverteilten Zufallszahlen. In *Jahrbuch Überblicke Mathematik 1993*, pages 25–44, Braunschweig, 1993. Vieweg.

[27] U. Dieter and J.H. Ahrens. An exact determination of serial correlations of pseudo-random numbers. *Numer. Math.*, **17**:101–123, 1971.

[28] U. Dieter and J.H. Ahrens. *Uniform random numbers.* Inst. f. Math. Stat., Technische Hochschule Graz, Graz, 1974.

[29] E.J. Dudewicz and T.G. Ralley. *The Handbook of Random Number Generation and Testing With TESTRAND Computer Code*, volume 4 of *American Series in Mathematical and Management Sciences.* American Sciences Press, Inc., Columbus, Ohio, 1981.

[30] W.F. Eddy. Random number generators for parallel processors. *J. Comp. Appl. Math.*, **31**:63–71, 1990.

[31] J. Eichenauer and J. Lehn. A non-linear congruential pseudo random number generator. *Statist. Papers*, **27**:315–326, 1986.

[32] J. Eichenauer-Hermann. Inversive congruential pseudorandom numbers avoid the planes. *Math. Comp.*, **56**:297–301, 1991.

[33] J. Eichenauer-Hermann and H. Grothe. A remark on long-range correlations in multiplicative congruential pseudo random number generators. *Numer. Math.*, **56**:609–611, 1989.

[34] J. Eichenauer-Herrmann. Statistical independence of a new class of inversive congruential pseudorandom numbers. *Math. Comp.*, **60**:375–384, 1993.

[35] K. Entacher. Selected random number generators in the run test. Preprint, Mathematics Institute, University of Salzburg.

[36] K. Entacher. Selected random number generators in the run test II: subsequence behavior. Article in preparation, Mathematics Institute, University of Salzburg.

[37] G.S. Fishman. Multiplicative congruential random number generators with modulus 2^{β}: an exhaustive analysis for $\beta = 32$ and a partial analysis for $\beta = 48$. *Math. Comp.*, **54**:331–344, 1990.

[38] G.S. Fishman and L.R. Moore. A statistical evaluation of multiplicative congruential random number generators with modulus $2^{31} - 1$. *J. Amer. Statist. Assoc.*, **77**:129–136, 1982.

[39] G.S. Fishman and L.R. Moore. An exhaustive analysis of multiplicative congruential random number generators with modulus $2^{31} - 1$. *SIAM J. Sci. Statist. Comput.*, **7**:24–45, 1986.

[40] M. Flahive and H. Niederreiter. On inversive congruential generators for pseudorandom numbers. In G.L. Mullen and Shiue P.J.-S., editors, *Finite Fields, Coding Theory, and Advances in Communications and Computing*, pages 75–80, New York, 1992. Dekker.

[41] P. Frederickson, R. Hiromoto, and T.L. Jordan. Pseudo-random trees in Monte Carlo. *Parallel Comput.*, **1**:175–180, 1984.

[42] H. Fuss. Simulating 'fair' random numbers. In G.C. Vansteenkiste, E.J.H. Kerckhoffs, L. Dekker, and J.C. Zuklervaart, editors, *Proceedings of the 2nd European Simulation Congress, Sept. 9–12, 1986*, pages 252–257, Belgium, 1986. Simulation Councils, Inc.

[43] M. Greenberger. An a priori determination of serial correlation in computer generated random numbers. *Math. Comp.*, **15**:383–389, 1961.

[44] M. Greenberger. Method in randomness. *Comm. ACM*, **8**:177–179, 1965.

[45] P.M. Gruber. Geometry of numbers. In P.M. Gruber and J.M. Wills, editors, *Handbook of Convex Geometry*, volume B, pages 739–764, Amsterdam, 1993. Elsevier Science Publishers B.V.

[46] P.M. Gruber and C.G. Lekkerkerker. *Geometry of Numbers.* Elsevier Science Publishers B.V., Amsterdam, 2nd edition, 1987.

[47] J.H. Halton. Pseudo-random trees: multiple independent sequence generators for parallel and branching computations. *J. Comp. Physics*, **84**:1–56, 1989.

[48] J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods.* Chapman and Hall, London, 1983.

[49] F. Härtel. *Zufallszahlen für Simulationsmodelle.* PhD thesis, Hochschule St. Gallen für Wirtschafts-, Rechts- und Sozialwissenschaften, St. Gallen, 1994.

[50] P. Hellekalek. General discrepancy estimates v: diaphony and the spectral test. Preprint, Mathematics Institute, University of Salzburg.

[51] P. Hellekalek. Study of algorithms for primitive polynomials. Report D5H-1, CEI-PACT Project, WP5.1.2.1.2, Research Institute for Software Technology, University of Salzburg, Austria, 1994.

[52] P. Hellekalek, M. Mayer, and A. Weingartner. Implementation of algorithms for IMP-polynomials. Report D5H-2, CEI-PACT Project, WP5.1.2.1.2, Research Institute for Software Technology, University of Salzburg, Austria, 1994.

[53] G.W. Hill. Cyclic properties of pseudo-random sequences of Mersenne prime residues. *Comput. J.*, **22**:80–85, 1979.

[54] E. Hlawka. Zur angenäherten Berechnung mehrfacher Integrale. *Mh. Math.*, **66**:140–151, 1962.

[55] E. Hlawka, editor. *Theorie der Gleichverteilung.* Bibliographisches Institut, Mannheim, 1970.

[56] L. Holmild and K. Rynefors. Uniformity of congruential pseudo-random number generators. Dependence on length of number sequence and resolution. *J. Comp. Physics*, **26**:297–306, 1978.

[57] M.H. Kalos and P.A. Whitlock. *Monte Carlo Methods*, volume 1: Basics. John Wiley, New York, 1986.

[58] D.G. Kendall and B. Babington-Smith. Randomness and random sampling numbers. *J. Royal Statist. Soc.*, **101**:146–166, 1938.

[59] D.E. Knuth. *The Art of Computer Programming*, volume 2: Seminumerical Algorithms. Addison-Wesley, Reading, MA, 2nd edition, 1981.

[60] J.F. Koksma. Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo 1. *Mathematica B (Zutphen)*, **11**:7–11, 1942/1943.

[61] A.N. Kolmogorov. Grundbegriffe der Wahrscheinlichkeitsrechnung. In *Ergebnisse der Mathematik und ihrer Grenzgebiete*, volume **2**, Berlin, 1933.

[62] H.M. Korobov. Anwendung zahlentheoretischer Methoden auf Probleme der numerischen Mathematik. *Moskau*, 1963.

[63] L. Kronecker. Die Periodensysteme von Functionen reeller Variablen. In *Mathematische Werke*, volume 3, pages 31–46, New York, 1968. Chelsea Publishing Company. Reprint.

[64] L. Kronecker. Näherungsweise ganzzahlige Auflösung linearer Gleichungen. In *Mathematische Werke*, volume 3, pages 47–109, New York, 1968. Chelsea Publishing Company. Reprint.

[65] J.C. Lagarias. Pseudorandom numbers. *Statistical Science*, **8** :31–39, 1993.

[66] G. Larcher. A class of low-discrepancy point-sets and its application to numerical integration by number-theoretical methods. *Grazer Mathematische Berichte*. To appear.

[67] G. Larcher and C. Traunfellner. On the numerical integration of Walsh-series by number-theoretical methods. *Math. Comp.*, **63**:277–291, 1994.

[68] P. L'Ecuyer. Efficient and portable combined random number generators. *Comm. ACM*, **31**:742–774, 1988.

[69] P. L'Ecuyer. Random numbers for simulation. *Comm. ACM*, **33**:85–97, 1990.

[70] H. Leeb. On the digit test. In P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Tagungsband zum 1. Salzburger Minisymposium über Pseudozufallszahlen und Quasi-Monte Carlo Methoden am 18. Nov. 1994*, Salzburg. Mathematics Institute, University of Salzburg. To appear.

[71] H. Leeb. Selected random number generators in the digit test. Preprint, Mathematics Institute, University of Salzburg.

[72] H. Leeb. Selected random number generators in the digit test II: subsequence behavior. Article in preparation, Mathematics Institute, University of Salzburg.

[73] D.H. Lehmer. Mathematical methods in large-scale computing units. In *Proc. 2nd Sympos. on Large-Scale Digital Calculating Machinery, Cambridge, MA, 1949*, pages 141–146, Cambridge, MA, 1951. Havard University Press.

[74] K. Leichtweiß. *Konvexe Mengen*. Springer, Berlin, 1980.

[75] R. Lidl and H. Niederreiter. *Finite Fields*. Addison-Wesley, Reading, MA, 1983.

[76] M.D. MacLaren and G. Marsaglia. Uniform random number generators. *J. Assoc. Comput. Mach.*, **12**:83–89, 1965.

[77] N.M. MacLaren. A limit on the usable length of a pseudorandom sequence. *J. Statist. Comput. Simul.*, **42**:47–54, 1992.

[78] G. Marsaglia. Random numbers fall mainly in the planes. *Proc. Nat. Acad. Sci. USA*, **61**:25–28, 1968.

[79] G. Marsaglia. Regularities in congruential random number generators. *Numer. Math.*, **16**:8–10, 1970.

[80] G. Marsaglia. A current view of random number generators. In L. Billard, editor, *Computer Science and Statistics: The Interface*, pages 3–10, Amsterdam, 1985. Elsevier Science Publishers B.V.

[81] N. Metropolis and S.M. Ulam. The Monte Carlo method. *J. Amer. Statist. Assoc.*, **44**:335–341, 1949.

[82] F. Neuman and C.F. Martin. The autocorrelation structure of Tausworthe pseudo-random number generators. *IEEE Trans. Comput.*, **C-25**:460–464, 1976.

[83] F. Neuman and R. Merrick. Autocorrelation peaks in congruential pseudorandom number generators. *IEEE Trans. Comput.*, **C-25**:457–460, 1976.

[84] J. von Neumann. *Memorandum on the program of the high speed computer.* Institute of Advanced Study, Princeton, NJ, 1945.

[85] H. Niederreiter. On a new class of pseudorandom numbers for simulation methods. *J. Comput. Appl. Math.* To appear.

[86] H. Niederreiter. Recent trends in random number generation and random vector generation. *Ann. Oper. Res.*, **31**:323–345, 1991.

[87] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods.* SIAM, Philadelphia, 1992.

[88] H. Niederreiter. oral communication, Nov. 18, 1994.

[89] W. Parry. *Topics in Ergodic Theory.* Cambridge University Press, Cambridge, MA, 1981.

[90] P. Peach. Bias in pseudo-random numbers. *J. Amer. Statist. Assoc.*, **56**:610–618, 1961.

[91] K. Petersen. *Ergodic Theory.* Cambridge University Press, Cambridge, MA, 1983.

[92] B.D. Ripley. The lattice structure of pseudo-random number generators. *Proc. Roy. Soc. London Ser. A*, **389**:197–204, 1983.

[93] B.D. Ripley. *Stochastic Simulation.* John Wiley, New York, 1987.

[94] B.D. Ripley. Uses and abuses of statistical simulation. *Mathematical Programming*, **42**:53–68, 1988.

[95] B.D. Ripley. Thoughts on pseudorandom number generators. *J. Comput. Appl. Math.*, **31**:153–163, 1990.

[96] B.D. Ripley and M.D. Kirkland. Iterative simulation methods. *J. Comput. Appl. Math.*, **31**:165–172, 1990.

[97] A. Rotenberg. A new pseudo-random number generator. *J. Assoc. Comput. Mach.*, **7**:75–77, 1960.

[98] S.S. Ryshkov. On the reduction theory of positive quadratic forms. *Dokl. Akad. Nauk SSSR*, **198**:1028–1031, 1971. = Soviet Math. Dokl. **12**, 946–950 (1971).

[99] S.S. Ryshkov. On Hermite, Minkowski and Venkov reduction of positive quadratic forms in n variables. *Dokl. Akad. Nauk SSSR*, **207**:1054–1056, 1972. = Soviet Math. Dokl. **13**, 1676–1679 (1973).

[100] G. Sawitzki. Another random number generator which should be avoided. *Statist. Soft. Newsl.*, **11**:81–82, 1985.

[101] W.C. Schmid. Zur numerischen Integration von Walsh-Reihen. Master's thesis, University of Salzburg, 1993.

[102] R. Schneider. *Convex Bodies: The Brunn–Minkowski Theory.* Press Syndicate of the University of Cambridge, Cambridge, MA, 1993.

[103] C.P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit*, volume 218 of *Lecture Notes in Math.* Springer, Berlin, 1971.

[104] C.S. Smith. Multiplicative pseudo-random number generators with prime modulus. *J. Assoc. Comput. Mach.*, **18**:587–593, 1971.

[105] I.M. Sobol'. The distribution of points in a cube and the approximate evaluation of integrals. *Zh. Vychisl. Mat. i Mat. Fiz.*, **7**:784–802, 1967. In Russian.

[106] I.M. Sobol'. *Multidimensional Quatrature Formulas and Haar Functions.* Izdat. "Nauka", Moscow, 1969. In Russian.

[107] H. Stegbuchner. Zur quantitativen Theorie der Gleichverteilung mod 1. Arbeitsberichte, Mathematisches Institut der Universität Salzburg, Salzburg, Austria, 1980.

[108] T.G. Stoelinga. *Convexe Puntverzamelingen.* PhD thesis, Groningen, 1932. See Zentralblatt für Mathematik und ihre Grenzgebiete, volume 5, pages 256–257, Springer, Berlin, 1933.

[109] R.C. Tausworthe. Random numbers generated by linear recurrence modulo two. *Math. Comp.*, **19**:201–209, 1965.

[110] W.E. Thomson. A modified congruence method for generating pseudorandom numbers. *Comp. J.*, **1**:83–86, 1958.

[111] P. Walters. *Ergodic Theory – Introductory Lectures*, volume 458 of *Lecture notes in Math.* Springer, Berlin, 1975.

[112] S. Wegenkittl. Empirical testing of pseudorandom number generators. Master's thesis, University of Salzburg, 1995. In preparation.

[113] A. Weingartner. Nonlinear congruential pseudorandom number generators. Master's thesis, University of Salzburg, 1994.

[114] B.A. Wichmann and I.D. Hill. An efficient and portable pseudo-random number generator. *Appl. Statist.*, **31**:188–190, 1982.

[115] B.A. Wichmann and I.D. Hill. Building a random-number generator. *Byte*, **12**:127–128, 1987.

[116] S.K. Zaremba. The mathematical basis of Monte Carlo and quasi-Monte Carlo methods. *SIAM Rev.*, **10**:303–314, 1968.

[117] P. Zinterhof. Über einige Abschätzungen bei der Approximation von Funktionen mit Gleichverteilungsmethoden. *Sitzungsber. Österr. Akad. Wiss. Math.-Natur. Kl. II*, **185**:121–132, 1976.

Curriculum vitae

Name: Hannes Leeb

Date of birth: September 4, 1968

Place of birth: Klagenfurt, Austria

Parents: Peter and Barbara Leeb

Graduations:
1979: Volksschule in Ebene Reichenau
1987: Matura, Gymnasium BRG in Salzburg